Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

DISSERTATION

Hate Speech Detection and Analysis for Low-Resource Languages: The Case for Amharic

Abinew Ali Ayele

Language Technology
Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences

University of Hamburg
Hamburg, Germany

A dissertation submitted for the degree of

Doctor rerum naturalium (Dr. rer. nat.)

December 2024

Hate Speech Detection and Analysis for Low-Resource Languages: The Case for Amharic

Dissertation submitted by: Abinew Ali Ayele

Date of Submission: 23.12.2024
Date of Disputation: : 28.04.2025

Supervisor: Prof. Dr. Chris Biemann, University of Hamburg
Co-Supervisor: Dr. Seid Muhie Yimam, University of Hamburg

Committee:
1st Examiner: Prof. Dr. Chris Biemann, University of Hamburg
2nd Examiner: Prof. Dr. Anne Lauscher, University of Hamburg
Chair: Prof. Dr. Stefan Wermter, University of Hamburg

University of Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Language Technology

Affidavit

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

06.05.2025

Date

Signature

(Abinew Ali Ayele)

This work is dedicated to my parents:
Ali Ayele Hassen and Tobia Hassen Abebe.

Acknowledgements

First and foremost, Alhamdulillah, praise be to God, the Almighty, for the numerous opportunities I have been blessed with.

I am grateful to the countless individuals who have contributed to my PhD journey. Their guidance and support have been essential to my success and made completing this dissertation possible. I sincerely appreciate everyone who supported me along this academic path, including those whose names I might not mention but whose impact has not gone unnoticed.

I want to give special thanks to my supervisor, Chris Biemann. You provided me with the fantastic opportunity to join the LT-Group, where I met so many friendly, disciplined, and collaborative colleagues. Your steadfast mentoring and encouragement have played a huge role in my learning and growth. Beyond your academic support, I will never forget how you helped fund my stays in Hamburg. You were always there for me, cheering on my successes and helping me through tough times.

I am equally grateful to my co-supervisor, Seid Muhie Yimam. Your continuous supervision has been absolutely invaluable, and your guidance throughout every stage of my dissertation has been nothing short of remarkable. You have always been available whenever I needed assistance, offering support for both my academic pursuits and social challenges. You truly are a man of solutions, and I can honestly say that you helped me build my foundation from the ground up. I am also grateful to my external co-supervisor Dr. Tesfa Tegegne Asfaw for his unreserved supervision.

I would also like to acknowledge Katrin and Adeline for their consistent support. Thank you both for being there to help me navigate various administrative matters and ensuring that everything ran smoothly during my studies.

I am very grateful to my LT colleagues. Your unwavering support has made a significant difference in my life, both academically and socially. I cherish the friendships I have formed within this community, as you exemplify kindness, support, and empathy. A special thank you goes to Steffan; I truly appreciate your support in translating the abstract into German and your contribution of the LaTeX dissertation template.

I want to take a moment to thank my friends who have been there for me, offering their encouragement and support throughout my study: Ebrahim Checkol, Mohammed Seid, Alelign Desalew, Birhanu Jibril, and Dereje Berihun. I would especially like to acknowledge Tilahun Abdissa, Birhanu Hailu, Adem Chanie, Abdu Mohammed, Abdulkerim Mohammed, and Rawda Assefa for their help in proofreading my dissertation draft, encouraging and supporting me a lot. Your support and encouragement truly means a lot to me! Tilish, you are a lot to me.

Lastly, I would like to express my profound gratitude to my family, who have been my pillars of strength throughout my study. Your support and encouragement have significantly impacted my journey. Without your understanding, I would not have been able to reach this stage of my academic endeavor and complete my study. Thank you to all my family members for standing by my side and believing in me throughout this journey.

Use of Third-Party Software

This dissertation is written using the LaTeX template, which is available on https://github.com/uhh-lt/thesis-template-uhh-lt-latex. The pictures are drawn using the free version Miro and Microsoft PowerPoint. Microsoft Excel is also used to create different charts. ChatGPT and the free version Grammarly are used to correct grammatical errors and rephrase sentences to improve readability. All of my publications included in this dissertation are written using the free version Overleaf template. Moreover, I customized WebAnno, POTATO and Yandex Toloka annotation tools to create the datasets utilized in this dissertation. The German version of the abstract was translated using Google Translate from the English abstract and later corrected by my German-speaking colleagues.

No one is born hating another person because of the colour of his skin, or his background, or his religion. People must learn to hate, and if they can learn to hate, they can be taught to love, for love comes more naturally to the human heart than its opposite.

— Nelson Mandela (1994)

Abstract

This dissertation addresses the pressing issue of hate speech in the digital age, particularly within the ever-evolving landscape of social media. Over recent decades, social media platforms have become fertile grounds for debates on various social and political issues. However, this openness has also facilitated the alarming proliferation of hateful and discriminatory messages.

Addressing hate speech requires the involvement of both automatic methods, which utilize natural language processing and machine learning models, and non-automatic methods, which involve human judgment and policy regulations. The complex and subjective nature of hate speech makes reliance on legal regulations alone insufficient, necessitating the development of AI-based linguistic and critical discourse analyses to effectively manage and mitigate these challenges. This dissertation primarily focuses on automatic approaches for detecting and analyzing hate speech, specifically within Ethiopia's dynamic social, political, and cultural context, where the challenges are especially pronounced.

The dissertation is structured to provide a comprehensive approach for detecting and analyzing hate speech by designing and implementing five main components: detection, target identification, intensity determination, multimodal detection, and detoxification. We present several datasets that have been compiled using various data annotation methods, aimed at mitigating the impact of hate speech. It also presents novel data sampling strategies and preprocessing pipelines, which addresses data imbalance problems in hate speech studies.

One way to obtain these datasets is crowdsourcing. We explore the viability of crowdsourcing as a method for annotating hate speech data and present hate speech datasets for Amharic, a low-resource language, and French, a high-resource language, utilizing the Yandex Toloka platform. With regard to crowdsourcing, our findings highlight the opportunities and challenges of crowdsourcing for different languages and emphasize the need for careful quality control mechanisms. Additionally, crowdsourcing poses greater challenges for low-resource languages due to the scarcity of crowd workers, which increases the likelihood of malicious users participating in the annotation task.

Another approach for data annotation is to employ an in-house annotation setup, which ensures the creation of high-quality annotated hate speech datasets for Amharic under a controlled setup. We present a dataset of 15.1k tweets, annotated using WebAnno, showcasing the benefits of controlled annotation environments over crowdsourcing by achieving higher inter-annotator agreement.

The insights from both annotation strategies, crowdsourcing and in-house, highlight that hate speech annotation is highly subjective, requiring diverse contextual background information. This addresses our first research question: *What are the main challenges in crowdsourcing and in-house hate speech annotation approaches?*

We introduce a new multidimensional dataset, specifically focusing on three tasks: category classification, target community identification, and intensity ratings. The findings highlight how hate speech in Ethiopia often targets ethnic and political identities, which reflects the complex socio-political dynamics of Ethiopia, addressing our second research question: *To what extent do hate speech disproportionately target specific vulnerable communities?* Additionally, experimental results illustrate how hate and offensive speech manifests itself as continuous values, rather than discrete, binary categories, requiring a more comprehensive and in-depth analysis of intensities while studying hate and offensive speech. These findings address our third research question: *How can hate and offensive speech be understood: as distinct categories or as values on a spectrum of varying intensities?*

The dissertation expands the research into a multimodal analysis by examining hate speech in Amharic memes, which aims to address the fourth research question: *To what extent do multimodality enhance the detection of hate speech compared to unimodal approaches?* The findings from multimodal experiments highlight the superiority of multimodal models over unimodal approaches.

Detecting hate speech, identifying the targets, and assessing its intensity can help content moderators to remove harmful messages from social media platforms. However, these measures alone are insufficient to address online abuse in a broader context. Thus, rewriting toxic content into a non-toxic form provides additional opportunities to enhance online safety. To this end, we introduce the first parallel dataset for Amharic, containing toxic textual inputs and their non-toxic counterparts, generated using text rewriting and rephrasing techniques. We investigate methods for rephrasing toxic content into more neutral language, highlighting the challenges large language models (LLMs) like GPT-4 encounter due to issues with inaccurate and incoherent outputs, addressing the fifth research question: *What challenges do large language models (LLMs) face in Amharic text detoxification task?*

In summary, this dissertation makes significant contributions by developing comprehensive datasets and methodologies to mitigate the pressing issue of online hate speech, within a low-resource languages context. It offers novel insights into the complex nature of hate speech, spanning detection, categorization, intensity prediction, and text detoxification efforts in the context of Ethiopia, a country having diverse social, political and cultural complexities. These contributions pave the way for future research and technological advancements in creating safer online environments, advocating a multi-faceted approach to combating online hate speech in situations where cultural and linguistic diversities are prominent.

Zusammenfassung

Diese Dissertation befasst sich mit dem Problem "Hassreden im digitalen Zeitalter“, insbesondere in der sich ständig weiterentwickelnden Landschaft der sozialen Medien. In den letzten Jahrzehnten sind die Plattformen der sozialen Medien zu einem fruchtbaren Boden für Debatten über verschiedene soziale und politische Themen geworden. Diese Offenheit hat jedoch auch die alarmierende Verbreitung von hasserfüllten und diskriminierenden Botschaften begünstigt.

Der Umgang mit Hassreden erfordert sowohl automatische Methoden, die natürliche Sprachverarbeitung und Modelle des maschinellen Lernens nutzen, als auch nichtautomatische Methoden, die menschliches Urteilsvermögen und politische Regelungen einbeziehen. Aufgrund der komplexen und subjektiven Natur von Hassreden sind gesetzliche Regelungen allein nicht ausreichend, so dass die Entwicklung von KI-basierten linguistischen und kritischen Diskursanalysen erforderlich ist, um diese Herausforderungen effektiv zu bewältigen und zu entschärfen. Diese Dissertation konzentriert sich in erster Linie auf automatische Ansätze zur Erkennung und Analyse von Hassreden, insbesondere im dynamischen sozialen, politischen und kulturellen Kontext Äthiopiens, wo die Herausforderungen besonders ausgeprägt sind.

Die Dissertation ist so strukturiert, dass sie einen umfassenden Ansatz zum Verständnis von Hassreden durch vier verschiedene Analyseebenen bietet: Erkennung, Zielidentifizierung, Intensitätsbewertung und Detoxifizierung. Wir stellen mehrere Datensätze vor, die mit verschiedenen Methoden der Datenannotation zusammengestellt wurden, um die Auswirkungen von Hassreden zu mildern. Darüber hinaus werden neuartige Strategien zur Datenerfassung und Datenvorverarbeitung vorgestellt, die sich mit Problemen des Datenungleichgewichts bei Studien zu Hassreden befassen.

Eine Möglichkeit, diese Datensätze zu erhalten, ist Crowdsourcing. Wir untersuchen die Durchführbarkeit von Crowdsourcing als Methode zur Annotation von Hassrededaten und präsentieren Datensätze für Amharisch, eine Sprache mit geringen Ressourcen, und Französisch, eine Sprache mit vielen Ressourcen, unter Verwendung der Yandex-Toloka-Plattform. Im Hinblick auf Crowdsourcing zeigen unsere Ergebnisse die Möglichkeiten und Herausforderungen von Crowdsourcing für verschiedene Sprachen auf und betonen die Notwendigkeit sorgfältiger Qualitätskontrollmechanismen. Darüber hinaus stellt Crowdsourcing für Sprachen mit geringen Ressourcen eine größere Herausforderung dar, da es nur wenige Crowdworker gibt, was die Wahrscheinlichkeit erhöht, dass böswillige Benutzer an dem Annotationsprojekt teilnehmen.

Ein weiterer Ansatz für die Annotation von Daten ist die Verwendung einer internen Annotationsumgebung, die die Erstellung hochwertiger annotierter Hassreden-Datensätze für Amharisch unter kontrollierten Bedingungen gewährleistet. Wir präsentieren einen Datensatz von 15,1 tausend Tweets, die mit WebAnno annotiert wurden, und zeigen die Vorteile von kontrollierten Annotationsumgebungen gegenüber Crowdsourcing, indem wir eine höhere Übereinstimmung zwischen den Annotatoren erreichen.

Die Erkenntnisse aus beiden Annotationsstrategien, Crowdsourcing und Inhouse, machen deutlich, dass die Annotation von Hassreden sehr subjektiv ist und verschiedene kontextbezogene Hintergrundinformationen erfordert. Daraus ergibt sich unsere erste Forschungsfrage: *Was stellt die größten Herausforderungen in Crowdsourcing- und Inhouse-Annotationsprojekten von Hassreden dar?*

Wir stellen einen neuen multidimensionalen Datensatz vor, der sich speziell auf drei Aufgaben konzentriert: Klassifizierung von Kategorien, Identifizierung der Zielgemeinschaft und Bewertung der Intensität. Die Ergebnisse zeigen, dass Hassreden in Äthiopien häufig auf ethnische und politische Identitäten abzielen, was die komplexe soziopolitische Dynamik Äthiopiens widerspiegelt und unsere zweite Forschungsfrage beantwortet: *Inwieweit richten sich Hassreden unverhältnismäßig stark gegen bestimmte gefährdete Gemeinschaften?* Darüber hinaus veranschaulichen die experimentellen Ergebnisse, wie sich Hass und beleidigende Äußerungen als kontinuierliche Werte und nicht als diskrete, binäre Kategorien manifestieren, was eine umfassendere und tiefgreifendere Analyse der Intensität bei der Untersuchung von Hass und beleidigenden Äußerungen erfordert. Diese Ergebnisse gehen auf unsere dritte Forschungsfrage ein: *Wie können Hass und beleidigende Äußerungen verstanden werden: als unterschiedliche Kategorien oder als Werte auf einem Spektrum unterschiedlicher Intensität?*

In der Dissertation wird die Forschung auf eine multimodale Analyse ausgeweitet, indem Hassrede in amharischen Memes untersucht wird, um die vierte Forschungsfrage zu beantworten: *Inwieweit verbessert Multimodalität die Erkennung von Hassrede im Vergleich zu unimodalen Ansätzen?* Die Ergebnisse der multimodalen Experimente unterstreichen die Überlegenheit der multimodalen Modelle gegenüber unimodalen Ansätzen.

Die Erkennung von Hassreden, die Identifizierung der Ziele und die Bewertung ihrer Intensität können Moderatoren dabei helfen, schädliche Nachrichten von Social-Media-Plattformen zu entfernen. Diese Maßnahmen allein reichen jedoch nicht aus, um Online-Missbrauch in einem breiteren Kontext zu bekämpfen. Daher bietet die Umformung toxischer Inhalte in eine nicht-toxische Form zusätzliche Möglichkeiten zur Verbesserung der Online-Sicherheit. Zu diesem Zweck stellen wir den ersten parallelen Datensatz für Amharisch vor, der toxische Texteingaben und ihre ungiftigen Gegenstücke enthält, die mithilfe von Techniken zum Paraphrasieren von Texten erzeugt wurden. Wir untersuchen Methoden zur Umformulierung toxischer Inhalte in eine neutralere Sprache und heben die Herausforderungen hervor, denen sich große Sprachmodelle (LLMs) wie GPT-4 aufgrund von Halluzinationen gegenübersehen, um die fünfte Forschungsfrage zu beantworten: *Welche Herausforderungen stellen sich großen Sprachmodellen (LLMs) bei der Detoxifikation von amharischen Texten?*

Zusammenfassend leistet diese Dissertation durch die Entwicklung umfassender Datensätze und Methoden zur Detoxifizierung von Online-Hassrede in ressourcenarmen Sprachen einen wichtigen Beitrag. Sie bietet neue Einblicke in die komplexe Natur von Hassreden, die die Erkennung, Kategorisierung, Intensitätsbewertung und Detoxifizierung von Texten in Äthiopien, einem Land mit vielfältigen sozialen, politischen und kulturellen Gegebenheiten, umfassen. Diese Beiträge ebnen den Weg für künftige Forschung und technologische Fortschritte bei der Schaffung sicherer Online-Umgebungen und befürworten einen vielschichtigen Ansatz zur Bekämpfung von Online-Hassreden in Situationen, in denen kulturelle und sprachliche Unterschiede im Vordergrund stehen.

Contents

Appendices

List of Figures

List of Tables

List of Abbreviations

5Js	5 Consecutive Junes (2018-2022)
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bi-directional Long Short Term Memory
CBOW	Contentious Bag of Words
CHrF	Character F-score
CNN	Convolutional Neural Network
FDRE	Federal Democratic Republic of Ethiopia
FLAIR	Framework for state-of-the-art NLP
GPT	Generative Pre-trained Transformer
IAA	Inter Annotator Agreement
J-Score	Joint Score
LLaMA	Large Language Model Meta AI
LLM	Large Language Model
LM	Language Model
LR	Logistic Regression
LSTM	Long Short Term Memory
NB	Naïve Bayes
NLP	Natural Language Processing
OCR	Optical Character Recognition
Potato	POrtable Text Annotation TOol
ResNet	Residual Network
RNN	Recurrent Neural Network
RoBERTa	Robustly optimized BERT
SHAP	SHapley Additive exPlanations
SIM	Content Similarity
STA	Style Transfer Accuracy
SVM	Support Vector Machine
TF-IDF	Term Frequency Inverse Document Frequency
Vgg	Visual Geometry Group

"Given the complexities of modern society, both domestically and internationally, the development of sophisticated tools in the computational social sciences will assume increasing importance."

— Kim L. Boyer

1 Social NLP in Ethiopian Languages

Contents

1.1 Introduction

The persistent pace of advancements in digitization, accompanied by the pervasive integration of large language models into various aspects of technology, has triggered a significant and complex transformation in the daily routines of individuals and communities (Griffin et al., 2023; Choi et al., 2023). This transformation spans through diversified domains of everyday life such as interpersonal or business communications to national or global policy decisions of government bodies. While individuals and communities are immersing themselves into the digital world shaped by transformative technological influences, huge volumes of data are generated through the interactions. This poses practical challenges to the research community, which requires complex and deeper explorations to unveil the hidden knowledge from such large data collections.

Currently, researchers are actively engaged in exploring the voluminous data with strong passion and dedication through the sophisticated applications of natural language processing (NLP) and machine learning methodologies. The surge in researchers' interest and commitment further emphasizes the dynamic and evolving relationship between technology and human language (Nityasya et al., 2023; Griffin et al., 2023).

More specifically, the steady increase in the number of online communities together with the emergence of various social media platforms, online forums, and blogs have intensified the amount of data which is generated on daily bases. The need to explore the ever increasing social media data in the contemporary lives of people highlights the growing importance of NLP applications such as sentiment analysis and hate speech detection task (Reuver et al., 2021).

1.2 Motivation of the Study

These days, the prevalence and influence of social media platforms are constantly expanding and easily reaching the global community. This growth occurs concurrently with the proliferation of a diverse spectrum of online content crafted by a multitude of contributors on various topics, which is readily available for consumption and active engagement by online users worldwide (Sazzed, 2023). Reports indicate, as of 2024, the number of active social media users has surpassed 5.2 billion people globally, constituting approximately 62% of the world's population. Besides, the average annual growth rate of individuals who are highly engaged in social media platforms has also exceeded a 5%, demonstrating a significant increase in their participation over time (Kemp, 2024).

Due to its diverse ranges of applications and its ability to facilitate communication, networking, information sharing, entertainment, business, and marketing, social media has become increasingly important for people in modern society, and it has become an integral part of daily life (Sazzed, 2023).

Social media platforms have created environments that foster the emergence of social movements, offering fertile ground for individuals to unite themselves towards a shared goal. The utilization of social media platforms has not only revolutionized the traditional methods of organizing social movements but also has fundamentally transformed the way people come together, plan and execute their shared goals, easily giving rise to the establishment of a multitude of diverse social movements (Rogers et al., 2019). The platforms have granted social movements abundant potential to innovate and develop novel methods of organizing protests which function within the digital domain and exhibit significant levels of engagement from people that go beyond traditional social movement organizational frameworks (della Porta and Diani, 2015).

The Arab Spring, which began in Tunisia after Mohamed Bouazizi, a young Tunisian street vendor, set himself on fire in front of a municipal office to protest against the government due to his ill treatment by local officials on December 17, 2010 can be a notable example. The event provoked the people of Tunisia to protest against the government, which eventually led to the overthrow of the Tunisian government in January 2011 and brought regime change within a month. The protests rapidly spread even to other countries such as Egypt and Yemen in January 2011, Bahrain and Libya in February 2011, and Syria in March 2011. These events are impressive testaments to the significant roles played by social media platforms in catalyzing extensive social movements and protests throughout North Africa and the Middle East (Rogers et al., 2019).

Similarly, in Ethiopia, movements on social media started in late 2011. The Ethiopian muslims protest against government involvement in the internal matters of the *Mejlis*, the Muslims religious administration organization, marked the beginning of the nonviolent demonstration with a moto of "*Let Our Voices be Heard!*", which occurred from 2011-2015. The protest happened every Friday after Juma'a prayer throughout the country in general and the grand *Anwar Mosque* in particular (Omar, 2020). The well known hidden social movement, "*Let Our Voices be Heard!*", organized the demonstrations on social media, which usually took place in the Mosque campus.

The other notable movements which stared in 2016 include the *Oromo Qeeroo* and the *Amhara Fanno* youth movements against government power abuse, maladministration, and corruption, which eventually changed to sever popular protests that bring a regime change in the Ethiopia's political history (Forsén and Tronvoll, 2021). Social media has emerged as the primary arena for orchestrating these social movements and protests, thereby exerting significant influences on the socio-political landscape of the nation (Abraha, 2017). Such protests have still continued to spread in social media, where its consequences characterized by the amplification of previously marginalized voices and intensified conflicts that have left a lasting scar and caused irreversible harm that can never be forgotten by many Ethiopians (Abraha, 2017; **Ayele**, Dinter, et al., 2022).

While social media has been used to bring political changes, at the same time online users are actively utilizing social media platforms as instrumental tools for generating and disseminating hate speech on the web (Bran and Hulin, 2023; Mathew et al., 2021; Davidson et al., 2017; Waseem and Hovy, 2016; **Ayele**, Yimam, et al., 2023). The ease of communication and the global reach of these platforms have enabled users to spread hateful and offensive content aggressively in wider circles across online communities (Zufall et al., 2022). The anonymity provided by social media platforms has allowed propagators of hateful messages to craft and dispatch harmful content while concealing their identities behind digital screens (Bran and Hulin, 2023; Kiritchenko et al., 2021; Zufall et al., 2022).

Online hate speech, which is the focus of this dissertation, can have real-world consequences, contributing to social divisions, fueling hostility, and incitement of violence (Abraha, 2017). Social media companies, policymakers, and researchers are increasing their emphasis on developing strategies to detect, combat, and mitigate the impact of hate speech on these platforms without compromising the principles of freedom of speech, user safety and privacy (Pavlopoulos et al., 2017; Ye et al., 2023). For the past couples of years, there has been increasing attention and interest in exploring hate speech among researchers from diverse academic disciplines, including social science, psychology, medicine, communication studies, and computer science (Tontodimamma et al., 2021; Davidson et al., 2017; Mathew et al., 2021; Davidson et al., 2019; Chekol et al., 2023; **Ayele**, Yimam, et al., 2023).

The drive behind conducting this dissertation on hate speech stems from our firm dedication to examine the complex dimensions of the topic. Our motivation emphasizes five key areas of importance: enhancing digital societal well-being, safeguarding vulnerable communities, advocating for digital inclusiveness, leveraging technological innovations, and creating accessible resources to address hate speech on social media platforms.

- **Enhancing digital societal well-being.** Mitigating online hate speech improves the digital societal well-being of individuals by cultivating a healthier digital ecosystem, which encourages them to engage in constructive dialogue, promotes positive user experience, and promotes collaboration among people.

- **Safeguarding vulnerable communities**. Our motivation extends to the protection of vulnerable groups, as hate speech disproportionately affects marginalized communities. By conducting research in this area, there is an aspiration to develop effective mechanisms to shield these groups from online discrimination and harm. Additionally, legal compliance plays a pivotal role, with the research seeking to support and reinforce existing regulations against hate speech, ensuring a legal framework that aligns with societal values.

- **Advocating for digital inclusiveness**. Mitigating online hate speech promotes digital inclusiveness by encouraging positive digital interactions and enhancing online environments that respect diverse perspectives. In essence, the multi-faceted motivation for hate speech detection research reflects a comprehensive commitment to societal welfare, online safety, the protection of vulnerable communities, legal adherence, ethical technology development, and the advancement of inclusiveness in the digital landscape.

- **Leveraging technological innovations**. Our motivation to develop and harness models as technological solutions in combating online hate speech contributes to safeguarding freedom of expression. This includes the development of sophisticated tools that are capable of identifying and addressing hate speech without imposing undue restrictions on the expression of ideas and opinions.

- **Creating accessible resources**. The development of accessible hate speech resources, including datasets, guidelines, models, source codes, and associated tools for low-resource languages such as Amharic, serves to advance research and address problems on social media. This contribution helps to alleviate the digital gap and linguistic disparities within the domain of advanced large language models.

The aforementioned five key areas of importance that drive our motivation are achieved through designing and implementing the hate speech detection and analysis components indicated in Figure 1.1. These components are:

- **Hate Speech Detection**. This component is designed to detect the presence of hate speech and classify it into categories, such as hate, offensive and normal.

- **Hatred Target Identification**. Once we detect the presence of hate speech, it becomes necessary to investigate which portion of the population is specifically targeted and affected by hate speech propagators on social media. Thus, this component is specifically designed to analyze how hate speech disproportionately targets specific communities.

- **Intensity Determination**. This component determines the intensities of hatefulness and offensiveness in social media texts using datasets annotated with a Likert rating scale. It explores the variation in intensity within hateful and offensive texts across a continuum.

- **Multimodal Hate Speech Detection**. The previous three components mainly relay on textual data while hate speech on social media manifests itself in multiple modalities, increases the difficulty level of detecting hate speech. Thus, the multimodal component detects hate speech from Amharic memes, utilizing both the image and textual features.

- **Detoxification**. This component focuses on rewriting or rephrasing toxic messages into a more neutral form, which can offer additional opportunities to content moderators in ensuring a peaceful and inclusive social media environment.

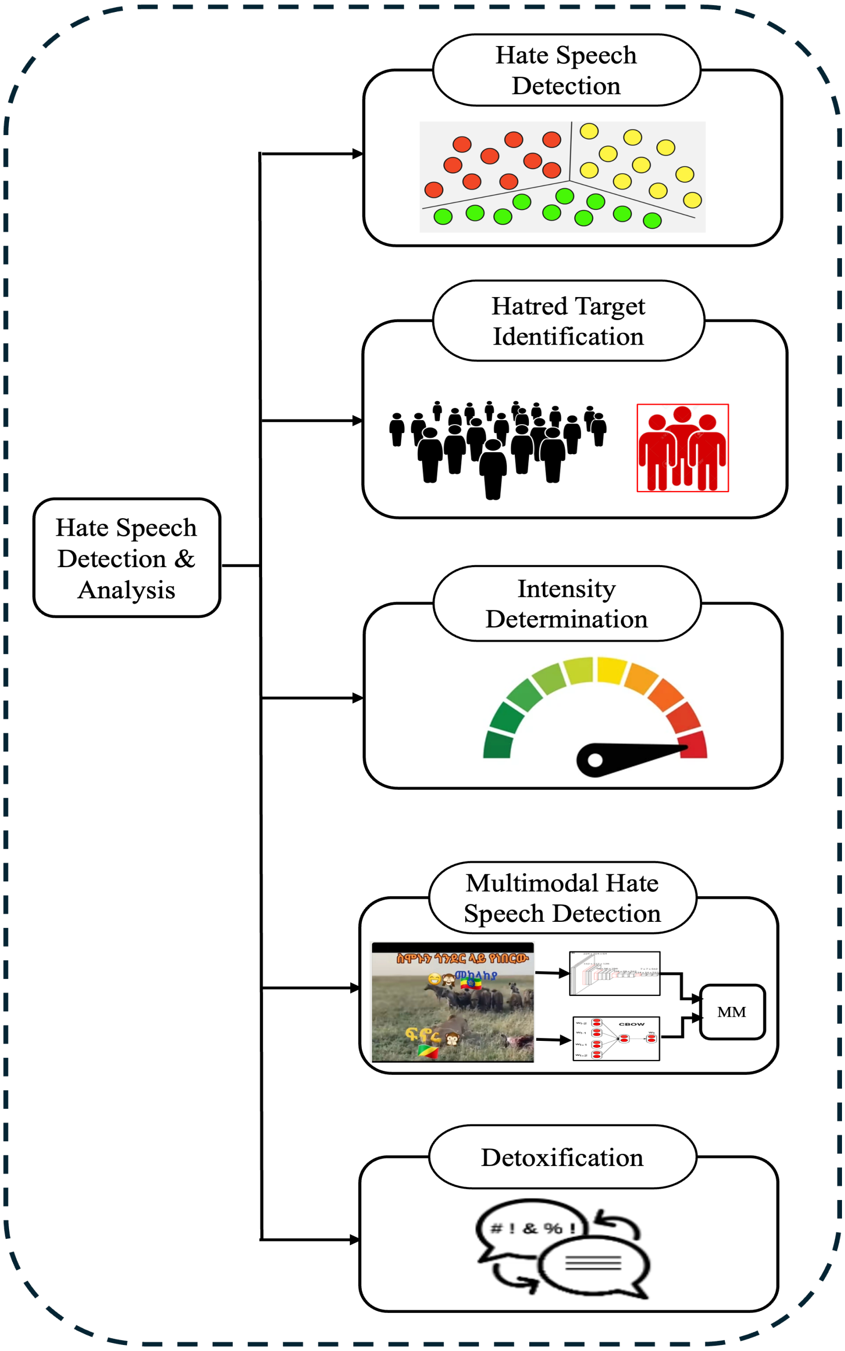

Hate Speech Detection & Analysis
Hate Speech Detection
Hatred Target Identification
Intensity Determination
Multimodal Hate Speech Detection
MM
Detoxification
# ! & % !

Figure 1.1: Building blocks of the dissertation.

Sections 1.3 presents a brief overview of Natural Language Processing (NLP) and its applications in the contemporary digital world. Besides, before delving into an in-depth study of hate speech analysis, we assess and briefly present the current status of social NLP tasks such as, sentiment analysis, in low-resource languages like Amharic.

1.3 The Status of Social NLP in Ethiopian Languages

These days, natural language processing is offering a lot of applications for people spanning across diverse domains to enhance communication, accessibility and improve performance on their daily routines. NLP provides people with assistive technologies that can improve their daily lives such as contextualized information retrieval services, language translation, text-to-speech and speech-to-text systems, conversational chatbots, healthcare insights, personalized education, social media analysis, and many more applications thereby revolutionizing the various aspects of human interaction (Montejo-Ráez and Jiménez-Zafra, 2022; Hovy and Yang, 2021).

Access to the advantages offered by NLP applications is often severely limited or entirely unavailable to the speakers of low-resource languages that are mainly found in developing nations with poor digital infrastructures (Röttger, Nozza, et al., 2022). This disparity arises primarily from the notable absence of crucial linguistic resources which are necessary for designing and implementing appropriate NLP services for these languages. The challenges of NLP in low-resource languages predominantly stems from the scarcity of available resources such as free text corpora, annotated datasets, lexicon entries, and adequately trained models (Hedderich et al., 2021). This scarcity of foundational resources significantly hinders the development and deployment of effective NLP tools tailored to low-resource languages. Consequently, these challenges aggravate digital disparities, linguistic inequalities, and further marginalizes speakers of the languages in the spheres of technology in general and human language technologies in particular (Tonja et al., 2023).

Among the main challenges to conduct natural language processing research for low-resource languages such as Amharic, is resource scarcity (i.e. datasets and developed NLP tools). The absence of a sizable and properly annotated research corpus for various natural language processing tasks is one of the main challenges for Amharic natural language processing tasks (Gezmu et al., 2017; Yimam et al., 2021). The lack of well designed and developed natural language processing tools and applications such as annotation tools and classification models, is also a significant bottleneck to conduct NLP related research in Amharic (Gezmu, Seyoum, et al., 2018; Yimam et al., 2021). Due to the limited availability of resources, including linguistic datasets, tools, and research support, the Amharic language is still categorized as one of the low-resource languages in Sub-Saharan Africa.

Another critical challenge in Amharic natural language processing research is attributed to the morphological complexity of the language (Mulugeta et al., 2012). According to Gezmu, Nürnberger, et al. (2018) and Yimam et al. (2021), the Amharic language poses several morphological challenges, which include variations in orthography, compound word formations, and the existence of homographs. These factors collectively contribute to the complex nature of its linguistic structures.

In the broader area of natural language processing, social NLP typically focuses on analyzing and understanding natural languages in the context of social, cultural, economical and political interactions among people in digital environments (Hovy and Spruit, 2016). Social NLP encompasses the utilization of NLP methodologies to explore the data generated from social media platforms, online forums, and blogs, covering a wide range of topics that reflect the diverse contexts of social interactions among individuals and communities in their day-to-day experiences (Hovy and Yang, 2021). Its primary emphasis lies in understanding the language patterns used in social contexts and extracting significant insights pertaining to the dynamic attributes of digital communications and interaction among people. Presently, social media platforms are generating vast volumes of data daily, necessitating thorough exploration and comprehension of these interactions by leveraging emerging Artificial Intelligence (AI) applications in both natural language processing and social sciences disciplines to gain a deeper understanding of social phenomena (Del Tredici et al., 2019; Hovy and Yang, 2021).

In contemporary times, there is a strong interest among researchers and developers in exploring novel approaches and gaining deeper insights to fully leverage the capabilities of natural language processing with data gathered from online social media platforms and digital forums (OpenAI, 2024). The interplay between evolving social dynamics, diverse digital landscapes, and advancing AI technologies significantly fuels the ongoing evolution of the social NLP domain. This evolution is driven by both advancements in NLP methodologies and the continuously shifting dynamics of communication and interaction mechanisms among individuals in cyberspace.

To sum up, the main emphasis in social NLP lies in enhancing contextual comprehension during conversations among individuals, considering their diverse cultural, social, political, and economic backgrounds (Hovy and Spruit, 2016). It involves the analysis of emotions, sentiments, hate speech, abusive content, fake news, and related subjects, primarily derived from data generated in social media platforms (Del Tredici et al., 2019; Hovy and Yang, 2021).

This briefly provides an overview of some social NLP topics, particularly offering a brief assessment of sentiment analysis tasks in Section 1.3.1, with a primary focus on hate speech in low-resource Ethiopian languages, such as Amharic, in subsequent chapters and sections.

1.3.1 Sentiment Analysis

Social media and sentiment analysis have a mutually beneficial connection. While social media platforms offer rich sources for sentiment analysis, sentiment analysis enables the extraction of valuable insights into the emotions and opinions conveyed within social media content (Kenyon-Dean et al., 2018). Sentiment analysis is a powerful natural language processing tool that has a broad range of applications across various fields and domains where understanding and responding to human sentiment is essential to achieve better success in particular contexts. Sentiment analysis is also a valuable tool for data-driven decision-making across diverse domains of studies (Roccabruna et al., 2022; Tabari et al., 2017; Kenyon-Dean et al., 2018). These domains including but not limited to business research, financial analysis, governance and politics, healthcare, education, and social media monitoring (Roccabruna et al., 2022). The following examples can showcase some of the application domains:

- **Businesses research**: sentiment analysis is utilized to monitor brand sentiments and marketing campaigns, analyze customer feedback and improve service quality, understand consumer sentiment towards products, brands, distribution channels (Carlos and Yalamanchi, 2012). It is also used to improve brand strategy, advertising, product development and track emerging trends. It helps to mitigate potential reputation crises and risks by analyzing sentiment in news articles, social media discussions, and other sources to mitigate risks and safeguard organizational interests (Hovy and Yang, 2021; Hovy and Spruit, 2016).

- **Financial analysis**: financial analysis assesses a company's financial health using tools like income statements, balance sheets, and cash flow statements. It also provides insights into the performance and value of a company or investment opportunity, enabling stakeholders to make well-informed decisions regarding investments, financing, and strategic planning (Tabari et al., 2017).

- **Governance and politics**: analyzing sentiments in political speeches, news articles, and social media discussions helps governments to understand public opinion, predict election outcomes, and guide campaign strategies, inform policy decisions, and foster citizen engagements. Policymakers can tailor communication strategies through enhancing the government's transparency and responsiveness to contribute inclusive democratic processes (Sanders and van den Bosch, 2020).

- **Healthcare**: Sentiment analysis in healthcare entails analyzing patient feedback, reviews, and social media discussions to gauge satisfaction levels, identify concerns, and improve service quality. It aids healthcare providers in understanding patient sentiments towards services, treatments, and facilities, enabling them to tailor care delivery and enhance patient experience (Bobicev and Sokolova, 2018). By analyzing sentiments, healthcare organizations can identify trends, monitor public health perceptions, and address issues promptly, thereby improving overall patient satisfaction and healthcare outcomes (Yadav et al., 2018). Additionally, sentiment analysis assists in evaluating healthcare interventions, tracking the effectiveness of communication campaigns, and identifying areas for improvement in healthcare policies and practices (Bobicev and Sokolova, 2018).

- **Education**: Educational institutions benefit from analyzing student feedback to improve courses and programs (Rakhmanov and Schlippe, 2022). Sentiment analysis in education involves analyzing student feedback to understand satisfaction levels and identify areas for improvement, aiding educators in enhancing teaching methods and student engagement. This approach enables educational institutions to monitor trends, assess program effectiveness, and address issues promptly, ultimately improving overall student satisfaction and academic outcomes (Hussiny and Øvrelid, 2023). During Covid-19 pandemic, many studies that utilized sentiment analysis have been conducted to analyze the perspectives of students, instructors, and families regarding the implications of the pandemic on their academic lives (Yıldırım et al., 2023; Kocaçınar et al., 2023).

In general, sentiment analysis serves as a valuable tool for facilitating a more informed and effective decision-making processes across various sectors through providing insights into the emotions and opinions existed within diverse datasets generated in those sectors, which offer data-driven decision-making capability (Roccabruna et al., 2022).

However, the task of sentiment analysis is less researched in the low-resource Ethiopian languages to utilize the benefits across sectors due to the previously mentioned reasons in Section 1.3.

In (Tonja et al., 2023), we assessed the status of sentiment analysis task in Ethiopian languages and organized the available resources such as datasets, source codes and models in to a common publicly accessible GitHub repository[1].

Languages	Author(s)	Size	Algorithm	Score	Dataset	Model
Amharic	Yimam et al. (2020)	9,400	F-Role2Vec	58.48	Yes	Yes
	Philemon and Mulugeta (2014)	600	Naïve Bayes	51.00	No	No
	Abeje et al. (2022)	2,000	LSTM	90.10 (acc)	Yes	No
	Alemneh et al. (2020)	30,000	hybrid	98.00(acc)	No	No
Oromo	Oljira (2020)	3000	Naive Bayes	93.00	No	No
	Rase (2020)	1,452	LSTM	87.70	No	No
	Wayessa and Abas (2020)	1,810	SVM	90.00	No	No
	Yadesa et al. (2020)	341	dictionary	86.10	No	No
Tigrinya	Tela (2020)	4,000	XLNet	81.62	No	No

Table 1.1: Summary of related works for selected Ethiopian languages in sentiment analysis tasks, `Size` shows the annotated dataset used during the experiment, `Score` shows the outperformed model results evaluated using F1 score, `Dataset` and `Model` shows the availability of dataset and models in publicly accessible repositories.

Table 1.1 summarizes recent studies on sentiment analysis tasks for selected Ethiopian languages, including Amharic, Oromo, and Tigrinya. The studies utilize various algorithms such as F-Role2Vec, Naïve Bayes, LSTM, SVM, hybrid, and XLNet.

For Amharic, Yimam et al. (2020) achieved the highest F1 score of 58.48% using F-Role2Vec with a dataset and a model publicly available, while Abeje et al. (2022) achieved the highest accuracy of 90.10% using LSTM. Among the sentiment analysis studies in Oromo language, the highest accuracy was achieved by Oljira (2020) using Naïve Bayes with an accuracy of 93.00%, while Rase (2020) achieved 87.70% accuracy using LSTM. Besides, Wayessa and Abas (2020) achieved 90.00% accuracy using SVM. For Tigrinya Language, Tela (2020) is the only available sentiment analysis study before Afrisenti (Muhammad, Abdulmumin, **Ayele**, et al., 2023). Tela (2020) achieved an F1 score of 81.62% using XLNet with a 4000 manually labeled dataset. None of the datasets and models for Oromo, Tigrinya, and most of the works for Amharic are publicly accessible, hence results are not also comparable. This suggests that more work needs to be done in creating publicly accessible datasets and models for sentiment analysis tasks in Ethiopian languages.

In conclusion, our survey studies discussed in Table 1.1 indicate the potential for sentiment analysis in Ethiopian languages. The results show that the models' performance varies depending on the algorithm, dataset, and model availability. As can be seen from Table 1.1, only 2 out of 9 works which is 22% of the studies shared their datasets publicly to promote further research. Only 1 out of the 9 works has publicly released both the datasets and models for future researchers to replicate the study. These findings highly signify the need to create publicly accessible datasets, design annotation guidelines, generate sentiment lexicon, build classification models, and ensure their availability in different applications, and help future researchers to replicate the tasks as well as to improve the performance of models.

[1] https://github.com/EthioNLP/Ethiopian-Language-Survey

	`amh`	`orm`	`tir`	`arq`	`hau`	`ibo`	`ary`	`pcm`	`pt-MZ`	`kin`	`swa`	`tso`	`twi`	`yor`
train	5,985	-	-	1,652	14,173	10,193	5,584	5,122	3,064	3303	1,811	805	3,482	8,523
dev	**1,498**	**397**	**399**	415	2,678	1,842	1,216	1,282	768	828	454	204	389	2,091
test	**2,000**	**2,097**	**2,001**	959	5,304	3,683	2,962	4,155	3,663	1027	749	255	950	4,516
Total	**9,483**	**2,494**	**2,400**	3,062	22,155	15,718	9,762	10,559	7,495	5,158	3,014	1,264	4,821	15,130

Table 1.2: Sizes and splits of the AfriSenti datasets. We do not allocate training splits for Oromo (`orm`) and Tigrinya (`tir`) due to the limited size of the data and only evaluate on them in a zero-shot transfer settings.

In (Muhammad, Abdulmumin, **Ayele**, et al., 2023), we presented the methods of data collection, annotation and the baseline models on the data sets. List of stopwords for each language (Amharic, Tigrinya and Oromo) were utilized to collect the tweets from Twitter. In order to get balanced collections of sentiment classes (positive/negative/neutral), we used a sentiment lexicon—a dictionary of positive and negative words when selecting samples for annotation. While collecting tweets using lists of Amharic words, we encountered tweets written in Tigrinya, which was a result of Amharic-Tigrinya code-mixing. This prompted us to utilize the Pycld2 library[2] for language detection.

Each tweet was annotated by two independent annotators and then curated by a third, more experienced individual, who decided on the final gold labels for three languages: Amharic, Tigrinya, and Oromo. The free marginal multi-rater (Randolph, 2005) was employed to compute the kappa agreement scores. The kappa scores of 0.47, 0.51 and 0.20 were achieved for Amharic, Tigrinya, and Oromo, respectively. Despite moderate agreements were achieved for Amharic and Tigrinya, we obtained a low agreement score for Oromo due annotation challenges such as noisy user generated data and difficulties in dealing with tone, digraphia, and code-switching (Yimam et al., 2020; Adebara et al., 2022). Table 1.2 shows the proportion of train, development and test data splits where our contributions of the three Ethiopian languages are presented in the first three columns (amh, orm, and tir) with bold fonts.

Figure 1.2 presented the distribution of labels (positive/negative/neutral) across all the three Ethiopian languages in the Afrisenti datasets. In the Amharic dataset, the number of tweets labeled as positive is relatively small compared to those labeled as negative or neutral. Meanwhile, in the Tigrinya dataset, negative labeled tweets dominate the other class labels, including neutral and positive classes.

For the baseline experiments, three experimental settings were considered:

- Monolingual baseline models based on multilingual pre-trained language models for 12 African languages with training data,
- Multilingual training of all 12 languages, and their evaluation on a combined test of all 12 languages,
- Zero-shot transfer to Oromo (`orm`) and Tigrinya (`tir`) from any of the 12 languages with available training data.

Table 1.3 illustrates the accuracy scores of monolingual baseline results for the 12 language that have training samples in the *AfriSenti* dataset, excluding Oromo and Tigrinya languages. Our contribution here is the Amharic language datasets and models

[2] https://pypi.org/project/pycld2/

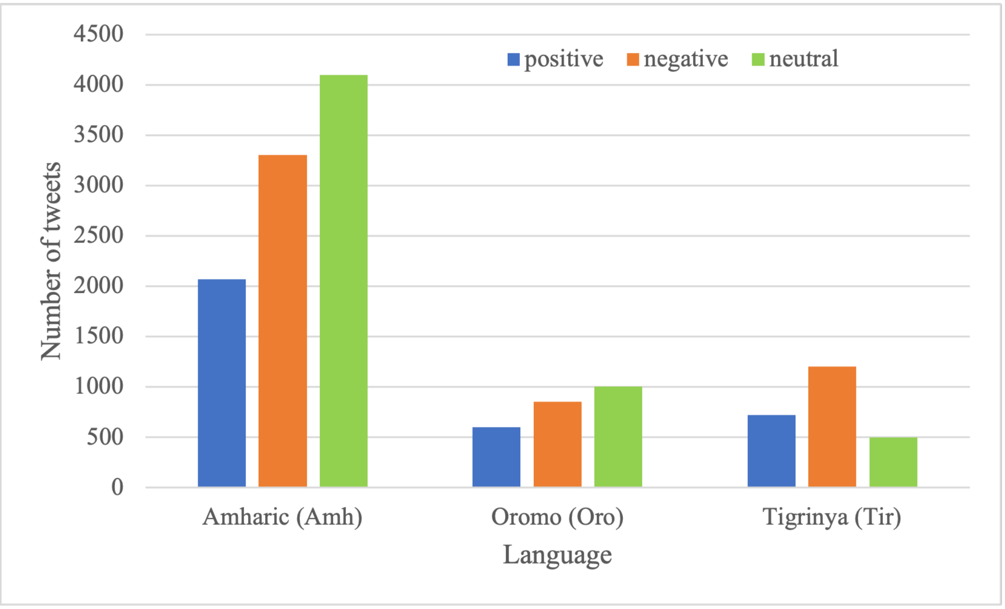


Figure 1.2: Sentiment label (positive/negative/neutral) distributions of the three Ethiopian languages (Amharic, Oromo and Tigrinya) in the *AfriSenti* datasets.

Lang.	AfriBERTa large	XLM-R base	AfroXLMR base	mDeBERTa base	XLM-T base	XLM-R large	AfroXLMR large
amh	56.9	60.2	54.9	57.6	60.8	**61.8**	61.6
arq	47.7	65.9	65.5	65.7	**69.5**	63.9	68.3
ary	44.1	50.9	52.4	55.0	**58.3**	57.7	56.6
hau	78.7	73.2	77.2	75.7	73.3	75.7	**80.7**
ibo	78.6	75.6	76.3	77.5	76.1	76.5	**79.5**
kin	62.7	56.7	67.2	65.5	59.0	55.7	**70.6**
pcm	62.3	63.8	67.6	66.2	66.6	67.2	**68.7**
pt-MZ	58.3	70.1	66.6	68.6	71.3	71.6	**71.6**
swa	61.5	57.8	60.8	59.5	58.4	61.4	**63.4**
tso k	51.6	47.4	45.9	47.4	**53.8**	43.7	47.3
twi	**65.2**	61.4	62.6	63.8	65.1	59.9	64.3
yor	72.9	62.7	70.0	68.4	64.2	62.4	**74.1**
AVG	61.7	61.9	63.9	64.2	64.7	63.1	**67.2**

Table 1.3: Accuracy scores of monolingual baselines for *AfriSenti* on the 12 languages with training splits. Results are averaged over 5 runs.

where XLM-R large model achieved the best accuracy result of 61.8%. AfroXLMR large, XLM-T base and XLM-R base obtained the second, third, and fourth best results in the Amharic dataset, respectively.

Since the datsets presented for Oromo and Tigrinya languages were relatively small, zero-shot cross-lingual transfer learning approach has been utilized. Table 1.4 shows the zeroshot cross-lingual transfer task performances from models trained on different source languages with available training data to the test-only languages Oromo and Tigrinya. While Hausa and Amharic were found to be the best source languages for

	Target Lang.		
Source Lang.	orm	tir	AVG
amh	46.5	62.6	54.6
arq	27.5	56.0	41.8
ary	42.5	58.6	50.6
hau	**47.1**	**68.6**	**57.9**
ibo	41.7	39.8	40.8
kin	43.6	64.8	54.2
pcm	26.7	58.2	42.5
por	28.7	21.5	25.1
swa	36.8	26.7	31.8
tso	21.5	15.8	18.7
twi	9.8	15.6	12.7
yor	39.2	67.1	53.2
multilingual	42.0	66.4	54.2

Table 1.4: Zeroshot evaluation on orm and tir. All SRC LANGs are trained on AfroXLMR-large.

Oromo, Hausa and Yoruba presented to be the best source languages Tigrinya. Hausa even outperformed in the multilingual trained model. The impressive performance for transfer between Hausa and Oromo may be because both are from the same language family and share a similar Latin script. Besides, Hausa has the largest training dataset in *AfriSenti*. Both linguistic similarity and size of source language data have been shown to correlate with successful cross-lingual transfer (YH Lin et al., 2019). However, it is unclear why Yoruba performs particularly well for Tigrinya despite the difference in script. One hypothesis is that Yoruba may be a good source language in general, as shown in (Adelani et al., 2022) where Yoruba is the second best source language for named entity recognition in African languages.

In (Muhammad, Abdulmumin, Yimam, et al., 2023), we organized a Sem-Eval shared task on sentiment analysis for 14 low-resource African languages, for which we contributed datasets for three Ethiopian languages: Amharic, Oromo, and Tigrinya. Most of the teams that participated in the *Monolingual Sentiment Classification Sub-Task* surpassed our *AfriSenti* baseline result for Amharic, with the highest score reaching 78.42% F1. Additionally, the best teams that participated in the *Zero-Shot Sentiment Classification Sub-Task* outperformed our *AfriSenti* baseline of 68.60% with a score of 70.80% for the Tigrinya language. However, our *AfriSenti* baseline of 47.10% outperformed the best team's score of 46.23% for Oromo.

Despite the main focus of this study is to explore hate speech on social media in the Ethiopian context, we have presented the status of sentiment analysis in Ethiopian languages and contributed baseline datasets, guidelines and models as a preliminary task in this dissertation. We still recommend that researchers contribute more datasets and explore the sentiment analysis task from diverse perspectives and applications for Ethiopian low-resource languages.

1.3.2 Hate Speech on Social Media

Hate speech on social media refers to the use of inappropriate languages in online communications and interactions, which can also incite violence. This type of language can take various forms, including insults, threats, harassment, or discriminatory remarks (Rapp, 2021). The prevalence of online hate speech on social media platforms has raised concerns about the well-being of online communities through creating negative psychological effects, fostering online harassment, eroding civil discourse, and damaging reputations (Vidgen et al., 2019). It can also lead to user disengagement, reduce platform trust, disproportionately affect marginalized groups, and in severe cases, it incites violence and conflict that cause property damages and even loss of human lives.

Addressing hate speech on social media requires efforts involving a combination of technological solutions, community engagement, and policy enforcement to create a more respectful and secure digital space (Davidson et al., 2017; Mathew et al., 2021). Technological solutions employ the implementation of sophisticated filtering algorithms, efficient user reporting tools, and utilization of machine learning models, while community engagement encompasses establishment of transparent community guidelines, deployment of human moderation teams, and initiation of educational campaigns (Shen and Rose, 2019). Policy enforcement also includes collaborations with stakeholders, designing and enforcing regulations, and providing overall legal support when necessary (LeGendre et al., 2022).

Due to the multifaceted and subjective nature of hate speech, there exists ongoing and extensive debates about the subject matter between the academia, the industry and other stakeholders (ElSherief et al., 2021). It is a complex phenomenon that is inherently tied to the dynamics between groups and is dependent on linguistic subtleties (Fortuna and Nunes, 2018).

Thus, the definitions of hate speech vary widely across regions and organizations, highlighting its challenges in achieving consensus (Luo et al., 2023). The discussions and debates among researchers, practitioners, and other stakeholders encompass the social, political, cultural, legal, and ethical considerations, which reflects its complex and diverse perspectives. The evolving nature of online communication platforms also adds extra layers of complexity to dealing with the topic of hate speech. Despite these challenges, addressing hate speech requires active engagement from stakeholders in both industry and academia to identify its impacts, shape policies, and foster inclusive discourse (Vidgen et al., 2019).

There is no universally accepted definition of hate speech, mainly because of the above mentioned reasons to determine whether a speech is normal, offensive, or conveys hate. Providing a precise and universally accepted definitions of hate speech by multiple stakeholders such as scholars and practitioners, who may come from diverse fields of studies, cultural and social backgrounds is highly challenging and even impossible (Papcunová et al., 2023; Luo et al., 2023). Fortuna and Nunes (2018) analyzed various definitions of hate speech proposed by associations, scientific community, and social media platforms, and suggested a more refined definition of hate speech for future researchers through exploring and scrutinizing the commonalities and differences in those diverse definitions.

For the purpose of this dissertation, we adapt the definition of hate and offensive speech proposed by Fortuna and Nunes (2018) and Casanovas and Oboler (2018). Thus, we define hate speech as language content that attacks, diminishes, incites violence or hate against groups, based on specific characteristics such as national or ethnic origin, physical appearance, religion, gender identity, disability, political and religious ideologies. Moreover, it can manifest in various linguistic styles, including subtle forms or instances involving humor. It indirectly or directly focuses on group identities and has a potential to incite violence (Casanovas and Oboler, 2018). On the other hand, we describe offensive speech as a speech that usually targets individuals with the intention to be offended, but the offense should not be due to the individual's group identity (Fortuna et al., 2020; Casanovas and Oboler, 2018). This is inline with the definitions of hate speech on the Ethiopian proclamation issued to regulate hate speech and misinformation on social media in 2020[3].

This dissertation focuses on employing machine learning models to address hate speech by deploying algorithms that are capable of analyzing patterns, contexts, and linguistic subtle distinctions for the automatic identification and filtering of harmful content (Fortuna et al., 2020; Davidson et al., 2017; Mathew et al., 2021). We train the models utilizing our new datasets, which contain instances of normal, offensive, and hate speech to learn and distinguish diverse forms of inappropriate expressions. Through progressive learning and adaptation, machine learning models can improve their precision and accuracy in detecting hate and offensive speeches. Effective models that are trained to combat hate speech can also support and enhance effectiveness of content moderation systems on social media platforms (Shen and Rose, 2019). This methodology offers a proactive and scalable solution to manage the dynamic nature of online communications and the evolving patterns of abusive behaviors, such as hate and offensive speech.

We collected tweets from X/Twitter from 2023 to 2024 and selected samples based on several criteria. We had human experts annotate the tweets to determine the labels for each instance, thus providing inputs for the entire process of our hate speech study. We built models utilizing these datasets to classify tweets into categories, identify hatred targets, rate intensity, and assess the toxicity of the tweets. Figure 1.3 provides highlights of the general structure of our proposed methodology employed to combat hate speech, which our models:

- Detect the presence of hate or offensive speech in a tweet and predict its label as either hate, offensive, or normal.
- Identify targets in hateful speech and classify as ethnic, political, religious, gender, disability related,e.t.c.
- predict intensities of hatefulness and offensiveness in tweets and predict numerical intensity scores ranging from one to five.
- Detect, rewrite and detoxify toxic messages, which result in generating more detoxified, non-toxic messages.

The detail descriptions and justifications of the tasks highlighted in Figure 1.3 will be presented in the subsequent Chapters.

Before we get deep into the main focus of this dissertation, which includes collection of quality labeled datasets, building models and analyzing their performances, we have

[3] https://www.article19.org/wp-content/uploads/2021/01/Hate-Speech-and-Disinformation-Prevention-and-Suppression-Proclamation.pdf

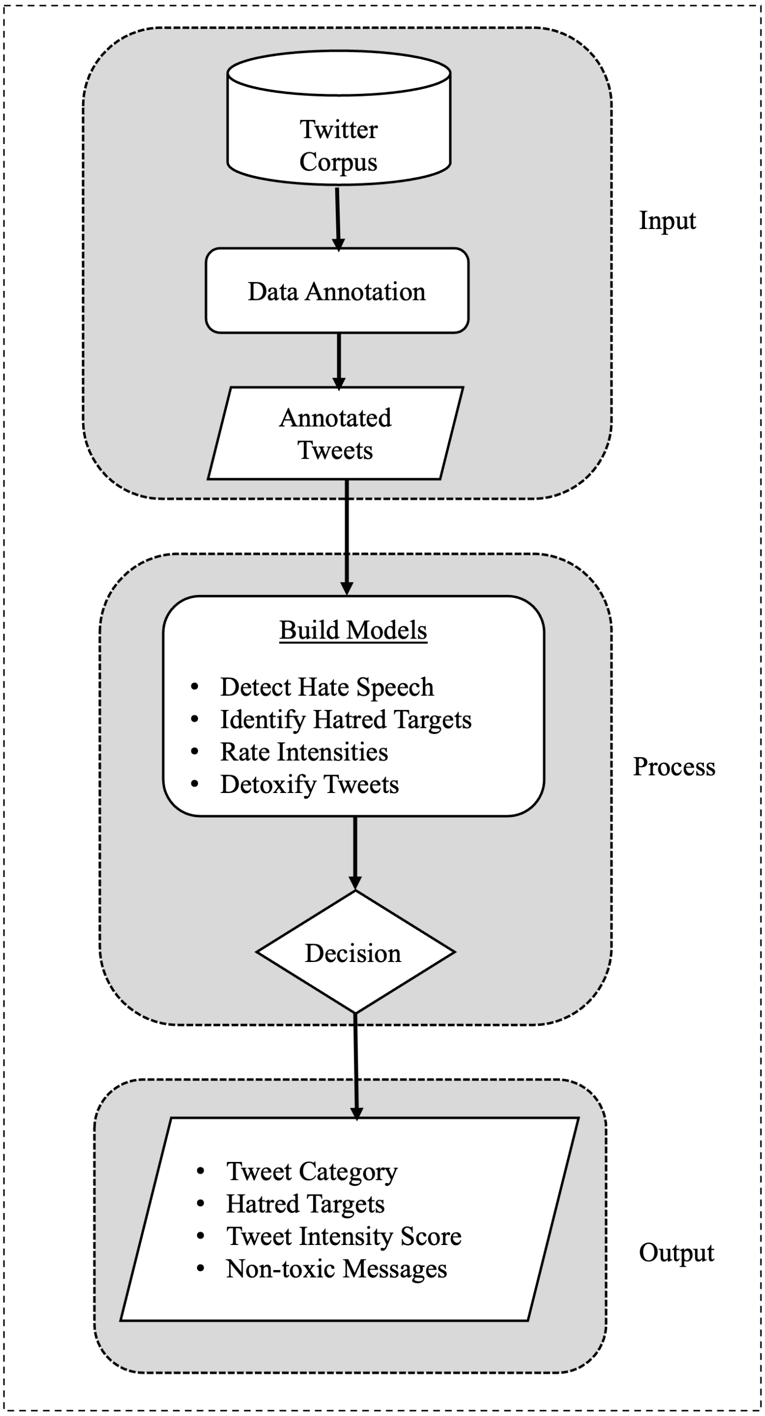


Figure 1.3: Overview of the task procedures.

conducted a brief survey of hate speech studies that have been conducted so far and identified the research gap within the topic.

1.4 Research Objectives

The main objective of this dissertation is to conduct an in-depth analysis of hate speech in low-resource languages, with a particular focus on Amharic. This dissertation aims to investigate the unique challenges posed by the lack of extensive linguistic resources, such as annotated datasets and computational tools, which are often available for high-resource languages. By examining the linguistic and socio-cultural variations that frame hate speech in Amharic, we aim to contribute to the broader understanding of how hate speech manifests in low-resource languages and to propose effective detection and mitigation strategies.

The specific objectives outlined help to comprehensively guide and structure this research endeavor, ensuring a thorough exploration of several key areas, including the challenges of data annotation, identification of targeted communities, the subjective nature of hate speech intensity, the impact of multimodal content such as memes, and the limitations faced by AI models in these context. The detail specific objectives are:

- To investigate the challenges of hate speech data annotation in low-resource languages, such as Amharic.
- To identify communities that are often targeted by hate speech in Ethiopia.
- To explore the subjective nature of hate speech with respect to intensity.
- To examine the impact of multimodality in detecting hate speech in Amharic memes.
- To analyze the challenges of AI models in detecting hate speech in low-resource languages, such as Amharic.

1.5 Research Questions

In this dissertation, we employ comprehensive approaches to analyze hate speech. We conceptualize the following research questions, thereby exploring the complex and subjective nature of hate speech. These include:

- **RQ1**: What are the main challenges in crowdsourcing and in-house hate speech annotation approaches?
- **RQ2**: To what extent do hate speech disproportionately target specific vulnerable communities?
- **RQ3**: How can hate and offensive speech be understood: as distinct categories or as values on a spectrum of varying intensities?
- **RQ4**: To what extent do multimodality enhance the detection of hate speech compared to unimodal approaches?
- **RQ5**: What challenges do large language models (LLMs) face in Amharic text detoxification task?

1.6 Publications Used in This Dissertation

This section provides a list of accepted papers that comprise the dissertation. Additionally, the contributions of authors in each of the accepted papers are presented in detail. I led and completed most of the tasks in each of the accepted papers, for which I am the first author except for (**Ayele**, Babakov, et al., 2024), in which the authors are listed in alphabetic order, and (**Ayele**, Dinter, et al., 2023), in which I and Dinter have equal contributions.

1.6.1 Accepted Papers Comprising This Dissertation

- **Abinew Ali Ayele**, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, Chris Biemann. 2022. **The 5Js in Ethiopia: Amharic Hate Speech Data Annotation Using Toloka Crowdsourcing Platform**. *In proceedings of International Conference on Information and Communication Technology for Development for Africa (ICT4DA2022). Pages 114-120. Bahir Dar, Ethiopia. IEEE.* (**Ayele**, Dinter, et al., 2022).

- **Abinew Ali Ayele**, Skadi Dinter, Seid Muhie Yimam and Chris Biemann. 2023. **Multilingual Racial Hate Speech Detection Using Transfer Learning**. *In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. Pages 41-48. Varna, Bulgaria. INCOMA Ltd.* (**Ayele**, Dinter, et al., 2023).

- **Abinew Ali Ayele**, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Tegegne Asfaw and Chris Biemann. 2023. **Exploring Amharic Hate Speech Data Collection and Classification Approaches**. *In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. Pages 49-59. Varna, Bulgaria. INCOMA Ltd.* (**Ayele**, Yimam, et al., 2023).

- **Abinew Ali Ayele**, Esubalew Alemneh Jalew, Adem Chanie Ali, Seid Muhie Yimam and Chris Biemann. 2024. **Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse**. *In Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024. Pages 167-178. Torino, Italia. ELRA and ICCL.* (**Ayele**, Jalew, et al., 2024).

- Melese Ayichilie Jigar, **Abinew Ali Ayele**, Seid Muhie Yimam and Chris Biemann. 2024. **Detecting Hate Speech in Amharic Using Multimodal Analysis of Social Media Memes**. *In Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024. Pages 85-95. Torino, Italia. ELRA and ICCL.* (Jigar et al., 2024).

- **Abinew Ali Ayele**, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, Xintong Wang, Matti Wiegmann, Seid Muhie Yimam, Eva Zangerle. 2024. **Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking**

Analysis, and Generative AI Authorship Verification Condensed Lab Overview. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Pages 231-259. Springer Nature, Switzerland.* (**Ayele**, Babakov, et al., 2024).

- Daryna Dementieva, Nikolay Babakov, Amit Ronen, **Abinew Ali Ayele**, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee and Alexander Panchenko. 2025. **Multilingual and Explainable Text Detoxification with Parallel Corpora**. *In Proceedings of the 2025 International Conference on Computational Linguistics (COLING 2025). Pages-not yet published. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.* (Dementieva et al., 2025).

- Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, **Abinew Ali Ayele**, Moges Ahmed Mehamed, Olga Kolesnikova and Seid Muhie Yimam. 2023. **Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities**. *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023). Pages 126-139. Dubrovnik, Croatia. Association for Computational Linguistic.* (Tonja et al., 2023).

- Shamsuddeen Hassan Muhammad, Idris Abdulmumin, **Abinew Ali Ayele**, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku and Steven Arthur. 2023. **AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages**. *In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Pages 13968-13981. Singapore, Singapore. Association for Computational Linguistics.* (Muhammad, Abdulmumin, **Ayele**, et al., 2023).

- Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, **Abinew Ayele**, Saif M Mohammad, Meriem Beloucif and Sebastian Ruder. 2023. **SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval)**. *In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). Pages 2319-2337. Toronto, Canada. Association for Computational Linguistics.* (Muhammad, Abdulmumin, Yimam, et al., 2023).

1.6.2 Comments on the Degree of Authorship

In our paper, **The 5Js in Ethiopia: Amharic hate speech data annotation using Toloka crowdsourcing platform** (**Ayele**, Dinter, et al., 2022), I conceived the research idea, design research questions, collected the datasets, managed the annotation task, designed and conducted transformer-based experiments, wrote the paper, and presented it at the conference. Dinter customized the Toloka annotation tool, while Belay conducted classical machine learning experiments. Asfaw, Yimam and Biemann provided overall supervisory guidance.

Our paper **Multilingual racial hate speech detection using transfer learning** (**Ayele**, Dinter, et al., 2023), is conceptualized from Dinter's masters thesis. I and Dinter collected the dasaset for three languages, French, German and Amharic focusing on the death of George Floyd, design annotation guidelines and manage the annotations. German and Amharic racial hate speech dataset are excluded from the study due to low annotation quality. Dinter sets up the data annotation tool and conducted the experiments while I provided technical supervisory contributions. I reformulated the research problem, designed the structure of the research, and wrote the draft paper, which Dinter proofread. I chose the conference venue, submitted the paper, addressed reviewer comments during the rebuttal, prepared the camera ready, and presented it at the conference. I and Skadi assumed equal contributions. Yimam and Biemann provided overall supervisory guidance.

In our paper **Exploring Amharic Hate Speech Data Collection and Classification Approaches** (**Ayele**, Yimam, et al., 2023), I conceived the research agenda, collected the datasets, designed sampling strategies, prepared the annotation guidelines, managed and executed the annotation task, conducted transformer-based experiments, wrote the paper, addressed reviewer comments during the rebuttal, prepare the camera ready, and presented it at the conference. Belay conducted classical machine learning experiments while Asfaw, Yimam and Biemann provided overall supervisory guidance.

In our paper **Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse** (**Ayele**, Jalew, et al., 2024), I conceptualized the research problem, collected the datasets, prepared the annotation guidelines, managed annotations, designed and conducted experiments, analyzed the results, wrote the draft paper, and delivered the presentation at the workshop. Jalew, Ali, Yimam and Biemann provided overall supervisory guidance.

Our paper **Detecting Hate Speech in Amharic Using Multimodal Analysis of Social Media Memes** (Jigar et al., 2024), is conceived from Jigar's masters thesis. Jigar collected the datasets, led the annotation task, and conducted deep learning experiments. I provided close technical supervisory contributions while Jigar condacted data collection, annotation, and deep learning based experiments. I reformulated the research problem to fit into a scientific research paper and conducted additional experiments utilizing six transformer-based architectures for text-only, image-only unimodal, and multimodal models. Besides, I wrote the paper, addressed reviewer comments during the rebuttal, prepared the camera-ready version, and presented the paper at the workshop. Yimam and Biemann provided overall supervisory guidance.

In our shared task paper, **Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview** (**Ayele**, Babakov, et al., 2024), which is a collaborative effort by 29 contributors listed in alphabetical order, my role is associated with all tasks related to the Amharic language. I designed data sampling strategies and collected the datasets, led annotations, and analyzed the experimental results. Additionally, I contributed to the overall writing of the paper.

In our paper **Multilingual and Explainable Text Detoxification with Parallel Corpora** (Dementieva et al., 2025), my role is mainly focused on creating Amharic datasets including annotation quality evaluation. In addition, I analyzed experimental results related to Amharic languages, explore insights and actively participated on the overall paper writing, mainly Amharic language related content.

The paper **Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities** (Tonja et al., 2023), a collaborative effort by several authors, analyzes the challenges, current status, and future directions of NLP tasks in low-resource Ethiopian languages. Tonja contributed to machine translation-related assessments in Ethiopian languages. Belay presented an evaluation of question answering tasks, while Azime contributed to the assessment of POS tagging and named entity recognition tasks in Ethiopian languages. I contributed sections related to sentiment analysis and hate speech and analyzed the challenges, current status, and future directions of these tasks in Ethiopian languages, including Amharic. Mehamed presented an assessment of news classification and summarization NLP tasks in Ethiopian languages, while Kolesnikova and Yimam provided overall supervisory guidance.

The paper **AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages** (Muhammad, Abdulmumin, **Ayele**, et al., 2023), is a collaborative work, which comprised 27 authors working on 14 African low-resource languages. My role in this paper is mainly focused on leading and coordinating the Ethiopian team, which worked on Amharic, Tigrinya, and Oromo languages. I collected the datasets, supervised the annotation procedures, evaluated annotation outputs and experimental results, analyzed the findings related to Ethiopian languages, participated in the overall paper writing, and drafted the sections related to Ethiopian languages.

The shared task paper **SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval)** (Muhammad, Abdulmumin, Yimam, et al., 2023) is a collaborative effort by multiple authors. I led the shared task, specifically for Ethiopian languages such as Amharic, Tigrinya, and Oromo. I collected the datasets, led annotations, evaluated the annotations and data quality, participated in the overall writing of the paper, and drafted the sections related to Ethiopian languages.

1.7 Organization of the Dissertation

This dissertation is organized into eight chapters. Chapter 1 introduces the basic concepts, objectives, research questions and briefly describes the topic under investigation. In addition, we have presented our preliminary studies conducted on social media datasets such as hate speech and sentiment analysis, which help us to conceptualize our main research problem focusing on hate speech (Tonja et al., 2023; Muhammad, Abdulmumin, **Ayele**, et al., 2023; Muhammad, Abdulmumin, Yimam, et al., 2023).

Chapter 2 presents three broad issues: literature review, data collection and annotation strategies, and machine learning models employed throughout the dissertation. Firstly, the literature review covers the state of the art approaches in hate speech studies and assesses the research status in low-resource languages such as Amharic, particularly within the context of social, cultural and political landscape of Ethiopia. Secondly, the data collection and annotation tasks present the specific strategies used and data quality evaluation procedures utilized across the entire study. Lastly, the chapter covers reviews of main machine learning approaches, which are utilized in the study.

In Chapter 3, we present crowdsourcing hate speech studies for a low-resource language, Amharic and a high-resource languages, French, within two broad sections. The chapter presents two case studies, utilizing Toloka crowdsourcing data annotation approaches. While Section 3.2 describes crowdsourcing Amharic hate speech data collection procedures (**Ayele**, Dinter, et al., 2022), Section 3.3 presents crowdsourcing hate speech in French language (**Ayele**, Dinter, et al., 2023).

Chapter 4 thoroughly discusses the procedures of data collection and sampling strategies, and the challenges of hate speech data annotation, focusing on an in-house approach. The chapter also describes hate speech detection and classification tasks utilizing various transformer models on such datasets which are produced in a controlled annotation setup (**Ayele**, Yimam, et al., 2023).

In Chapter 5, we present a more complex dataset annotated for three different tasks: classification, targeted community identification, and hatefulness and offensiveness intensities within tweets (**Ayele**, Jalew, et al., 2024).

Chapter 6 mainly explores multimodal hate speech datasets consisting of Amharic memes and extracted texts. The chapter presents and compares models in unimodal settings such as Image-Only and Text-Only, and in multimodal approach (Jigar et al., 2024).

Chapter 7 introduces parallel datasets of toxic input texts and non-toxic counterparts which are rephrased in a more neutral way. The Chapter presents detoxification results from generative models and showcases the applicability of text rewriting to tackling toxic content on social media platforms (**Ayele**, Babakov, et al., 2024; Dementieva et al., 2025).

Lastly, Chapter 8 presents the concluding remarks drawn from the entire dissertation and introduces future research directions.

Darkness cannot drive darkness; Light can do that. Hate cannot drive out hate; Love can do that.

— Martin Luther King (1963)

2

Related Work, Background and Methodology

Contents

In this chapter, we briefly discuss the related literature on hate speech studies, as well as the data collection, annotation, and data quality evaluation strategies employed throughout this dissertation. In addition, we also shortly present various machine learning models used in the dissertation.

2.1 Review of Literature

This section presents an overview of the Amharic language, the state of the art in hate speech detection studies and briefly discusses the status of low-resource Ethiopian languages.

2.1.1 Amharic Language

Amharic is the second most widely spoken language in the Semitic language family, following Arabic. Amharic is the working language of the Federal Democratic Republic of Ethiopia (FDRE) and many regional states within the country such as Amhara, Addis Ababa, Dire Dawa, South Ethiopia, South West Ethiopia, Benishangul-Gumuz, and Gambela (Salawu and Aseres, 2015; Gezmu, Seyoum, et al., 2018). Nearly one-third of non-Amharic native speakers in towns across Ethiopia speak Amharic as their second or third language, in addition to their own mother tongues (Khan et al., 2011). Moreover, Amharic is used in governmental administration, public media, mass communication, and nationally used for commercial transactions. Amharic scripts are originated from the Ge'ez alphabet which is called Fidäl or 'Ethiopic script'. It has 34 core characters each having seven different variations to represent vowels, which coexisted with the consonants, as presented in Figure 2.1. It has also unique special characters for the majority of the core symbols. The numerals in Amharic constitute 20 unique symbols or digits, which all the remaining other numbers are produced through the combinations of these 20 unique symbols. Additionally, Amharic has its own peculiar punctuation markers, which are utilized in diverse context.

Amharic is a morphologically rich language, where the structure of words and the patterns of word formation are highly complex. A single Amharic word can contain a lot of information about grammatical features such as tense, mood, number, and gender identifiers that can be expressed through affixes (Gezmu, Seyoum, et al., 2018). There are even cases where a single alphabet can be a taken as a complete sentence containing the subject, the verb and the object of the sentence, which also convey full information. Amharic is also one of the highly inflected languages, which follows Subject-Object-Verb (SOV) word order.

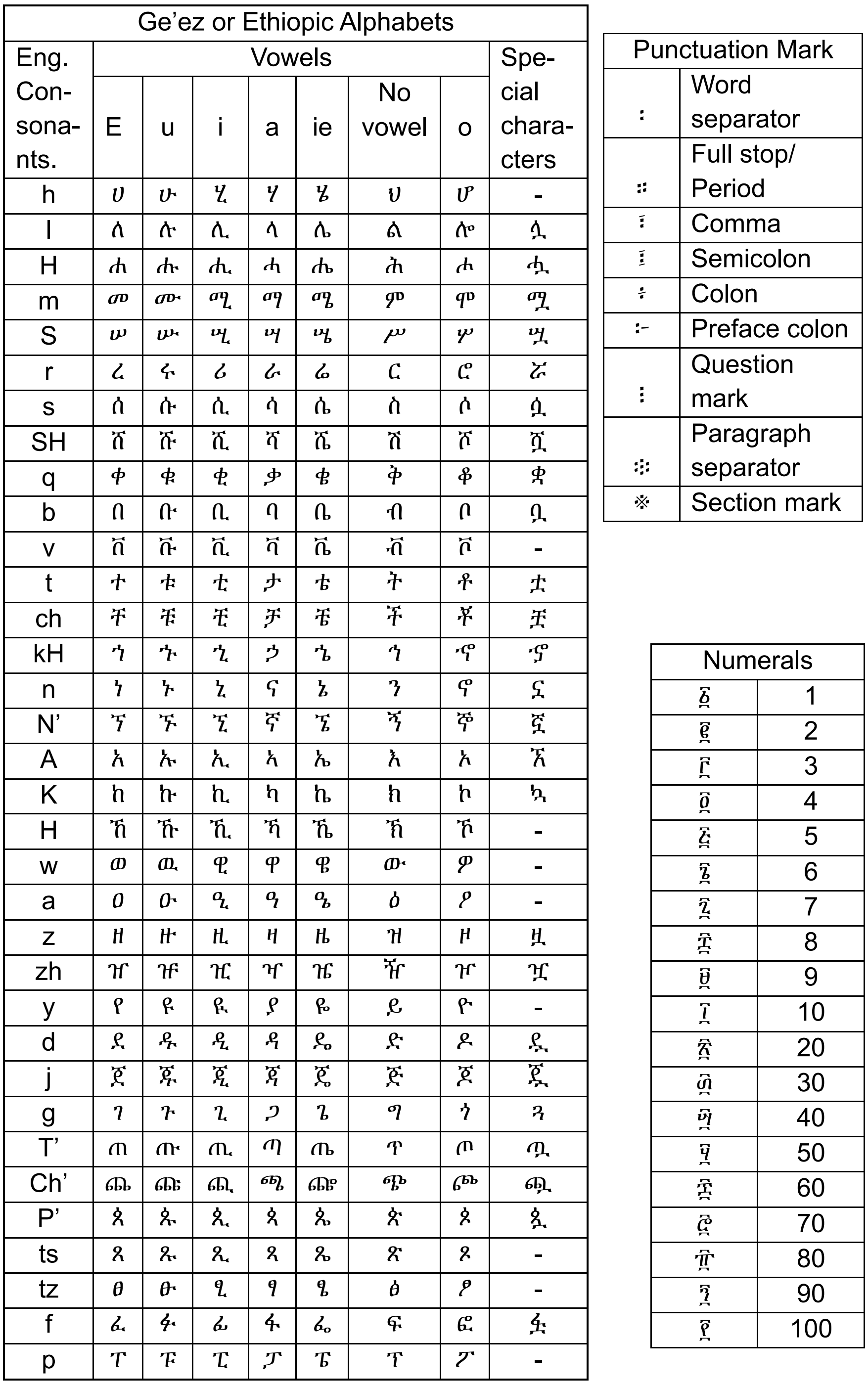
The models built for high-resource and even for other Semitic languages, could not work well for Amharic without extensive experimentation and evaluation of performances with Amharic datasets due to the above aforementioned factors (Yimam et al., 2021).

2.1.2 State of the art in Hate Speech Studies

Various investigations have been undertaken with the primary objective of exploring the complex landscape of hate speech, which is widely spreading on online platforms and mainly targeting marginalized and minority groups. These studies covered a diverse array of disciplines and methodologies, aiming not only to elucidate the mechanisms and manifestations of hate speech but also to develop effective strategies for mitigating its prevalence and harmful consequences within digital environments. On the studies, there have been debates and developments on the definition and conceptualization of hate speech, detection and classification approaches, contextual analysis, and policy implications for regulatory bodies.

Conceptualizations of Hate Speech

Hate speech is a complex concept that has been evolving for decades, which resulted in its diverse definitions among multitudes of researchers, content moderators, policy makers, and regulatory bodies. These stakeholders debate about how to define hate speech and



Ge’ez or Ethiopic Alphabets								
Eng. Consonants	Vowels							Special characters
	E	u	i	a	ie	No vowel	o	
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	-
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	ሏ
H	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	ሗ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	ሟ
S	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	ሧ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ	ሯ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	ሷ
SH	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ	ሿ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቋ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ
v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ	-
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ	ቷ
ch	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ	ቿ
kH	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኋ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ	ኗ
N’	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ	ኟ
A	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ	ኧ
K	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኳ
H	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ	-
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ	-
a	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ	-
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ	ዟ
zh	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ	ዧ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ	-
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ	ዷ
j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ	ጇ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጓ
T’	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ	ጧ
Ch’	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ	ጯ
P’	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ	ጷ
ts	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ	-
tz	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ	-
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ	ፏ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ	-

Punctuation Mark	
፡	Word separator
።	Full stop/ Period
፣	Comma
፤	Semicolon
፥	Colon
፦	Preface colon
፧	Question mark
፨	Paragraph separator
※	Section mark

Numerals	
፩	1
፪	2
፫	3
፬	4
፭	5
፮	6
፯	7
፰	8
፱	9
፲	10
፳	20
፴	30
፵	40
፶	50
፷	60
፸	70
፹	80
፺	90
፻	100

Figure 2.1: The Ge’ez or Ethiopic Alphabets, special symbols, punctuation marks, and numerals.

its boundaries (Zufall et al., 2022; Madukwe et al., 2020). They often distinguish between hate speech and offensive speech, emphasizing the intent to harm or discriminate based on characteristics like race, religion, ethnicity, gender, disability, sexual orientation, etc. Most of the definitions of hate speech include phrases specifying that hate speech is a deliberate or intended attack against specific groups based on their particular group identities (Gagliardone et al., 2014; Zufall et al., 2022; Beyhan et al., 2022). Hate speech lies in a complex relationship with freedom of expression, the promotion of hatred, and the incitement of violence (Gagliardone et al., 2014). This complexity necessitates an in-depth examination of various historical, social, and political contexts in which hate speech arises.

Hate Speech Detection and Classification Approaches

Natural language processing and machine learning techniques have been extensively employed to detect and classify hate speech for the last decades. These methods often involve collecting and annotating large datasets utilized to train models that can automatically identify hate speech in text, image, or audiovisual content. The first category of attempts include early hate speech detection and classification studies such as Warner and Hirschberg (2012), Waseem and Hovy (2016), Nobata et al. (2016), Davidson et al. (2017), Davidson et al. (2019), and ElSherief et al. (2018). These studies mainly employed NLP techniques and statistical machine learning methods such as support vector machine, linear regression, Naïve Bayes, Random Forest, and Decision Trees, which utilized feature extraction techniques. The statistical approaches mainly employ manual engineering to extract relevant attributes, which help the classifier to detect and classify the content as hate or non-hate. Early hate speech studies mainly used feature extraction techniques such as n-gram features (unigrams, bigrams,...) , linguistic features (length of words, number of occurrences of specific terms,...), syntactic features (POS tags, dependency relations,...), and semantics features (like skip-grams) (Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Nobata et al., 2016; Davidson et al., 2017; Davidson et al., 2019; ElSherief et al., 2018). Term Frequency-Inverse Document Frequency (TF-IDF) is also one of the statistical feature engineering method used in classical machine learning and NLP to evaluate the importance of a word in a document relative to a collection of documents, which consists of two main components namely term frequency and inverse document frequency (Robertson, 2004).

The second category of hate speech studies include Ousidhoum et al. (2019), Founta et al. (2019), Winter and Kern (2019), and Kapil and Ekbal (2020), which employed deep learning models that are capable of extracting features automatically to feed inputs to the algorithms. Deep learning models utilized bag-of-words (BoW) and word2vec as input representation techniques that generates important attributes from the raw data and provided as inputs to train the models. Besides, these methods are relatively expensive and complex, which require high computational powers and huge collections of manually labeled datasets to train themselves and produce better results (Tiwari et al., 2018).

The third category incorporates the studies by Demus et al. (2022), Mathew et al. (2021), Röttger, Seelawi, et al. (2022), Röttger, Nozza, et al. (2022), Demus et al. (2022), Logacheva et al. (2022), Floto et al. (2023), Park et al. (2023), Geleta et al. (2023), and El-Sayed and Nasr (2024). Works in this category employ transformer networks, which are specifically fine-tuned for hate speech detection and classification tasks. Transformers have improved the task of hate speech detection through offering better performance in capturing contextualized information, enabling transfer learning, and offering scalability for handling large-scale datasets (Jain et al., 2024).

Majority of the studies on hate speech mainly focused on detecting and classifying in to binary categories (hate or non-hate) or multiple classes such as hate, offensive, abusive, or normal/neutral (Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Nobata et al., 2016; Davidson et al., 2017; Winter and Kern, 2019).

Other studies attempts to identify specific communities or groups who are the recipients of targeted hate speech attacks and subjected to hostility or discrimination

(Ousidhoum et al., 2019; Founta et al., 2019; Davidson et al., 2019; ElSherief et al., 2018; Kapil and Ekbal, 2020),

There are also studies that are mainly focused on measuring hatefulness intensities of messages to evaluate the extent of hostility or harm directed at particular groups (Beyhan et al., 2022; Göhring and Klenner, 2022; Geleta et al., 2023; Tillmann et al., 2023; Sanguinetti et al., 2018; Dahiya et al., 2021; Meng et al., 2023)

Recently, the advancements in generative models have brought novel mechanisms to traditional hate speech detection and classification tasks (OpenAI, 2024; Das et al., 2024) that recommends content moderators to delete hateful or offensive messages (Demus et al., 2022; Floto et al., 2023; Logacheva et al., 2022). The generative models provide text detoxification capabilities, which detoxify and rewrite toxic messages without losing the original meaning of the content. This methods assist content moderators to take actions only on hateful messages that can not be detoxified.

2.1.3 The Status of Hate Speech Studies in Ethiopian Languages

The majority of hate speech studies conducted so far have mainly focused on languages that possess a lot of resources, which include English, French, German, Spanish, and Chinese. However, there are limited research attempts for low-resource African languages such as Amharic, which has faced various machine learning and NLP challenges due to resource scarcities (Yimam et al., 2021).

In our work, Tonja et al. (2023), we explored hate speech detection studies conducted so far in Ethiopian languages such as Amharic, Afan Oromo, Tigrinya, and Wolaytta. We collected and organized the resources in our publicly available GitHub repository[1].

Languages	Author(s)	Size	Algorithm	Score	Dataset	Model
Amharic	Mossie and Wang (2018)	6,120	Word2Vec	85.34	No	No
	Mossie and Wang (2020)	14,266	CNN-GRU	97.85	No	No
	Abebaw et al. (2022b)	2,000	MC-CNN	74.50	Yes	No
	Getaneh (2020)	30,000	BILSTM	90.00	No	No
	Ayele, Dinter, et al. (2022)	5,267	RoBERTa	50.00	Yes	Yes
Oromo	Ababu and Woldeyohannis (2022)	12,812	BiLSTM	88.00	No	No
	Defersha and Tune (2021)	13,600	L-SVM	63.00	No	No
	Kanessa and Tulu (2021)	2,780	SVM+TF-IDF	96.00	No	No
Tigrinya	Bahre (2022)	7,793	NB+TF-IDF	79.00	No	No

Table 2.1: Summary of related works for selected Ethiopian languages in hate speech tasks, *Size* shows the number of sentences used during the experiment, *Score* shows the model results evaluated using F1 score, *Dataset* and *Model* shows the availability of dataset and models in publicly accessible repositories.

Table 2.1 presents a summary of the related works in hate speech detection for selected Ethiopian languages. The table includes the name of the language, the author(s) of the paper, the size of the dataset used, the algorithm used, the score obtained, and the availability of the dataset and model in publicly accessible repositories.

For *Amharic*, five studies were conducted with various approaches and datasets. Mossie and Wang (2018) used word2Vec to detect hate speech and reported an F1 score of 85.34 for binary classification tasks, hate or non-hate labels. In another study,

[1] https://github.com/EthioNLP/Ethiopian-Language-Survey

Mossie and Wang (2020) employed CNN-GRU approach and attempted to explore hatred targeted communities in Ethiopia Abebaw et al. (2022b) emphasized on exploring the MC-CNN method for hate speech studies in Amharic, but with relatively smaller datasets. Lastly, **Ayele**, Dinter, et al. (2022) used datasets annotated on crowdsourcing setups, with low-inter annotator agreement that impacts the quality of the dataset, which is reflected in the model's performance.

For *Oromo*, three studies were conducted, and none of them made their datasets or model publicly accessible. Ababu and Woldeyohannis (2022) annotated 12,812 comments and explored the fine-grained hate speech categories. Defersha and Tune (2021) and Kanessa and Tulu (2021) employed classical machine learning algorithms such as SVM to study Oromo hate speech detection tasks. Bahre (2022) studied hate speech detection task to mitigate its rapid spread on social media. The datasets and models used in these studies were not publicly accessible. In summary, the table shows that hate speech detection in Ethiopian languages is one of the topics of research interest. However, similar to other tasks there is still a lack of publicly accessible datasets and models, which could hinder the development and evaluation of future research. It is worth noting that only two of the nine studies made their dataset and model publicly accessible.

One of the main contributions of this dissertation is to build various datasets that can address multiple domains of hate speech to bridge the research gap in the Ethiopian context in general and the Amharic language in particular. The datasets cover hate speech category classification, target detection, intensity scaling, and detoxification tasks, collected using both crowdsourcing and lab-based annotation techniques.

2.1.4 Unveiling Hate Speech within the Ethiopian Context

Ethiopia is a museum of cultural diversity, serving as the ancestral home to more than eighty distinct nations and nationalities, each with its own unique historical, cultural, and linguistic characteristics (Mengistu, 2015). This diversity contributes to the vibrant mosaic of Ethiopia's cultural heritage, showcasing the richness and depth of its societal fabric (Ayalew, 2020).

Hate speech in Ethiopia has historical roots intertwined with these diversities, dating back to the modern Ethiopian history including the periods of colonial attempts, particularly in the late 19th and early 20th centuries (Ayalew, 2020). During these times, there were various rivalries among ethnic and religious groups who were repeatedly engaged in wars and conflicts. Verbal exchanges between different tribes and religious groups often involve derogatory terms or insults aimed at undermining the other's dignity or promoting their own superiority. Particularly, racism and slavery were highly practiced undermining some tribes and labeling them as 'Barya', which means 'slave', who were discriminated with their skin color, kinky hair, flat noses, thick lips (Ayalew, 2020). During the Italian occupation of Ethiopia, from 1935 to 1941, ethnic divisions and planned prejudices among various ethnic groups were intensified by the invaders to create divisions among the people and weakening their unity and struggle (Ayalew, 2020).

After the Second World War, most struggles for political freedom in Ethiopia were strictly organized along ethnic lines, involving various nations and nationalities such as the Oromo Liberation Front (OLF), the Tigray People's Liberation Front (TPLF), and the Eritrean Liberation Front (ELF). The rise of ethnic based questions that require political representation during both the imperial and Derg regimes have already nurtured oppressed versus operator narrations of "we" and "them" utilizing derogatory expressions against other ethnic groups to mobilize their counterparts. The policies of assimilation and the suppression of ethnic identities ignite the seeds of ethnic hostilities during the imperial era. Similarly, the brutal policies of the derg rigime including the red terror campaign which targeted many ethnic groups have also aggravated ethnic grievances and caused the long civil war in Ethiopia, which lasted for 17 years (Chekol et al., 2023; Mostafa and Meysam, 2023).

Since 1991, Ethiopia has adopted a federal system based on ethnic identities, providing administrative autonomy to the various nations and nationalities spread across all regions. This federal administrative arrangement aimed primarily to address historical marginalization, which had fueled enduring conflicts and civil wars between the centralized government and ethnic insurgencies lasted over three decades, resulting in a significant shift in the country's political administration paradigm (Mostafa and Meysam, 2023; Ayalew, 2020). The new federal system provides various advantages to people, offering government services such as education, health, courts, and other administrative matters in their own language. However, politicians excessively utilized the federal administrative structure, which is organized along ethnic lines. They consistently criticize previous leaders, including the community they came from in order to cover their administrative weaknesses. These narrations aggravated negative sentiments among ethnic identities, which foster a fertile ground for propagating hate speech in Ethiopia. Particularly, ethnic identity became a basis for political organization and competition, which lead to inter-ethnic conflicts and the escalations of ethnically charged rhetoric (Taye, 2017; Mengistu, 2015).

The advancements in social media foster the formation of informal social movements organized to demand freedom and equitable economic opportunities in the country, which mobilized demonstrations and created tensions on the government. For instance, the Ethiopian Muslims' demonstration known as "Dimtsachin yisema" or "Let our voice be heard," as well as youth movements like the Oromo Qeerro and the Amhara Fano demonstrations were typical examples.

During the recent political changes in 2018, Ethiopia underwent significant political reforms, including the rise to power of Prime Minister Abiy Ahmed. While these reforms brought hope for increased freedoms and reconciliation, they also stirred tensions as various political and ethnic factions vied for power and influence. Ethnicity-based conflicts and violence became common incidences across the nation inspired by the propagation of hostilities on social media. There were some hateful terms created every time to label ethnic identities such as "Yeqen Jib" to mean dangerous hyena, and "Junta" to mean dangerous militants to to label Tigrians, "Oromuma or Teregna" to label the government as greedy Oromo lead, and "Jawusa" to label the Amhara youth as cowards and "Neftegnma" to label the Amhara as warriors.

Recently, hate speech has gained greater attention from various stakeholders including the government and researchers to mitigate its impact, particularly in the Ethiopian context. While the Ethiopian parliament has issued regulations to control hate speech and misinformation on social media, researchers from both social science and computer science have attempted to explore and address the spread of hate speech extensively.

This dissertation mainly focuses on mitigating the spread of hate speech across social media platforms through employing machine learning models and NLP applications. Thus, we propose the following general architecture, which shows the entire procedure and the components involved in tackling the spread of online hate speech.

The architecture presented in Figure 2.2 has the following main components:

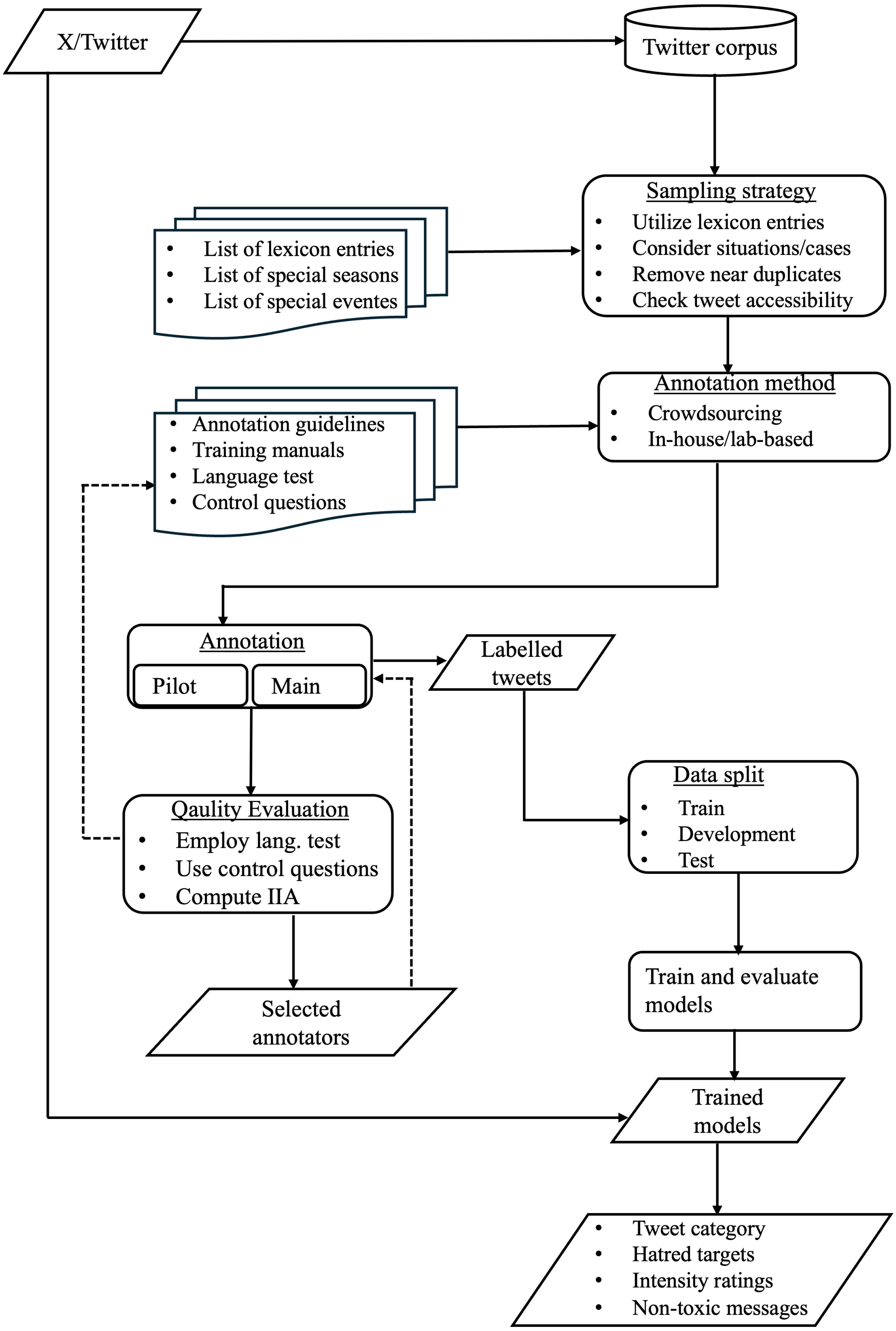


Figure 2.2: Architecture of the proposed system.

- **X/the former Twitter**: is our data source, a free social networking website where users broadcast short posts called tweets, which can contain texts, videos, photos or links[2]

[2] https://x.com

- **X/Twitter Corpus**: is our repository consisting of over 18 million tweets, organized in a relational database, which has been collecting tweets daily since 2014. Our access was blocked in May 2023 when X/Twitter changed its regulations regarding public API access.

- **Sampling Strategy**: We use heuristic mechanisms to select samples for annotation from our repository, such as utilizing lexicon entries, considering various seasons and special events, and excluding retweets and inaccessible tweets that had been deleted by X/Twitter for various reasons.

- **Annotation Strategy**: In this dissertation, we employ both crowdsourcing and in-lab annotation approaches. We designed annotation guidelines and training manuals for each annotation task.

- **Annotation**: in each of the tasks, we conduct both pilot and main task annotation tasks. The pilot annotation is a preliminary phase where a small subset of data is annotated to test and refine guidelines which can ensure the clarity and consistency of the task before full-scale annotation begins.

- **Quality Evaluation**: we employ procedures to supervise the quality of annotation results in each batch and take corrective actions for the next batch. We use both control and language test questions to manage and exclude malicious annotators, especially in crowdsourcing setups. Inter-annotator agreement was also computed for each batch to oversee the overall annotation process.

- **Annotator Selection**: annotators who understand the task well are selected for the full scale annotation.

- **Labeled Data**: the annotated data are processed and prepared for experimentation.

- **Data Split Strategy**: involves dividing the data into train, development and test sets, which is crucial for evaluating the performance of machine learning models effectively.

- **Model Training and Evaluation**: phase where machine learning models learn from datasets and their performances are assessed using various metrics and validation techniques.

- **Trained Models**: once trained, machine learning models can be utilized to predict or classify new data inputs into the corresponding output values.

- **Final Outputs**: the final outputs from the trained models can be categories, hatred targets, intensities or regenerated non-toxic tweets.

2.2 Data Collection and Annotation Strategies

Data is the fuel that powers machine learning algorithms, and its quality and quantity significantly impact the performance and effectiveness of the resulting models.

This section concisely illustrates the overall data collection and sampling strategies utilized in this dissertation. It also presents annotation approaches employed to prepare labeled datasets that were used for all experiments.

2.2.1 Data Collection

Substantial amounts of data are required to train machine learning models and explore the patterns and trends inherent in the data. Collecting and properly labeling datasets are the primary procedures for designing and implementing machine learning models to accurately and efficiently investigate insights from the data.

Since 2014, we have been collecting and storing Amharic tweets on a daily basis, creating a Twitter dataset in a relational database utilizing the Twitter API. Our scripts scrape large numbers of tweets that are written in Amharic, Awgni, Guragigna, Ge'ez, Tigrinya, or other Semitic languages that use the Fidäl script. Currently, we have collected and stored more than 18 million tweets. As indicated in Figure 2.3, the number of tweets stored in our repository showed a substantial increase since 2020 due to the evolving economic, social, and political dynamics in Ethiopia (**Ayele**, Yimam, et al., 2023). Particularly in the years 2020, 2021, 2022, and 2023, until Twitter suspended our API in April 2023, there was a significant increase in the number of tweets collected every day. The major reasons for the surge include, but are not limited to:

- The prevalence of the Covid-19 pandemic and its global impacts,
- Ethiopia's Tigray region holds a regional election in defiance of the federal government,
- The escalations of various national socio-political problems in Ethiopia,
- The conflict between the federal government and the Tigray People's Liberation Front (TPLF) in the Tigray region,
- The 6th Ethiopian national election,
- The assassination of artist Hachalu Hundessa and the imprisonment of opposition political party leaders in Oromia region due to the mass demonstrations and violence in the region following the death of the artist, and
- The Grand Ethiopian Renaissance Dam (GERD) dispute between Ethiopia and Egypt reached a high peak. The GERD case was even taken to the UN security council despite Ethiopia's complaints that it was not a security issue at all.

2.2.2 Data Annotation

Data annotation is the process of adding descriptive metadata or labels carefully to raw data, which can encompass various forms such as text, images, audio, or video (Demrozi et al., 2023). This process involves human annotators who precisely assign tags, categories, or attributes to each piece of data to enable machines understand and interpret the input information accurately. Data annotation plays a pivotal role in training machine learning models by providing labeled datasets that serve as the basis for automatic learning and pattern recognition (Demrozi et al., 2023; Rottger et al., 2022). It requires careful consideration of context, accuracy, and consistency to ensure the resulting annotated dataset is of high quality and can effectively support the development of robust and accurate machine learning models. Data are one of the critical challenges for implementing AI applications. It is either created manually

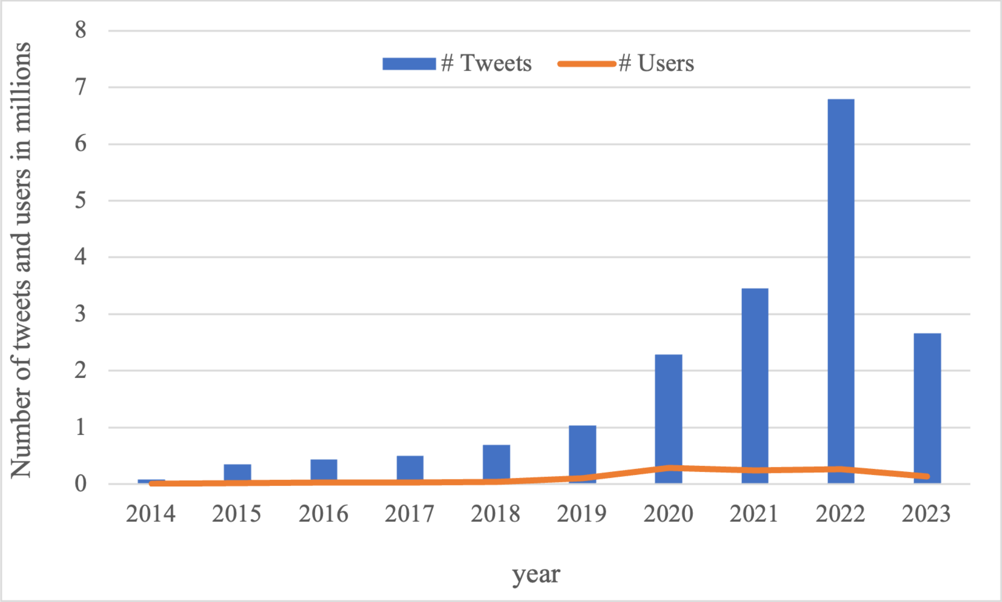


Figure 2.3: Number of tweets and users scraped per year.

through employing human expert annotators or automatically by utilizing advanced machine learning algorithms or tools (Tan et al., 2024). It is an integral part of various artificial intelligence (AI) applications, which is one of the most time-consuming and labor-intensive parts of machine learning projects (Rottger et al., 2022).

The data annotations presented in Chapter 3 were conducted using *Yandex Toloka* crowdsourcing platform, while WebAnn was employed to create the annotation presented in Chapter 4. We utilized POTATo annotator to produce the datasets presented in Chapter 5 and Chapter 7.

2.2.3 Data Quality Evaluation

Data quality, in the domain of machine learning (ML), significantly contributes to determining the performance and reliability of models (Baledent et al., 2022). The success of machine learning models in accurately predicting outcomes or recognizing patterns is significantly tied to the quality of the datasets utilized during the training phase. Therefore, ensuring the quality of data in the aspects of integrity, completeness, and accuracy has great significance for achieving optimal results and fostering trust in the predictive capabilities of machine learning models (Plank, 2022).

We employed various strategies to ensure data quality such as utilizing control questions, an initial language proficiency test, and computing inter annotator agreement. Since a lot of anonymous online users have the opportunity to join tasks in crowdsourcing annotations, preparing language test is used for an initial screening of potential annotators. Those who pass the language test are allowed to join the training tasks. On the other hand, we manually created control questions with the help of human expert annotators. These control questions are shuffled randomly with the task items to filter potential malicious annotators on the crowd who scam to get quick financial wins.

Another mechanism to ensure data quality is to compute inter annotator agreement at each annotation phase: pilot and main annotations. Cohen's kappa, Fleiss' kappa and free marginal kappa statistics methods can be employed to measure agreement among annotators. For this , we employed Cohen's and Fleiss' kappa inter annotator agreement computation methods.

While computing inter-annotator agreement:

- a kappa value of 1 indicates perfect agreement between the raters.
- a value of 0 suggests agreement that is no better than what would be expected by chance alone.
- a negative value indicates that the observed agreement is worse than what would be expected by chance, implying systematic disagreement.

The calculation of Cohen's kappa involves comparing the observed agreement between raters P_o with the agreement that would be expected by chance alone P_e.

The formula for Cohen's kappa is:

$$k = \frac{p_o - p_e}{1 - p_o}$$

where:

- P_o represents the proportion of observed agreement between the raters.
- P_e is the expected agreement between the raters if their judgments were completely independent.

Fleiss' kappa is also another statistical measure used to assess the agreement between multiple raters when categorizing items into mutually exclusive categories. It is an extension of Cohen's kappa, which is designed only for two raters, to be applied in situations involving more than two raters (Fleiss', 1971; Fleiss' et al., 2013). Fleiss' kappa is capable of handling complex relationships when evaluating inter-rater reliability by accommodating multiple raters and categories. It considers the proportion of agreement observed among all raters beyond what would be expected by chance alone. As such, it serves as a suitable metric that can be utilized across a broader range of applications.

Free-Marginal Multirater kappa (Kfree) is another kind of statistical metrics used to evaluate agreement among multiple raters or observers, which is deigned improve the limitations of Fleiss' kappa (Randolph, 2005). Fleiss' kappa can be affected by prevalence and bias, which can potentially lead to the paradox of high agreement but low kappa values. Additionally, it operates under the assumption that raters are constrained in their distribution of cases across categories, which is not commonly observed in many studies assessing inter-rater agreement (Randolph, 2005). Kfree is specifically used when the marginal totals are not fixed or known in advance, unlike in the case of Fleiss' Fixed-Marginal Multirater kappa, where they are fixed and known. Kfree incorporates both the agreement observed among raters and the agreement expected purely by chance, thereby providing a comprehensive evaluation of inter-rater agreement. Kfree and Fleiss' kappa differ in how they handle the expected agreement by chance.

According to (Cohen, 1960; Fleiss', 1971; Gisev et al., 2013), the kappa agreement scores have the various category ranges that spans form poor to perfect agreement as indicated in Table 2.2.

Kappa	Agreement
≤ 0.00	Poor
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table 2.2: Kappa statistics metric categories.

2.3 Machine Learning Models

Machine learning models use algorithms to automatically learn patterns and apply them to make predictions from previously unseen data instances. These models are capable of analyzing data, extracting meaningful insights, detecting implicit patterns, and making informed decisions without requiring human intervention (Tiwari et al., 2018). Most models, especially neural network-based ones, can be trained through iterative learning processes to progressively improve their performance through experience without explicit instructions. Over time, they gradually become more proficient. In addition to exploring and fine-tuning the various transformer-based model, we use the classical machine learning and deep learning algorithms to build hate speech detection models.

2.3.1 Statistical Machine Learning

Statistical machine learning combines statistical methods with machine learning algorithms to enable computers to learn from data and make decisions based on data. This approach utilizes manually engineering features and employ algorithms to learn patterns from the data, which is contrary to deep learning models that relies on neural networks with several layers to automatically learn features from data (James et al., 2013). Even though deep learning models have gained significant attention in recent years due to its capability to automatically learn features from data, statistical machine learning methods also remain relevant and widely used, especially in situations where data is limited or model explainablity or interpretability is significant (Linardatos et al., 2020). Naïve Bayes, logistic regression, and support vector machine are some of thse algorithms, which are also utilized in this dissertation.

Naïve Bayes

Naïve Bayes (NB) is a probabilistic classifier that operates on the principles of Bayes' theorem, which assumes conditional independence between all pairs of features. Despite the "naïve" assumption of feature independence, NB classifiers are widely utilized for their simplicity and efficiency in various machine learning tasks (James et al., 2013). Naïve Bayes models not only presume that all data features are conditionally independent of each other, given the class label, but also assume that all features exhibit a normal distribution within each class and possess equal importance in predicting the class label. It is one of the most popular and straightforward algorithms, which is effective to solve text classification problems (Bishop, 2006). Naïve Bayes Classifier can be trained easily and quickly to be used as benchmark model and perform even better than other statistical models such as logistic regression and SVM (James et al., 2013).

Naïve Bayes can be considered as one of the strongest baseline algorithm for text classification such as hate speech detection due to its simplicity, efficiency, and decent performance on text data despite its sensitivity to class imbalance problems. Various hate speech studies utilized Naïve Bayes such as (Chakravartula, 2019; Chakravarthi, 2020; Vargas et al., 2022; Mossie and Wang, 2018; **Ayele**, Dinter, et al., 2022).

The Bayes' theorem to find probability of event A, given the event B (the evidence) is true, which is stated mathematically in the following equation as:

$$\mathbf{P(A|B)} = \frac{\mathbf{P(B|A)P(A)}}{\mathbf{P(B)}} \tag{2.1}$$

where:

- A and B are events and $P(B) \neq 0$.
- $P(A \mid B)$ is the posterior probability of event A given event B.
- $P(B \mid A)$ is the likelihood of event B given event A.
- $P(A)$ is the prior probability of event A.
- $P(B)$ is the prior probability of event B.

Support Vector Machine

Support Vector Machine (SVM) classifier utilizes algorithms to identify the optimal hyperplane that maximizes the margin between the two closest support vectors and lies equidistant from these data points.

This procedure involves the systematic classification of the dataset into its respective classes by carefully calculating the margin that delineates each class, where vectors are strategically positioned to strengthen and delineate this margin region, thereby acquiring the title of support vectors (Cortes and Vapnik, 1995; Lee et al., 2012; Bishop, 2006).

SVM is a better choice for certain types of datasets and tasks due to its strong capability to capture complex relationships in the data, especially when there is a need for robust generalization and effective handling of high-dimensional data. The effectiveness of SVM classifier depends on various factors such as the choice of kernel function and the nature of the dataset (Lee et al., 2012). SVM can be good choices specially in binary text classification tasks to serve as benchmark models, but it might not capture all the complexities of language and context. SVM uses kernels to project data into a higher-dimensional space when working with text classification as data is often not linearly separable in its raw form (Evgeniou and Pontil, 2001; Cortes and Vapnik, 1995). The textual inputs shall be transformed into numerical vectors using feature extraction techniques such as Bag-of-Words, TF-IDF, and word embednings. SVM has been used in several hate speech studies such as Chakravartula (2019), Chakravarthi (2020), Vargas et al. (2022), Mossie and Wang (2018), **Ayele**, Dinter, et al. (2022), and Chandra et al. (2020), which indicate promising performance results. Figure 2.4 depicts a typical two dimensional space SVM architecture.

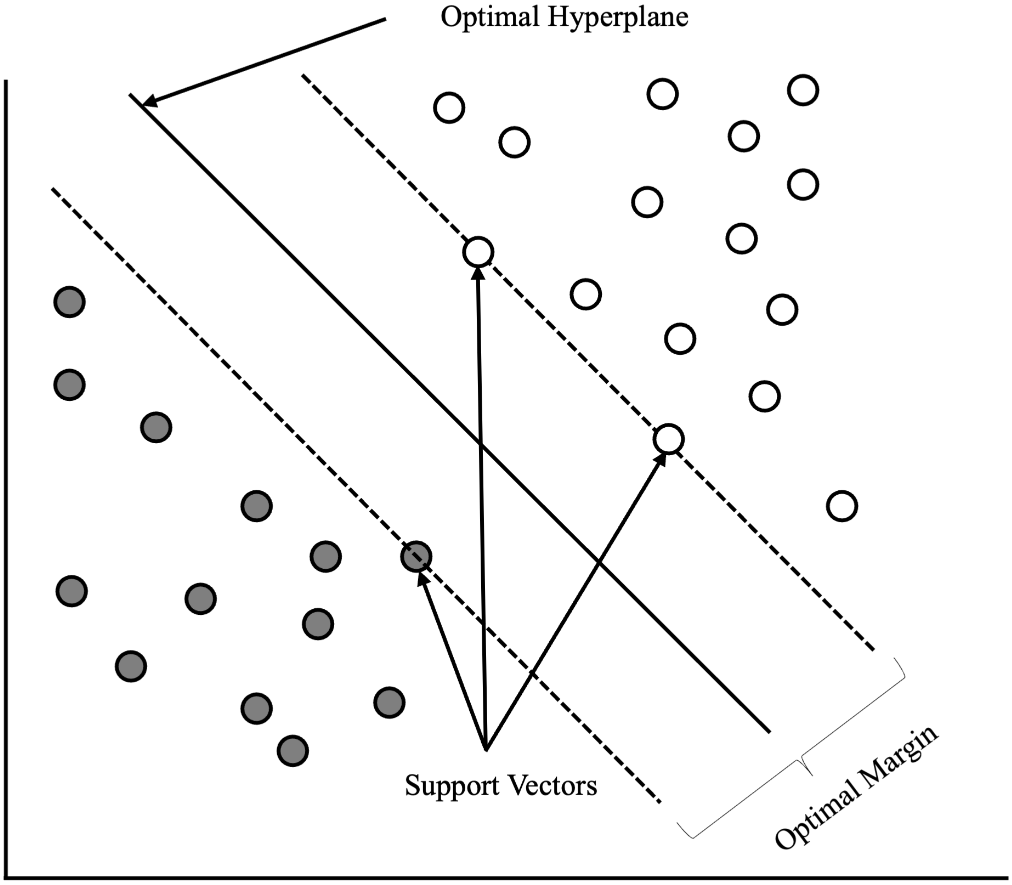


Figure 2.4: Typical SVM classifier.

Logistic Regression

Logistic regression is a statistical technique utilized to predict the likelihood of an event or occurrence based on a set of independent variables in a given dataset, determining whether it belongs to a particular category or not. It models the relationship between multiple independent variables (predictors) and a dependent variable (the outcome), where the outcome variable represents probability values bounded between 0 and 1 (Park, 2013). Logistic regression is easier to implement, provides valuable insights, and performs better on datasets that are linearly separable. It employs the sigmoid function to map predictions and their probabilities of real values to a range between 0 and 1. Logistic regression has been employed by various hate speech researtchers and showed promising results, these include Chakravartula (2019), Chakravarthi (2020), Vargas et al. (2022), Mossie and Wang (2018), Chandra et al. (2020), **Ayele**, Dinter, et al. (2022), and Baruah et al. (2019). Logistic regression is computed based on the following equation as:

$$y = \frac{\mathbf{e}^{(\mathbf{b}_0+\mathbf{b}_1\mathrm{X})}}{\mathbf{1} + \mathbf{e}^{(\mathbf{b}_0+\mathbf{b}_1\mathrm{X})}} \tag{2.2}$$

Where:

$$\begin{aligned} P_{(x)} &= \text{predicted output.} \\ x &= \text{ input value.} \\ b_0 &= \text{bias or intercept term.} \\ b_1 &= \text{coefficient for input (x).} \end{aligned}$$

Feature Extraction in Statistical Machine Learning

In statistical machine learning, feature extraction is an important step employed to identify and extract relevant features from raw data elements. The extracted features are then used to improve the informativeness of the dataset for machine learning algorithms. Feature extraction in natural language processing typically transforms raw text data into a format that machine learning algorithms can better understand the underlying problem and process. These techniques include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word Embeddings, and n-grams.

In this dissertation, we have employed the following feature extraction methods:

- **Bag of Words (BoW)**: BoW is an effective technique in NLP where the words (i.e. features) used in a text document can be extracted and categorized based on their usage frequency. It is the simplest method to represent text in numbers, which transforms a sentence into a bag-of-word vector. A vector space is a multi-dimensional space in which each individual word becomes a separate dimension. BoW assigns one dimension for each unique word in the document set and plots each separate text document as a point in the vector space (Salton et al., 1975). A vector of word counts represents each document, where the machine learning algorithms utilizes the word count as an input (Manning et al., 2008).

- **Term Frequency-Inverse Document Frequency ($\mathrm{TF-IDF}_{(t,d)}$)**: $\mathrm{TF-IDF}_{(t,d)}$, an extension of BoW, is an NLP feature extraction technique that employs a numerical measure to indicate how important a word is to a document in a collection or corpus. Compared to BoW, $\mathrm{TF-IDF}_{(t,d)}$ considers not only the frequency of a word in a single document, but also all other documents in the corpus. Unlike the bag-of-words method, which emphasizes words solely based on counts, $\mathrm{TF-IDF}_{(t,d)}$ combines term frequency ($\mathrm{TF}_{(t,d)}$) and inverse document frequency ($\mathrm{IDF}_{(t)}$) to weight the importance of terms in the document (Manning and Schütze, 1999; Manning et al., 2008).

  The Equations 2.3 and 2.4 depict how the $\mathrm{TF-IDF}_{(t,d)}$ features are extracted from text inputs.

$$\mathrm{IDF_t} = \log(\frac{\mathrm{N}}{\mathrm{DF_t}}) \tag{2.3}$$

$$\mathrm{TF-IDF_{(t,d)}} = \mathrm{TF_{(t,d)}} * \log(\frac{\mathrm{N}}{\mathrm{DF_t}}) \tag{2.4}$$

where:

$$\begin{aligned} \mathrm{TF}_{(t,d)} &= \text{the frequency of term t in document d.} \\ \mathrm{DF}_{(t)} &= \text{number of documents containing term t in corpus.} \\ \mathrm{IDF}_{(t)} &= \text{inverse document frequency.} \\ \mathrm{N} &= \text{number of documents.} \end{aligned}$$

- **N-grams**: N-grams represent contiguous sequences of n items from a given sample of texts to capture linguistic patterns in sequential data, such as text. N-grams are combinations of words, characters, or subwords that create a unit of meaning from the structures and contexts of words or phrases in the data to improve performances of machine learning algorithms (Manning and Schütze, 1999). N-grams are capable of capturing the sequential structure and dependencies in language, which helps NLP models to understand context better and make more accurate predictions. The choice of n affects its computational complexity and the level of details in capturing more context. Figure 2.5 depicts the N-gram feature extraction architectures such as unigrams, bigrams and trigrams, which are typically used in statistical NLP.

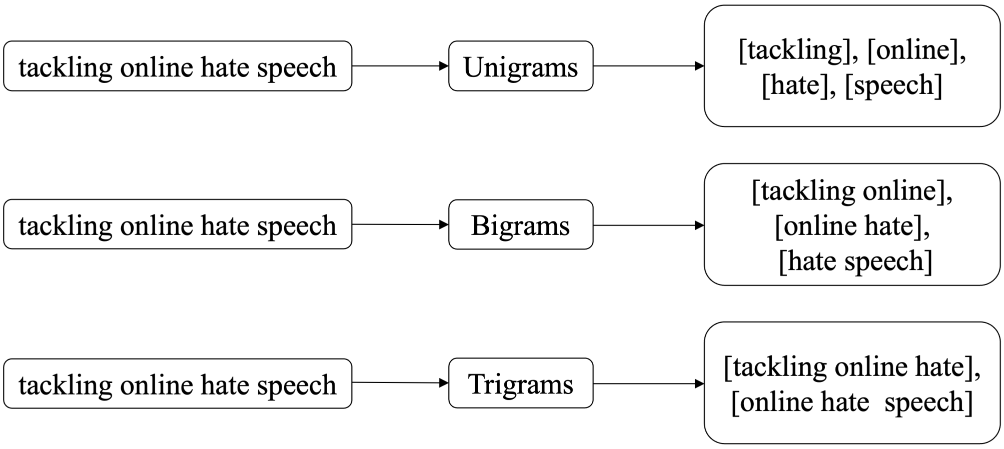

tackling online hate speech
Unigrams
[tackling], [online], [hate], [speech]
tackling online hate speech
Bigrams
[tackling online], [online hate], [hate speech]
tackling online hate speech
Trigrams
[tackling online hate], [online hate speech]

Figure 2.5: N-gram feature extraction procedure.

2.3.2 Deep Learning Models

Deep learning is a subfield of machine learning that emphasizes on modeling artificial neural networks (ANNs) consisting of multiple layers (known as deep neural networks), which are capable of learning complex representations of data through hierarchical composition of simpler features. The development of these algorithms is mainly inspired by the functionality of neurons to extract features automatically and perform computations in tasks that require complex decision-making processes. In this dissertation, we employed deep learning models such as recurrent neural network (in Chapter 4), convolutional neural networks, long short-term memory, and bidirectional long short-term memory in Chapter 3, 4 and 6.

Convolutional Neural Network (CNN)

CNN is a type of deep neural network architecture specifically tailored for handling structured grid-like data such as images. This algorithms learns directly from the input data without the need for manual feature extraction procedures (LeCun et al., 1989). A typical CNN architecture comprises convolutional, pooling, and fully-connected layers. The convolutional layer serves as an N-gram feature extractor in text classification tasks. In CNN, firstly, the input sequence undergoes conversion into a 2-D matrix through an embedding layer. Subsequently, a convolutional layer, along with dropout and max pooling layers, will convert the embedding matrix into a one-dimensional vector. Finally, the last layer produces probability distributions across the target classes (Hassan and Mahmood, 2017). Despite CNN is more suited for computer vision tasks, it has been used in text classification tasks such as hate speech and sentiment analysis, and achieved better results. Various hate speech and classification studies have used CNN (Abebaw et al., 2022b; Alshaalan and Al-Khalifa, 2020; **Ayele**, Dinter, et al., 2022; T Tran et al., 2020; Kamble and Joshi, 2018). Figure 2.6 presents a typical CNN architecture.

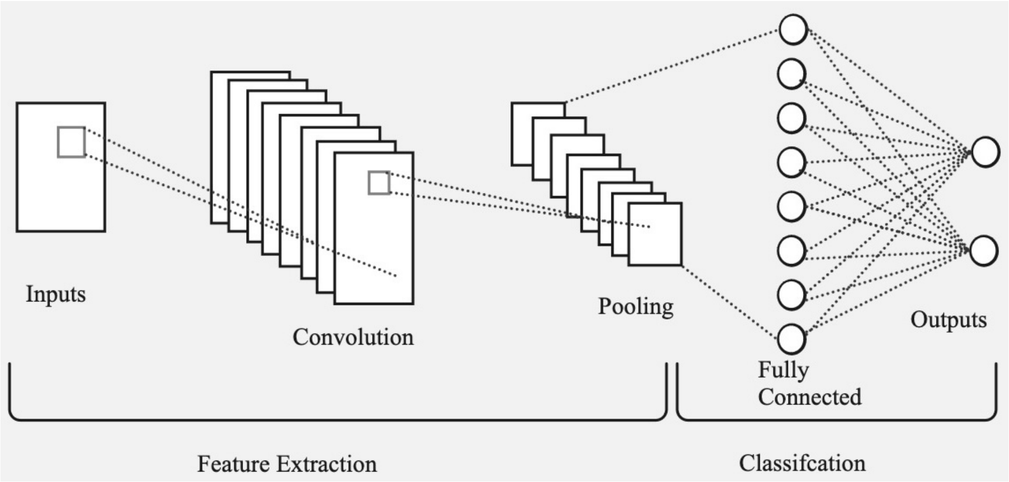


Figure 2.6: CNN architecture.

Recurrent Neural Network (RNN)

RNN is a special type of neural network, which is typically designed to handle sequential data, suited for textual data classification tasks. This algorithm introduces the concept of memory, which enables the network to preserve information about previous inputs in the network. Memory in the RNN architecture is important for various language understanding and generation tasks where context matters.

An RNN architecture, consisting of an input layer, a hidden layer, and an output layer, is designed to process sequential data by maintaining a hidden state that captures information about previous inputs. RNNs have recurrent connections that allow information to flow within the network (Hochreiter and Schmidhuber, 1997). At each time step, the RNN updates its hidden state using the current input and the previous state, capturing dependencies within the input sequence through these recurrent connections. The output is then computed from the hidden state (Mienye et al., 2024).

The standard RNN architecture is shown in Figure 2.7.

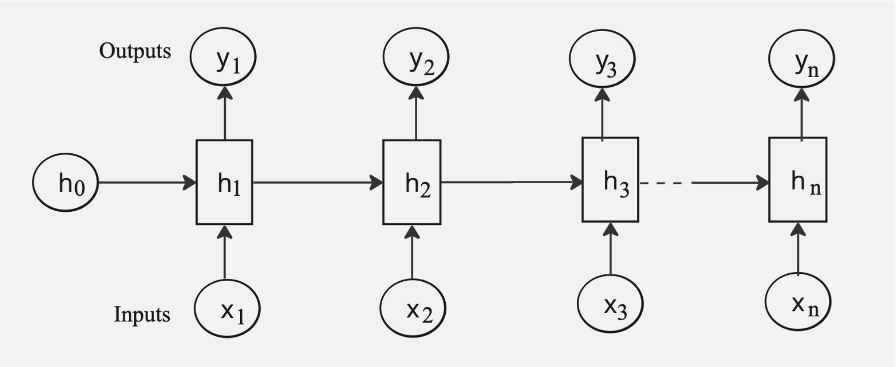


Figure 2.7: RNN architecture.

RNN models commonly face challenges stemming from the vanishing and exploding gradient problems, particularly when dealing with longer input sequences (James, 2020). As gradients are propagated backward in time, they can either diminish (vanish) or grow exponentially (explode), making it difficult for the network to learn long-term dependencies (Mienye et al., 2024). Variants of RNNs, such as LSTMs and gated recurrent units (GRUs), have been developed to mitigate these issues. These variants introduce gating mechanisms that regulate the flow of information and gradients through the network, enabling better handling of long-term dependencies (Hochreiter and Schmidhuber, 1997). Studies such as (Alshaalan and Al-Khalifa, 2020) employed RNN methods for hate speech detection.

Long Short-Term Memory (LSTM)

LSTM is a special type of recurrent neural network architecture, which is designed to address the vanishing gradient problems and capture the long-term dependencies in sequential input data. An LSTM architecture contains specialized mechanisms that allow it to store and retrieve information over long input sequences (Hochreiter and Schmidhuber, 1997; James, 2020). An LSTM architecture consists of three main gates that control the flow of information into, out of, and within the cell's memory, allowing it to learn how to retain or forget information over time.

The input gate controls the flow of information into the cell state and learns to accept or reject incoming data, while the forget gate determines what information to retain or discard from the previous cell state. The output gate controls the information used to generate the output and decides what part of the cell state to disclose to the external environment at each time step (James, 2020). Chandra et al. (2020), T Tran et al. (2020), and Kamble and Joshi (2018) have utilized LSTM methods in their hate speech classification task. The architecture of LSTM is shown in Figure 2.8.

Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTM is another variant of RNN predominantly utilized in natural language processing tasks. Unlike the conventional LSTM, it processes input from both directions, which enables it to receive information from both past and future contexts. BiLSTM is capable of capturing sequential relationships among words and phrases in a bidirectional manner employing an extra LSTM layer that operates in reverse direction to facilitate backward information flow within the input sequence (Graves and Schmidhuber, 2005).

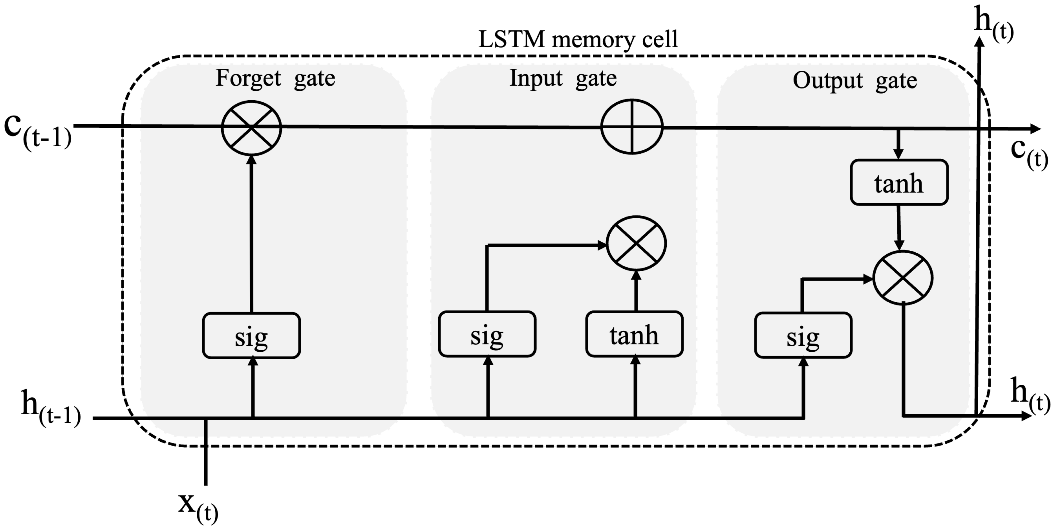


Figure 2.8: LSTM architecture.

Subsequently, the outputs from both LSTM layers are combined using various methods such as averaging, summation, multiplication, or concatenation. BiLSTM has been used by various hate speech studies especially before the introduction of transformer models such as (Baruah et al., 2019; T Tran et al., 2020; Kamble and Joshi, 2018). Figure 2.9 presented typical BiLSTM deep learning architecture.

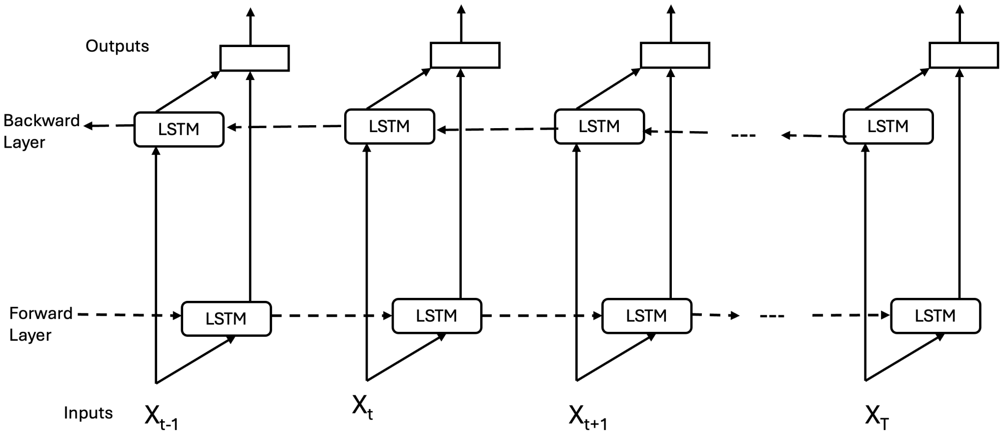


Figure 2.9: Typical BiLSTM architecture.

2.3.3 Transformer Networks

Transformer is a deep learning architecture, which operates on the basis of self-attention mechanism that weights the importance of each part of the input data differently. The Transformer architecture employs an encoder-decoder framework that does not depend on recurrence or convolutions to produce output. The encoder transforms an input sequence into a sequence of continuous representations. Subsequently, the decoder utilizes both the encoder's output and its own previous output to generate a sequence of outputs (Vaswani et al., 2017).

After tokenizating the inputs, the transformers architecture employs embedding to convert the tokenized inputs in to numeric vectors, which then will be combined into one vector and proceed through positional encoding. The transformer block, which composed of the attention and the feedforward components adds contexts to each input vector and predicts probability scores for each output. The final component of the transformer network is the softmax layer, which refines the probability scores obtained from the attention component and assigns the highest scores to correspond to the highest probabilities (Vaswani et al., 2017). In text classification, the softmax layer assigns scores or probabilities to each class, representing the likelihood that the input text belongs to each category. The class with the highest probability is usually selected as the final prediction. The general transformer architecture is shown in Figure 2.10.

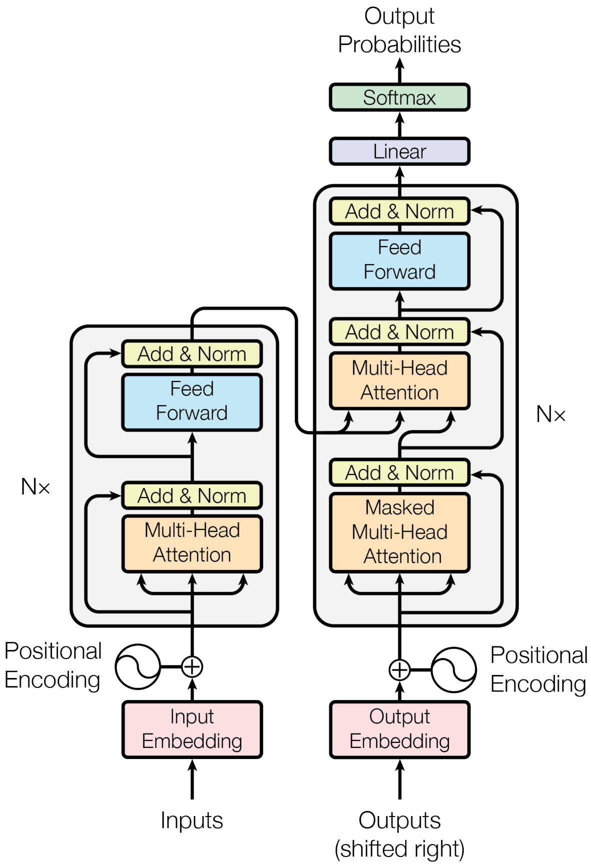


Figure 2.10: Transformer architecture (Vaswani et al., 2017).

In this dissertation, we employ various transformer models, which had been fine-tuned for low-resource African and Ethiopian languages including Amharic. These models include Am-RoBERTA and Am-Flair (Yimam et al., 2021), Afro-XLMR-large (Alabi et al., 2022), XLMR-Large (Conneau et al., 2020), variants of AfriBERTa such as AfriBERTa-small, AfriBERTa-base, and AfriBERTa-large (Ogueji et al., 2021), and AfroLM-Large with active learning setup (Dossou et al., 2022). These transformer-based models are fine-tuned specifically to reflect low-resource language contexts including Amharic.

Utilizing fine-tuned language models in low-resource settings improves performance, efficiency and accuracy with limited datasets. It enhances accessibility and cost-effectiveness, allows faster deployment, leverages transfer learning, and adapts better to specific domains through providing customized solutions for local needs and contexts (Grießhaber et al., 2020).

Am-Flair is a contextualized FLAIR embedding model typically trained on Amharic corpus from scratch with parameters of sequence length 250, minimum 100 batch size, and 10 maximum epochs, which is implemented based on sequences of characters where words are contextualized by their surrounding texts (Yimam et al., 2021).

Similar to Am-Flair, **Am-RoBERTa** is a RoBERTa-based language model that has been fine-tuned specifically with the Amharic language dataset with parameter settings of 5 epochs, train batch size of 8 per gpu, and block size of 512, which make it well-suited for downstream tasks and applications involving Amharic text (Yimam et al., 2021).

Afro-XLMR-large is a multilingual language model tailored for African languages including Amharic, which demonstrated exceptional performance in various natural language processing tasks for African languages.

XLM-R Large is a pre-trained language model developed by Facebook AI, based on the RoBERTa architecture. It excels at understanding and processing text in multiple languages, thanks to its cross-lingual capabilities. With its large size and extensive training data, XLM-R Large is a powerful tool for various natural language processing tasks across different languages, from translation to sentiment analysis, without the need for language-specific models (Conneau et al., 2020).

AfriBERTa and its variants are based on a BERT (Bidirectional Encoder Representations from Transformers) language model designed for understanding and processing African languages, which is tailored to the complex linguistic characteristics of those languages in continent. It is used to address the need for language models trained on 11 African languages including Ethiopian languages such as Amharic, Oromo and Tigrinya for small, medium and large model variants, which facilitates the various natural language processing tasks specific to African contexts (Ogueji et al., 2021).

AfroLM-Large (w/ AL) stands for AfroLM-Large with active learning, which combines a large-scale language models designed for African languages with active learning techniques. This integration optimizes the learning process by selecting the most informative data points for annotation, improving the model's performance while minimizing the need for labeled data (Dossou et al., 2022).

We have discussed the machine learning models employed in this dissertation. Now, let us turn our attention to obtaining the labeled training datasets, which will be discussed in the subsequent chapters. One way to obtain labeled datasets is through crowdsourcing, which will be explored in Chapter 3.

"Crowdsourcing is a great way to approach creation because in any given point there's always somebody on the Internet who knows something better than you do."

— Guy Kawasaki

3

Crowdsourcing for Hate Speech Annotations

Contents

3.1 Introduction to Crowdsourcing

Crowdsourcing offers opportunities to exploit the collective efforts of numerous individuals in the online crowd to collect, annotate, and validate large-scale datasets (Z Wang et al., 2015). This approach is particularly useful for tasks that require substantial amounts of labeled data. For instance, it is extensively utilized in natural language processing tasks that often require large datasets to properly train machine learning models (Suhr et al., 2021).

Data annotation is a fundamental task for numerous NLP applications such as part of speech tagging, named entity recognition, sentiment analysis, hate speech and text classification tasks. It involves labeling and organizing data to train models that heavily rely on well-annotated datasets to learn and make precise predictions (Öhman, 2020). Crowdsourcing annotation provides numerous advantages, making it an effective method for data collection and labeling across various fields, especially in machine learning and natural language processing (Garcia-Molina et al., 2016; Suhr et al., 2021; Öhman, 2020; Z Wang et al., 2015). The following are among the advantages of crowdsourcing annotation approach:

- **Scalability**: the crucial aspect of crowdsourcing is its ability to ensure rapid annotation of large datasets through distributing tasks among diverse contributors in the online crowd (Kunchukuttan et al., 2013).

- **Diverse Perspectives**: crowdsourcing allows various contributors from diverse backgrounds and provide a wide range of viewpoints, which can enhance the quality and robustness of the annotated data, especially in social NLP tasks such as hate speech and sentiment analysis (Li and Fukumoto, 2019). Crowdsourcing allows numerous anonymous end-users to participate from different backgrounds and locations (Garcia-Molina et al., 2016).

- **Cost-Effectiveness**: crowdsourcing utilizes a large pool of annotators on the crowd who are often paid per task (Garcia-Molina et al., 2016). Crowdsourcing is less costly and is more affordable than hiring dedicated teams of full-time employees for data annotation (Kunchukuttan et al., 2013).

- **Time Efficiency**: crowdsourcing is capable of ensuring timely completion of large scale data annotations through utilizing the power of online crowd community who participated and completed annotation tasks in short periods (Kunchukuttan et al., 2013; Li and Fukumoto, 2019). It reduces time and cost related to experimental facilities, in-lab personnel, and traditional participant recruitment schemes.

- **Flexibility**: crowdsourcing platforms provide Flexibility in the design and execution of annotations, enabling researchers to create specific data annotation tasks that match their project requirements and easily update or modify these tasks as needed (Feizabadi and Padó, 2014).

Despite crowdsourcing has numerous advantages that enable researchers to reach a wider crowd audience for conducting annotations, there remain various issues challenging its success. These challenges include but are not limited to:

- **Trustworthiness**: The diverse pool of contributors, despite its advantages, may introduce biases and errors in the annotation task, posing challenges to the trustworthiness of the dataset (Hsueh et al., 2009)

- **Reliability issues**: the accuracy of crowdsourcing annotations can also be inconsistent due to erroneous or dishonest responses of untrained participants, raising questions about the reliability of the data being collected (Suhr et al., 2021; Hsueh et al., 2009).

- **Lack of adequate support**: most of crowdsourcing platforms lack adequate support, specially for low-resource languages which hinders inclusiveness and further widens the digital divide.

- **Malicious users**: the presence of a significant number of malicious annotators who exploit tools like Google Translate to gain more financial rewards adds another layer of complexity and mistrust to crowdsourced projects (Z Wang et al., 2015).

Addressing the challenges of crowdsourcing annotation platforms particularly for low-resource languages, requires comprehensive strategies that prioritize quality control, linguistic diversity, and user accountability (Chida et al., 2022; Z Wang et al., 2015).

In this chapter, we discuss the procedures for crowdsourcing data annotation, along with the associated challenges and opportunities. This method has been evaluated in both low-resource and high-resource language context, as detailed in Sections 3.2 and 3.3 of our works (**Ayele**, Dinter, et al., 2022) and (**Ayele**, Dinter, et al., 2023), respectively.

3.2 Crowdsourcing Amharic Hate Speech Dataset

3.2.1 Introduction

The majority of natural language processing research studies disproportionately concentrated on only 20 of the world's 7,000 languages (Magueresse et al., 2020). This limited focus leaves a vast number of languages, particularly found in Africa and Asia regions significantly understudied and underrepresented (Magueresse et al., 2020; A Wang et al., 2013). As a result, these low-resource languages lack technological advancement and linguistic support when compared with more widely studied high-resource languages such as English (Aji et al., 2023; Avetisyan and Broneske, 2023). This imbalance in research attention highlights a critical need for a broader linguistic inclusivity and digital equity, to ensure that language technology tools and resources are accessible to the speakers of all languages (Grützner-Zahn et al., 2024). The reason behind these less researched languages might be due to the limitations of crowdsourcing platforms in Asian and African countries including Ethiopia, lack of online infrastructures such as payment methods, internet connectivity problems, shortage of native crowd performers/users, and the lack of awareness about online jobs (Öhman, 2020).

Most NLP studies in low-resource languages focus on in-house annotation approaches to gather datasets due to the aforementioned limitations facing crowdsourcing platforms (Chida et al., 2022). This issue is also reflected in Ethiopian languages such as Amharic, where attempts so far have primarily focused on in-house annotation datasets for various downstream NLP tasks, including hate speech. There have been attempts to build hate speech datasets and classification models for the Amharic language (Abebaw et al., 2022b; Mossie and Wang, 2018, 2020; Tesfaye and Kakeba, 2020).

In previous studies, data annotation for Amharic hate speech has been conducted in a controlled environment with a few personnel and limited contexts of user opinions. The work by Mossie and Wang (2018, 2020) collected Amharic hate speech datasets from the Facebook pages of individuals and organizations. The authors conducted the annotation in a controlled in-house data labeling scheme and reported a kappa score of 0.57 inter annotator agreement. Similarly, the work conducted by Abebaw et al. (2022b) collected Facebook comments and posts in the same manner as Mossie and Wang (2018) and achieved Cohen's kappa score of 0.8 with two annotators. Since these datasets were collected from a few annotators, they incorporate only limited perspectives, which may have limitations for its applicability due to the subjective nature of the topic under study, hate speech.

Studies by Mathew et al. (2021), Del Vigna et al. (2017), and Ousidhoum et al. (2019) have utilized crowdsourcing annotation techniques to label hate speech data in multiple languages, including English, Arabic, German, and French. Despite the application of these techniques, the findings showed low inter-annotator agreement, generally falling below a kappa score of 0.25. This suggests significant variability and inconsistency among annotators in identifying and classifying hate speech in these languages. An exception to this trend is the work by Mathew et al. (2021), which reported a comparatively higher inter-annotator agreement score of 0.46. These findings highlight the challenges and limitations of using crowdsourcing for hate speech annotation, especially in achieving consistent and reliable annotations results across different languages.

3.2.2 Selecting Appropriate Crowdsourcing Platform for Amharic Hate Speech Data Annotation

Identifying and selecting appropriate crowdsourcing platform for annotating low-resource languages such as Amharic should be the initial step to conduct accurate and efficient annotation projects.

Crowdsourcing platforms such as Amazon Mechanical Turk (MTurk)[1], Yandex Toloka[2], and Crowdflower[3] are the leading platforms employed in various research projects to facilitate large-scale data annotation efforts on the crowd, offering diverse opportunities over the classical in-house annotation strategies (Drutsa et al., 2019).

MTurk is widely utilized in the research community due to its availability of a global workforce around the world. This accessibility makes it an attractive option for researchers requiring diverse and continuous input for their studies (Callison-Burch and Dredze, 2010). However, despite its broad usage, Amazon MTurk poses significant challenges for researchers outside of the United States, Europe, and some parts of Asia (Vaughan, 2017). Access to the platform from these regions is often restricted, which complicated efforts of researchers in other parts of the world to leverage this resource effectively.

Moreover, funding requirements for using MTurk are substantial barriers particularly for junior researchers in developing nations. Obtaining the necessary financial resources

[1] https://www.mturk.com/

[2] https://toloka.yandex.com/

[3] http://crowdflower.com/

to pay for crowdsourced tasks on MTurk is often challenging and limit these researchers to utilize full advantages of the platform's capabilities (Öhman, 2020). This financial constraint further aggravates the challenge in conducting large-scale data annotation projects, ultimately hindering the progress of research in regions already facing resource limitations (Drutsa et al., 2019).

These challenges highlight the need for more inclusive and accessible crowdsourcing solutions that can support researchers globally, regardless of geographic and economic barriers.

In (**Ayele**, Dinter, et al., 2022), we utilized the Yandex Toloka crowdsourcing platform for the annotation of Amharic hate speech dataset. Yandex Toloka, similar to MTurk, is an increasingly prominent crowdsourcing platform. It achieves a substantial workforce of over 25k performers, who engage in approximately 6 million tasks across more than 500 projects on a daily basis, as highlighted by (Garcia-Molina et al., 2016). Toloka presents a favorable option for low-resource languages due to several reasons (Drutsa et al., 2021). Firstly, it offers a cost-effective solution, which is particularly beneficial for projects with limited budgets. Additionally, Toloka facilitates annotation from developing countries, enhancing inclusivity and diversity in data collection efforts. Moreover, the platform provides a training facility for performers, ensuring the quality and consistency of annotations. Lastly, Toloka allows for the filtering of performers based on language proficiency or country of origin, further customizing the annotation process to meet specific project requirements. These attributes collectively position Toloka as a preferred choice for tasks involving low-resource languages.

Section 3.2 proposed a crowdsourcing annotation scheme for hate speech data collection and explored the challenges associated with low-resource languages in general and Amharic language in particular. The dataset of this research is collected spanning across 5 years, 2018 to 2022 on X/ the former Twitter, particularly focusing in each of the June months every year. The data collection focused on 5 consecutive Ethiopian Junes, known as the **"5Js"** where there were controversial socio-political problems happened in Ethiopia which draw the attention of both main stream national medias and social media platforms as well.

This section mainly explores the challenges of crowdsourcing hate speech annotations and its feasibility in a low-resource language setting, such as Amharic.

Section 3.2 presents analysis of Yandex Toloka crowdsourcing platform for low-resource languages such as Amharic. The following are the main contributions presented in this section:

1. **Dataset**: Amharic hate speech dataset collected in a crowdsourcing setup.

2. **Challenges**: the research explored the possible challenges of Yandex Toloka crowdsourcing platform in low-resource languages, particularly Amharic.

3. **Advantages**: the study investigate the potential benefits of crowdsourcing annotations for low-resource language such as Amharic.

4. **Models**: various classification models are developed utilizing the crowdsourced hate speech datasets.

3.2.3 Data Collection

The source of the dataset used in this section is gathered from our X/Twitter data repository, which has been collected since 2014 (Yimam et al., 2019). At the time of annotation, in 2021, the size of repository surpassed 12 million tweets. A total of 5,400 tweets were sampled for annotation, covering a span of five years. This dataset included an initial set of 400 pilot tweets, chosen based on seed keywords. The data is processed and filtered based on the general data collection procedures and strategies described in Section 2.2 and depicted in Figure 2.2 to extract tweets that are written in Amharic, contain sample lexicon entries, exclude re-tweets, remove near duplicate and deleted tweets.

We designated the term "**5Js**" to represent the series of consecutive incidents occurring annually in June from 2018 to 2022. Remarkably, each of these five successive Junes witnessed significant events characterized by increased levels of violence and conflicts within the nation. These occurrences attracted substantial attention and coverage from both mainstream and social media platforms.

In June 2018, an assassination attempt was made on the then-elected Prime Minister of Ethiopia, Dr. Abiy Ahmed Ali, while he was commemorating his first 100 days in office alongside a gathering of people at Maskel Square in Addis Ababa.

In June 2019, a tragic event unfolded as both the Ethiopian army chief and three high-ranking officials from the Amhara region, including the head of state, were assassinated simultaneously in a coordinated manner. These devastating incidents occurred in separate locations, one in Addis Ababa and the other in Bahir Dar, creating a significant and dark moment in the nation's history.

In June 2020, the prominent Oromo artist, Hachalu Hundessa, was tragically killed near his residence. This event triggered widespread violence throughout the Oromia region, leading to the deaths of 86 ethnic Amhara civilians, numerous injuries, and the extensive destruction of properties, including hotels and buildings, which were burned and destroyed. The unrest prompted the government to arrest prominent leaders of the Oromo affiliated opposition party, accusing them of inciting and intensifying the violence and conflicts that occurred.

In June 2021, the 6th Ethiopian national election was conducted under complex situations. The country was in the midst of a severe conflict, with the Tigray People's Liberation Front (TPLF) engaged in a war against the federal government in the northern region. Adding to the complexity, Oromo affiliated opposition parties decided to boycott the election, opting not to participate in the electoral process which the ruling party a sole participant in Oromoia region. These factors contributed to a tight and volatile political environment during the election period. Besides, it is in June 2021 that TPLF rebel forces in Ethiopia's northern Tigray region have recaptured its capital, Mekelle from the national army, after seven months of struggle.

Finally, in June 2022, a tragic massacre occurred in Qellem Wellega in the Oromia region, where hundreds of ethnic Amhara people were killed by the militant Oromia Liberation Army, known as Shenie. Government media reported over 160 deaths and several injuries.

We collected tweets related to each June incident, beginning from the date the incident occurred and continuing for approximately one month. As part of our data preparation process, we removed retweets to avoid duplication, anonymized user identities to protect privacy, and performed all necessary preprocessing tasks. This comprehensive approach ensured that the data was clean, reliable, and ready for annotation.

3.2.4 Data Annotation

The task of annotating hate speech is inherently complex, involving careful judgments and context-sensitive interpretations (Fortuna et al., 2022). This complexity is worsen in low-resource languages, where the challenges are intensified by a scarcity of annotators, the absence of robust annotation frameworks, and a lack of expert researchers dedicated to advancing hate speech research in these languages (**Ayele**, Dinter, et al., 2022). We addressed these challenges by employing three annotators for each tweet to ensure consistency and completeness in the annotation process. To guide the annotators, we carefully prepared comprehensive annotation guidelines including explanatory training examples, which we then uploaded to the Yandex Toloka crowdsourcing annotation platform. The aim of providing guidelines is to standardize the annotation process and enhance the quality of the annotations, despite the inherent difficulties posed by working with hate speech, particularly in low-resource languages.

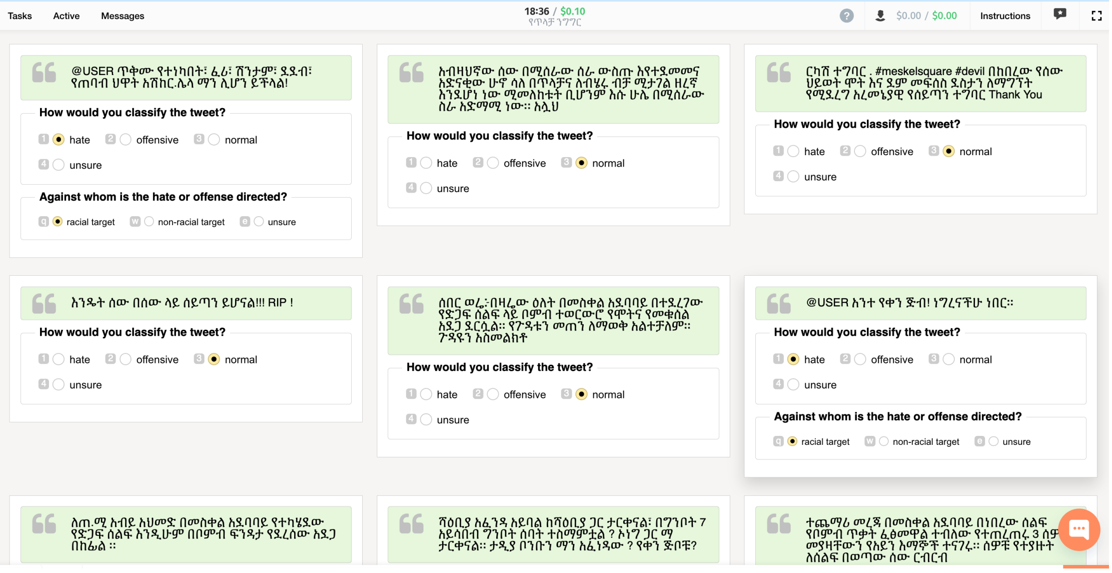
The annotation interface within the Toloka platform is illustrated in Figure 3.1. In order to prepare users, known as *performers* in Toloka, for the annotation task (Drutsa et al., 2021), we first presented them with 20 tweets as a training exercise. This preliminary step ensured that annotators were familiar with the guidelines and the nature of the task before commencing the actual annotation process. Additionally, to maintain the quality and reliability of the annotations, we incorporated 50 control tweets that are crafted by expert annotators with predefined gold labels to identify and filter out potentially malicious annotators.

Each annotation task presented to users consisted of 15 tweets, with one tweet randomly selected from the control tweets using Yandex Toloka's smart mixing technique. This method ensured that every task included a known quality control measure, helping us monitor and assess the performance of the annotators. For measuring inter-annotator agreement (IAA), we employed Fleiss' Kappa, which is appropriate for scenarios involving three overlapping annotators, ensuring the reliability and consistency of the annotations.

Figure 3.2 showcases a sample Toloka interface for one of the annotation pools, providing basic information about the performers. The data presented in the figure indicates that the average time taken to submit an assignment was 5.16 minutes, which breaks down to approximately 0.34 minutes per tweet. The figure also reveals that 88 users showed interest in participating in the annotation task, with 85 users actually participating and submitting at least one assignment. Additionally, it is noted that the average number of assignments submitted by each user was 2.54, equating to nearly 38 tweets per user.

Figure 3.2 provides a snapshot of the Toloka user interface for one of the annotation pools, including basic information about the performers. According to the data depicted in Figure 3.2, the average time taken to submit one task assignment is 5.16 minutes, indicating that approximately 0.34 minutes are required to annotate each tweet. Beside, Figure 3.2 indicates that 88 users showed interest in participating in this particular task pool, while 85 actually participated and submitted at least one assignment. It also indicates that the average number of assignments submitted by each user was 2.54, equating to nearly 38 tweets per user.

The procedures we followed not only ensured that annotators were reliable and well-prepared for the task, but also provided a robust mechanism for monitoring and maintaining the quality of the overall annotation process.



Tasks Active Messages 18:36 / $0.10 የጥላቻ ንግግር $0.00 / $0.00 Instructions

@USER ጥቅሙ የተነካበት፡ ፈሪ፡ ሽንታም፡ ደደብ፡ የጠባብ ህዋት አሽከር.ሌላ ማን ሊሆን ይችላል!

How would you classify the tweet?

1 ● hate 2 ○ offensive 3 ○ normal
4 ○ unsure

Against whom is the hate or offense directed?

q ● racial target w ○ non-racial target e ○ unsure

አብዛህኛው ሰው በሚሰራው ስራ ውስጡ እየተደመመና አድናቂው ሁኖ ሳለ በጥላቻና ለብሄሩ ብቻ ሚታገል ዘረኛ እንደሆነ ነው ሚመለከቱት ቢሆንም እሱ ሁሌ በሚሰራው ስራ አድማሚ ነው። አሏህ

How would you classify the tweet?

1 ○ hate 2 ○ offensive 3 ● normal
4 ○ unsure

ርካሽ ተግባር . #meskelsquare #devil በከበረው የሰው ህይወት ሞት እና ደም መፍሰስ ደስታን ለማግኘት የሚደረግ አረመኔያዊ የሰይጣን ተግባር Thank You

How would you classify the tweet?

1 ○ hate 2 ○ offensive 3 ● normal
4 ○ unsure

እንዴት ሰው በሰው ላይ ሰይጣን ይሆናል!!! RIP !

How would you classify the tweet?

1 ○ hate 2 ○ offensive 3 ● normal
4 ○ unsure

ሰበር ወሬ፦በዛሬው ዕለት በመስቀል አደባባይ በተደረገው የድጋፍ ሰልፍ ላይ በምብ ተወርውሮ የሞትና የመቁሰል አደጋ ደርሷል። የጉዳቱን መጠን ለማወቅ አልተቻለም። ጉዳዩን አስመልክቶ

How would you classify the tweet?

1 ○ hate 2 ○ offensive 3 ● normal
4 ○ unsure

@USER አንተ የቀን ጅብ! ነግረናችሁ ነበር።

How would you classify the tweet?

1 ● hate 2 ○ offensive 3 ○ normal
4 ○ unsure

Against whom is the hate or offense directed?

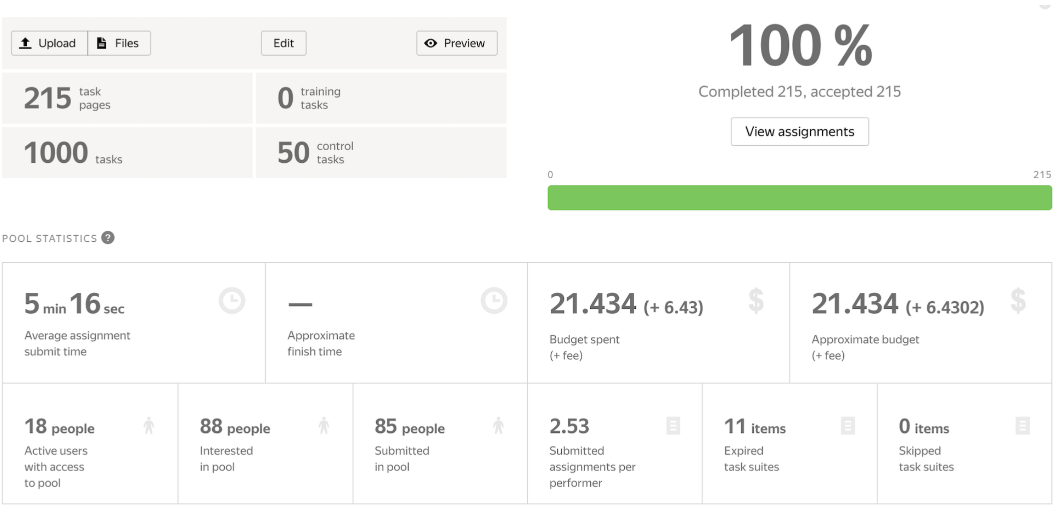
q ● racial target w ○ non-racial target e ○ unsure

ለጠ.ሚ አብይ አህመድ በመስቀል አደባባይ የተካሄደው የድጋፍ ሰልፍ እንዲሁም በቦምብ ፍንዳታ የደረሰው አደጋ በከፊል ፡፡

ሻዕቢያ አፈንጻ አደባል ከሻዕቢያ ጋር ታርቀናል፣ በግንቦት 7 አይሳበብ ግንቦት ሰባት ተስማምቷል ? ኦነግ ጋር ማ ታርቀናል። ታዲያ በንቡን ማን አፈነዳው ? የቀን ጅቦቹ?

ተጨማሪ መረጃ በመስቀል አደባባይ በነበረው ሰልፍ የቦምብ ጥቃት ፈፅመዋል ተብለው የተጠረጠሩ 3 ሰዎ መያዛቸውን የአይን እማኞች ተናገሩ። ሰዎቹ የተያዙት ለሰልፍ በወጣው ሰው ርብርብ

Figure 3.1: The Toloka UI for Amharic hate speech annotation.



215 task pages
0 training tasks
1000 tasks
50 control tasks

100 %

Completed 215, accepted 215

POOL STATISTICS

5 min 16 sec Average assignment submit time	— Approximate finish time	21.434 (+ 6.43) Budget spent (+ fee)	21.434 (+ 6.4302) Approximate budget (+ fee)

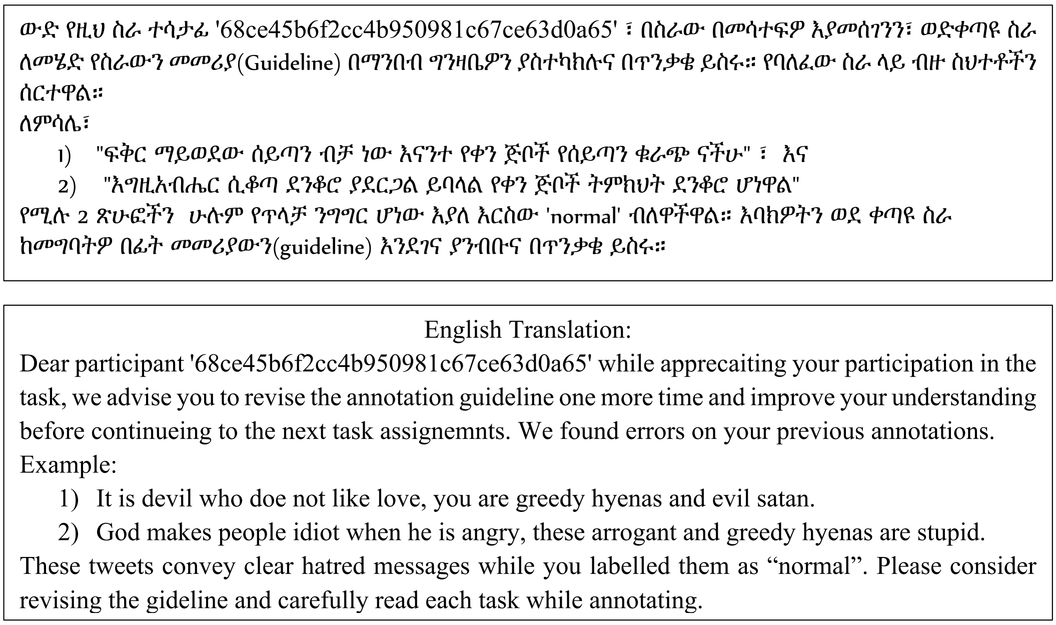
18 people Active users with access to pool	88 people Interested in pool	85 people Submitted in pool	2.53 Submitted assignments per performer	11 items Expired task suites	0 items Skipped task suites

Figure 3.2: Sample Toloka interface that shows performers participated and are interested in one of the pools.

3.2.5 Pilot Annotation

We conducted two rounds of pilot annotations, utilizing a total of 400 tweets, with 200 tweets assigned to each pilot. In the first pilot, 14 annotators were participated in the annotation task and completed assignments. Upon completion of this round, we calculated the inter-annotator agreement and obtained an agreement level of 0.15, highlighting the initial variability and potential challenges in achieving consensus among the annotators.

We conducted a thorough examination of the annotation outcomes, employing a detailed review process that incorporated control tweets to identify individuals who potentially relied on Google Translate, specifically from Amharic to a language they understand, in their annotation efforts. In response, we took decisive action by implementing measures to prevent these performers from further participation in subsequent tasks. Eventually, we sent reminder messages to four identified performers, directly addressing their oversight during the annotation process. These messages included concrete examples extracted from their annotations to illustrate the problems. Moreover, we encouraged these performers to revisit the annotation guidelines and undergo additional training to enhance their understanding and accuracy in future tasks. A screenshot of one of the messages is presented in Figure 3.3.



ውድ የዚህ ስራ ተሳታፊ '68ce45b6f2cc4b950981c67ce63d0a65' ፣ በስራው በመሳተፍዎ እያመሰገንን፣ ወድቀጣዩ ስራ ለመሄድ የስራውን መመሪያ(Guideline) በማንበብ ግንዛቤዎን ያስተካክሉና በጥንቃቄ ይስሩ። የባለፈው ስራ ላይ ብዙ ስህተቶችን ሰርተዋል።
ለምሳሌ፣

1) "ፍቅር ማይወደው ሰይጣን ብቻ ነው እናንተ የቀን ጅቦች የሰይጣን ቁራጭ ናችሁ" ፣ እና
2) "እግዚአብሔር ሲቆጣ ደንቆሮ ያደርጋል ይባላል የቀን ጅቦች ትምክህት ደንቆሮ ሆነዋል"

የሚሉ 2 ጽሁፎችን ሁሉም የጥላቻ ንግግር ሆነው እያለ እርስው 'normal' ብለዋችዋል። እባክዎትን ወደ ቀጣዩ ስራ ከመግባትዎ በፊት መመሪያውን(guideline) እንደገና ያንብቡና በጥንቃቄ ይስሩ።

English Translation:

Dear participant '68ce45b6f2cc4b950981c67ce63d0a65' while apprecaiting your participation in the task, we advise you to revise the annotation guideline one more time and improve your understanding before continueing to the next task assignemnts. We found errors on your previous annotations.
Example:

1) It is devil who doe not like love, you are greedy hyenas and evil satan.
2) God makes people idiot when he is angry, these arrogant and greedy hyenas are stupid.

These tweets convey clear hatred messages while you labelled them as “normal”. Please consider revising the gideline and carefully read each task while annotating.

Figure 3.3: Sample messages sent to individual performers who showed lower performances.

Simultaneously, for the remainder of the annotators, we issued a general message, emphasizing the importance of diligence and careful consideration when annotating tweets for the subsequent tasks. This communication served as a general reminder to all annotators, reinforcing the need for attentiveness and precision throughout the annotation process.

Despite our proactive measures, it is important to recognize the challenge posed by malicious annotators who may use Google Translate or any other automated tools to manipulate the annotation process. While we strive to mitigate such incidents, this problem remains a persistent concern, requiring continuous attention and adaptation of our strategies to validate the integrity of our annotation procedures.

This first pilot provided valuable insights into the annotation process, allowing us to identify areas for improvement in our guidelines and training procedures for the subsequent rounds of annotations.

In the second pilot task, a total of 29 users participated in the annotation process, with the majority being new performers. The inter-annotator agreement achieved for this second pilot was 0.25. When the results from both pilot tasks were combined, the overall inter-annotator agreement was 0.20. Several factors could potentially explain these low inter-annotator agreement scores. One significant factor might be the low payment rate per task, which could affect the motivation of annotators (Garcia-Molina et al., 2016; Huang et al., 2016). Another contributing factor could be the limited or entirely absent training provided to the Toloka performers, leading to inconsistencies in their understanding and execution of the annotation task (Öhman, 2020).

Additionally, the shortage of sufficient annotators proficient in low-resource languages might also play a crucial role, as it limits the pool of available performers capable of accurately conducting the annotations (Hsueh et al., 2009). Lastly, the presence of random annotators primarily motivated by the desire to earn more rewards without a genuine commitment to quality could further contribute to the low agreement levels (Huang et al., 2016). These issues highlight the complex challenges in achieving high inter-annotator agreement and underscore the need for improved strategies to enhance the accuracy and reliability of the annotation process.

3.2.6 The Main Annotation Task

For the main annotation task, we created five pools, each containing 1,000 tweets drawn from datasets spanning the years 2018 to 2022. These datasets focused on the controversial and distressing events that occurred in Ethiopia during the month of June each year. Within each pool, new performers joined the task, while some existing performers were banned from the project. In total, 579 performers from 27 different countries participated in the task. Among these participants, 17 users were blocked from the Toloka crowdsourcing system while 154 users were specifically prohibited from our annotation project.

The majority of participants were originated from three countries, Ethiopia, Pakistan, and the United States. Most of these users participated in only a few tasks. Toloka offers the flexibility to choose annotators based on either the country they reside in or the language they speak. For our project, we opted to select users who could speak Amharic, regardless of their country of residence. Figure 3.4 showed the list of countries which contributed at least five or more participants/performers.

For each annotation task, which consisted of 15 tweets, we offered a payment of $0.10. This compensation rate was set to attract a diverse range of annotators while maintaining our budgetary constraints. Despite the varying levels of participation and the geographical distribution of the annotators, this approach helped us compile a comprehensive and diverse set of annotations for the controversial events of the five Junes in Ethiopia, spanning from 2018 to 2022.

After analyzing the first and second round of pools, we found a significant number of malicious annotators who are engaged in the annotation task. To address this issue, we decided to use the filtering option to select performers based on their country of residence. Initially, we restricted participants only from Ethiopia and the United Arab Emirates. However, no one had begun or engaged in the task for about two days. In response, we revised our strategy and expanded the pool to include all performers who can speak Amharic, regardless of their country of residence. This change brought an immediate and positive impact, with the task being completed within an hour and a half. We checked this strategy twice and obtained the similar results each time.

We presented the biographical information of participant users in the annotation task in Table 3.1. We achieved an IAA of 0.34 for the overall dataset, which demonstrates promising results for crowdsourcing annotation approaches. This suggests that banning

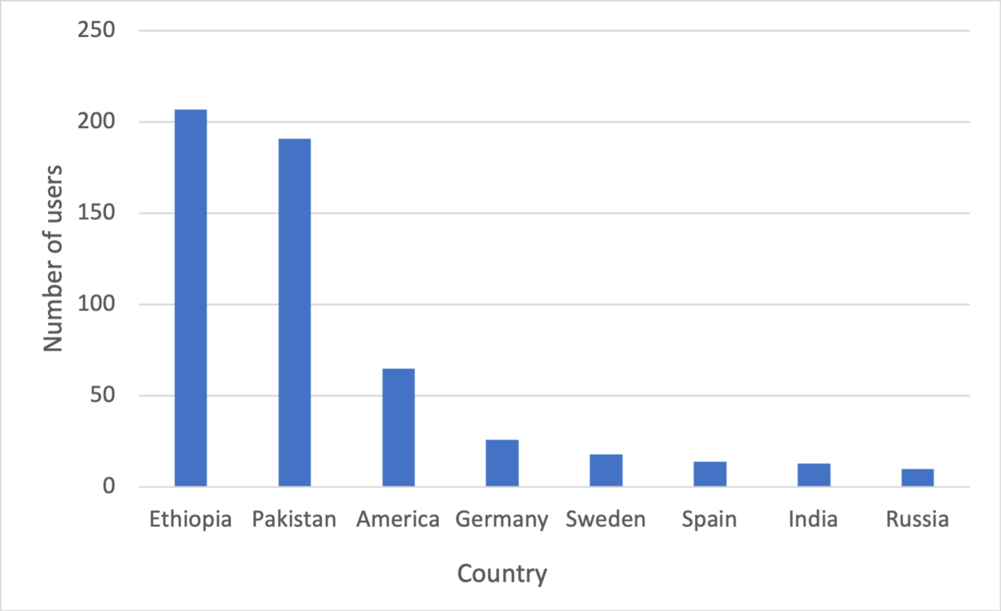


Figure 3.4: Distribution of Toloka performers by country of origin.

parameter	# Count
Total performers participated	579
Number of countries	27
Performers blocked from projects	154
Performers blocked from the system	17
The average age of performers	30 years

Table 3.1: Performers' Basic Information.

suspected malicious users from the project has contributed to the improvement of inter-annotator agreement scores. In our comparison with other related studies, we observed that the work conducted by Del Vigna et al. (2017) documented an inter-annotator agreement score of 0.26 on an Italian dataset while the study by Ousidhoum et al. (2019) reported agreement scores of 0.153, 0.202, and 0.244 for English, Arabic, and French datasets, respectively, using Amazon MTurk. Besides, in a study by Mathew et al. (2021), an inter-annotator agreement score of 0.46 was reported for an English dataset, signifying a moderate level of consensus among annotators. The Fleiss' kappa agreement score of 0.34 observed in our annotation task aligns moderately with similar tasks documented in the literature.

The gold labels for the annotated tweets are determined through a rigorous process of majority voting scheme. When at least two annotators agreed on a label, that label is assigned to the tweet based on the majority consensus. However, out of the 5,400 instances evaluated, 801 were left without a majority label since three annotators provided different category labels for a single tweet. To address this issue, we conducted a fourth round annotation to obtain the majority label. Despite this additional round of annotation process, 134 instances remained without a majority label, as performers consistently opted for the remaining fourth class label (among the 4 labels, namely hate, offensive, normal and unsure). As a result, these instances were excluded from further

label	# annotated	# train	# dev	# test
hate	2,325	1,859	233	233
normal	1,942	1,557	192	193
offensive	617	492	62	63
unsure	382	304	39	39
Total instances	5,276	4,212	527	527

Table 3.2: Dataset distribution

analysis. The dataset used for the experiment comprised 5.3k annotated instances. As it is observed in Table 3.2, the proporation of offensive and usure classes are sparely represented in the dataset.

3.2.7 Annotation Error Analysis

We randomly selected a subset of annotated tweets that presented disputes among the four annotators. These tweets proved that crowdsourcing hate speech annotations are challenging, even after undergoing four rounds of re-annotation. Despite these adjudication efforts, deciding a gold label for these tweets remained difficult. To showcase some of annotation errors encountered, examples are provided in Table 3.3 to highlight the complexities and inconsistencies that faced crowdsourcing hate speech annotations in achieving consensus among annotators.

We choose some of the annotated tweets randomly in the pilot study that contains disputes among the four annotators and become difficult to decide on the gold label even after re-annotating for the fourth time. The annotation errors are presented with examples as depicted in Table 3.3. The gold labels for the three tweets presented in Table 3.3, should have been *normal*, *offensive*, and *hate*, respectively. Despite these examples conveyed less subjective and clear messages as evaluated by human experts, they lack at least two consensus votes out of the four annotations needed to determine their gold labels. This inconsistency could stem from various factors. One possible reason is that some annotators might be malicious users who do not understand Amharic. Alternatively, they might understand the language but choose labels without carefully reading the tweets, motivated by a desire to accumulate as many rewards as possible.

	Tweet	Anno1	Anno2	Anno3	Anno4
1	አፈጉባኤ ወይዘሮ ሙፍሪያት ካሚል የደኢህዴን ሊቀመንበር ሆነው ተመረጡ።	offensive	unsure	normal	hate
	Speaker Mrs. Mufriat Kamil was elected as a chairperson of the SPDM party.				
2	@USER ወሬኛ ፈሳም ነገር ነህ አንተን ብሎ ተሳዳቢ።	unsure	offensive	normal	hate
	@USER You are talkative and coward, what an abuser you are				
3	@USER ባንዳ አህያ ነህ የባንዳዎች ዘር።	Offensive	normal	unsure	hate
	@USER You are a renegade ass. The race renegades.				

Table 3.3: Sample tweets where the gold label can not be determined even after fourth round annotations (Anno: Annotator).

Such behaviors of performers undermine the reliability of the annotation process, leading to discrepancies in the assigned labels.

In consultation with human experts, we also selected a sample of tweets that were incorrectly annotated either due to negligence or because they contained idiomatic and poetic expressions that might not be directly understood by readers due to the contextual variations. Table 3.4 presents these tweets alongside the labels assigned by the three annotators and their computed gold labels.

	Tweet	Anno1	Anno2	Anno3	Gold
1	@USER አንተ ደደብ ቄሮ የሚባል ከብትን ግራ። @USER You idiot, guide your cuttle, the so called qeerro.	normal	normal	offensive	normal
2	ነፍጠኛ ከኦሮሚያ ከኢትዮጵያ ካልጠፋ ሰላም የለም። There won’t be peace unless the musketeers are dismissed from Oromia and Ethiopia.	normal	normal	normal	normal
3	አማራነትን መርጦ የዘር ጭፍጨፋ ማድረግ ይቁም!! Stop genocide targeting the Amharas!!	hate	hate	hate	hate
4	@USER ተጠያቂነት ካልሰፈነ ጭፍጨፋው ይቀጥላል። @USER If there is no accountability, the massacre will continue.	hate	hate	hate	hate
5	የተበተነው ዱቄት አውሎ ንፋስ ሆኖ መጣ። The scattered powder came in a whirlwind.	normal	normal	offensive	normal
6	@USER አንተ ቀልድ አህያውን ፈርቶ ዳውላውን። @USER You are joking, while fearing the donkey, you deal with what it carries.	hate	hate	hate	hate

Table 3.4: Annotation error analysis (Anno: Annotator).

The first two tweets (tweet-1 and tweet-2) were misclassified as ’normal’ when the gold labels should have been ’hate’. These tweets contain clear abusive content targeting ethnic groups, which signifies the performers’ failure to recognize harmful content accurately. This type of error highlights the importance of thorough training and careful attention to context in the annotation process to prevent such misclassifications.

Similarly, tweets-3 and tweet-4 were labeled as ’hate’ by all annotators, despite not containing any hate/abusive content. This resulted in incorrect gold labels, demonstrating how biases or misunderstandings can lead to erroneous annotations. Such errors indicate the need for better guidelines and perhaps a review mechanism to ensure that performers understand and apply the labeling criteria correctly.

Additionally, tweets-5 and tweet-6 contained poetic and idiomatic expressions, which can be particularly challenging to annotate accurately. Tweet-5 ("the dispersed flour comes as a storm") is a *poetic expression* that requires contextual knowledge to interpret correctly. This phrase was used in July 2021 when TPLF rebels captured North Wollo, following the Ethiopian government’s announcement of the complete destruction of TPLF rebels during the law enforcement period, November 2020-2022. Without understanding this historical and socio-political context, performers might misinterpret the expression, leading to incorrect labels.

The last example, tweet-6 ("@USER you are joking; while fearing the donkey, you deal with what the donkey carries"), is an *idiomatic expression* used to describe the act of focusing on indirect trivial issues instead of addressing the main problem directly. Performers who are unfamiliar with this idiom might misinterpret its meaning, leading to inaccurate labeling.

These examples illustrate the complexities and challenges inherent in the annotation process, particularly when dealing with content that requires a deep understanding of social, political and cultural contextual variations.

3.2.8 Experimental Settings

The annotated tweets are further categorized into training, development, and test sets, following an 80:10:10 split, as dipicted in Table 3.2. The development dataset plays a crucial role in fine-tuning and optimizing the learning algorithms. Therefore, all results provided in subsequent sections are based on the instances within the test dataset.

In the computation of trainable parameters or weights for the deep learning models, we employ specific hyperparameters, such as an embedding dimension of 100, 10 epochs, a batch size of 64, a softmax activation function, and the adam optimizer. These parameters are mainly important in determining the effectiveness and efficiency of the learning process.

3.2.9 Results and Discussion

This section presents various experiments conducted for Amharic hate speech classification. This experimental analysis and exploration encompasses classical machine learning models like logistic regression, support vector machine, and Naïve Bayes, alongside deep learning counterparts such as long short term memory, bidirectional long short term memory, and convolutional neural networks. Additionally, we also explored fine-tuned Amharic transformer models like Am-FLAIR and Am-RoBERTa into our analysis. To assess the performance of these models, we employed the key evaluation metrics including F1 score, precision, recall, and accuracy.

As shown in Table 3.5, Am-RoBERTa outperformed other classifiers, achieving an F1 score of 50%. We further explored the predicted test files for error analysis which indicates that the embedding from Am-RoBERa model effectively understands context in tweets, such as idiomatic expressions. Among the classical classifiers, logistic regression and support vector machine perform better than Naïve Bayes, both attaining an F1 score of 49%. The deep learning classifiers, including long short term memory, bidirectional long short term memory, and convolutional neural network, achieve an F1 score of 44%, which is less accurate as compared to the classical classifiers.

The inter-annotator agreement score of 0.34 for our crowdsourcing-based dataset is significantly lower than the typically higher agreement rates achieved through in-lab annotation approaches employed in previous related studies (Mossie and Wang, 2020, 2018; Abebaw et al., 2022b). This issue greatly affects the performance of our models, as the quality and consistency of the dataset are vital for the accuracy and reliability of classification models. The lower quality of the crowdsourced annotations introduces noise and inconsistencies that impede the models' learning effectiveness.

Moreover, our analysis of the predicted test file reveals that the model has particular difficulty with idiomatic expressions. These expressions are context-dependent and often require a deep understanding of cultural and linguistic variations, which the model lacks. This limitation highlights the challenges inherent in the Amharic hate speech classification task, where the complexity of the language and the variations of expression can hinder accurate classification. Addressing these challenges is essential for enhancing model performance and achieving more reliable hate speech detection in Amharic.

Classifier	Precision	Recall	Accuracy	F1 score
Log.Reg	49%	54%	54%	49%
SVM	48%	54%	54%	49%
NB	52%	52%	52%	46%
LSTM	43%	46%	46%	44%
BiLSTM	43%	46%	46%	44%
CNN	43%	48%	48%	44%
Am-FLAIR	46%	51%	52%	48%
Am-RoBERTa	**49%**	**51%**	**51%**	**50%**

Table 3.5: Amharic hate speech classifier models result.

3.2.10 Conclusion

In Section 3.2, we presented a crowdsourcing-based Amharic hate speech dataset, which is the first of its kind to the low-resource Ethiopian languages in general and Amharic hate speech in particular. We presented 5.3k tweets annotated into hate, offensive, normal, and unsure classes. The dataset can be a benchmark dataset for the crowdsourcing-based Amharic hate speech detection task.

We present various models, including transformer-based contextual embedding models, trained on the crowdsourced hate speech dataset. The transformer-based contextual embedding model, Am-RoBERTa, outperformed all classical and deep learning models, achieving an F1 score of 50%. To advance research in Amharic hate speech detection and classification, the dataset, models, and source codes are publicly available under a permissive license at or GitHub repository[4].

3.3 Crowdsourcing Racial Hate Speech in French

3.3.1 Introduction

The increasing prevalence of racist content across various social media platforms is becoming more noticeable due to the rapid developments in such platforms and the anonymity of diverse online communities (Modha et al., 2018; Schwelb, 1966). These developments observed since a couple of years have fostered the spread of such content that convey discrimination, threats, or harassment against minorities (Chiril et al., 2020). The issue is still increasingly worsening on a global scale, inflicting severe physical and psychological harms, thereby significantly affecting the professional, social, economic, and emotional well-being the targeted communities (Tao and Fisher, 2022).

Racism is a type of hate speech, typically targeting people based on their racial identity (Davidson et al., 2019). As it is one type of hate speech, racism is a complex and subjective concept that is difficult to provide a universally accepted definition (Cercas Curry et al., 2024). Our working definition describes racism as a concept that includes all forms of expressions related to stereotypes or prejudice, involving derogatory terms, slurs, or any other form of verbal or written expression intended to humiliate or incite hostility against a person's identities (such as race, ethnicity, color, and hair

[4] https://github.com/uhh-lt/ethiopicmodels

texture). These expressions result in explicit acts of hatred and systemic discrimination, perpetuating inequality (Schwelb, 1966; Rosa and Flores, 2017). Racism can occur between individuals of different races (inter-racial) or within the same race (intra-racial), and it can manifest in various ways, including legal or illegal, direct or indirect, and overt or covert (Krieger, 1999; Matamoros-Fernández and Farkas, 2021). It utilizes linguistic and racial hierarchies to justify notions of the inferiority of racialized people's social, cultural, and linguistic practices, often resulting in economic exclusion (Field et al., 2021).

Perceptions of racial sentiment are deeply influenced by context and individual subjectivity, which makes detecting the prevalence of racism in textual sources inherently challenging (Hasanuzzaman et al., 2017). The complex and often implicit nature of racial discrimination in written content requires a sophisticated understanding of the cultural situations and contextual factors intertwined within the messages (Davidson et al., 2019).

Collecting data from diverse perspectives, potentially from a broad audience, is crucial for addressing and navigating these complexities (Öhman, 2020). In this context, employing crowdsourcing leverages the collective intelligence of a large group of people to identify, label, and categorize hate speech, which is particularly effective for high-resource languages such as English, Spanish, and French (Garcia-Molina et al., 2016; Suhr et al., 2021; Öhman, 2020). This approach is particularly useful for generating large, diverse datasets necessary for training machine learning models (Z Wang et al., 2015).

In (**Ayele**, Dinter, et al., 2023), we collected data from the former Twitter, now known as X, and annotated it using the Toloka crowdsourcing platform. To gather datasets concerning the topic of racism, we consider messages written after the death of George Floyd on May 25th, 2020, for approximately one month.

Social media platforms primarily employ content moderation systems, which are collaborative human-machine systems designed to detect and manage hate speech automatically. However, content moderation systems have limitations in effectively controlling and managing online hate speech and racism (Horta Ribeiro et al., 2021).

3.3.2 Motivation

This research is driven by our interest in exploring the challenges and opportunities involved in annotating racial hate speech data through crowdsourcing, specifically high-resource languages such as French.

To enhance the progress of developing hate speech detection models including racial hate speech across multiple languages, we undertake a study to extend the existing English hate speech detection model from HateXplain (Mathew et al., 2021) to accommodate the French language. This extension involves leveraging our own annotated dataset tailored specifically for French. Through this approach, we aim to contribute to the advancement of hate speech detection technology, making it more inclusive and effective across diverse linguistic contexts. Despite there are various types of bias and discrimination among people, we limit the scope of our research to racial discrimination which is one of the most critical problems in society (Vanetik and Mimoun, 2022) and rapidly escalating in social media (Schwelb, 1966).

This section analyzes the capabilities of BERT and HateXplain models to efficiently adapted for French racial hate speech detection tasks. Additionally, it explores the main challenges of racial hate speech data annotation on the Toloka crowdsourcing platform.

In this section, we have adopted a comprehensive research approach by incorporating a crowdsourcing-based method for annotating racial hate speech datasets. This involves utilizing the Yandex Toloka crowdsourcing platform, a widely recognized tool for such tasks, to gather annotations from a diverse pool of contributors. By harnessing the collective efforts of this crowd, we aimed to ensure a broad and representative coverage of racial hate speech instances within our French dataset.

Furthermore, to adapt HateXplain (Mathew et al., 2021), a BERT-based classification model initially designed for English tweets, and carefully fine-tuned to the French language. This process of fine-tuning involved training HateXplain on a corpus of French tweets, allowing the model to learn the specific linguistic patterns and contextual clues associated with hate speech in the French language. Through this fine-tuning approach, we aimed to enhance the model's accuracy and effectiveness in detecting and categorizing hate speech within French-language social media content.

The main contributions presented in this section include the following but not limited to:

1. Collecting racial hate speech dataset in French,
2. Exploring the annotation challenges of racial hate speech annotation on the Yandex Toloka crowdsourcing platform, and
3. Adaptation of a racial hate speech detection model for the French Twitter dataset.

3.3.3 Data Collection

Most existing hate speech datasets mainly focus on the classical hate speech categories, overlooking racial hate speech, which is a significant challenge in social media. The dataset utilized in this research was carefully collected from Twitter, concentrating on tweets that were published over a one-month period following the death of George Floyd[5]. This timeframe was chosen to capture the surge in public discourse and the significant increase in social media activity related to issues of racism. By focusing on this specific period, we aimed to gather a comprehensive and representative sample of tweets that reflect the diverse opinions of people expressed immediately after this pivotal event.

In the aftermath of this event, social media platforms such as Twitter, Facebook, and YouTube became prominent arenas for expressing opinions, sharing information, and organizing activism (Wirtschafter, 2021). However, these platforms also witnessed a significant increase in hate and offensive speech, with a notable rise in racial hate speech (Carvalho et al., 2022). This spike in harmful content affected online discussions and dialogues during a time of heightened emotions and widespread mobilization (Wirtschafter, 2021). By analyzing this dataset, we aim to understand the dynamics of online hate speech in general and racism in particular through the development of more effective detection and classification models.

We employed 3,473 French hate speech lexicon entries adapted from the work of Stamou et al. (2022) and Chiril et al. (2020) to filter the tweets that might contain racial hate speech content from the total 200m tweet corpus. We used the Python language detection[6] tool to filter tweets that are only written in French. We also removed truncated tweets since such tweets lack complete information and may confuse the annotators during annotation, and the model during experimentation. We removed

[5] The New York Times: How George Floyd died

[6] Python Language detection library:
https://pypi.org/project/langdetect/

retweets and kept only unique tweets that are not duplicated. Moreover, usernames and URLs are anonymized and replaced with <USER> and <URL> respectively. A total of 5k tweets are annotated using three independent annotators on Yandex Toloka crowdsourcing platform.

3.3.4 Data Annotation

The task of annotation is inherently complex that require careful consideration of various linguistic, cultural, and contextual factors (Davidson et al., 2017). The complexity is worsen when annotating for hate speech, particularly racial hate speech due to the absence of complete background context for texts collected from social media platforms (Davidson et al., 2017; **Ayele**, Belay, et al., 2022). In the absence of complete context, annotators can not understand the underlying intent and potential implications of the speech within the broader social and cultural context.

We invest significant effort in the design of annotation guidelines and training examples to ensure clarity and ease for annotators performing the task. This ensures data annotating and enhances the quality of annotated datasets essential for developing effective models to detect racial hate speech.

We have annotated 5k tweets using the Toloka crowdsourcing platform, with each tweet independently annotated by three different Toloka performers. This approach ensures high standards of accuracy and reliability through incorporating multiple perspectives into the dataset.

We selected 50 tweets and had them manually annotated by three French native speakers to create control questions. These annotations were further evaluated by subject experts for accuracy and objectivity. These control tweets were used to monitor the quality of annotations throughout the annotation process.

Each task presented to Toloka performers consists of 15 tweets, with one of these tweets designated as a control question. The inclusion of control questions is a deliberate strategy to ensure the accuracy and reliability of the annotations. Incorporating a known reference point within each task can help to monitor the performance of annotators by identifying discrepancies or random annotations and ensure quality of the dataset.

We requested Toloka performers to classify tweets into *hate*, *offensive*, *normal* and *unsure*. Additionally, performers were asked to further identify *hateful tweets* into *racial*, *non-racial* and *unsure*. Whenever the *hate* class is selected, the targets *racial*, *non-racial*, and *unsure* pops up immediately for further selection. The *unsure* label is provided to give performers the opportunity to indicate that a tweet is very difficult to classify. This option allows annotators to acknowledge cases where the content may be ambiguous, lack sufficient context, or fall into a gray area that does not clearly align with predefined categories. According to the work by Ross et al. (2017), providing basic definitions and detailed task descriptions for the annotation project beforehand significantly improves the alignment of annotators' opinions on the class labels. This preparatory step ensures that all annotators have a clear understanding of the criteria and expectations, leading to more consistent and reliable annotations. We presented the annotation guidelines to provide a complete description of the annotation task. Additionally, two training task pools, structured in the same way as the actual task, were made available for Toloka performers to complete before joining the main annotation task. These procedures ensure that Toloka performers gain sufficient knowledge and understanding of the annotation task, leading to more accurate and consistent annotations.

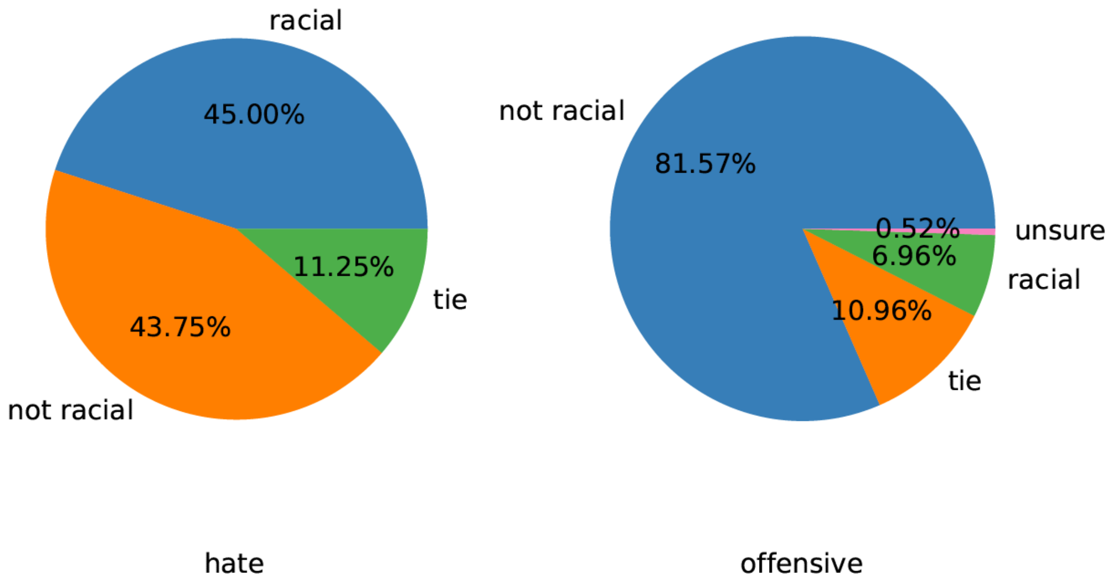
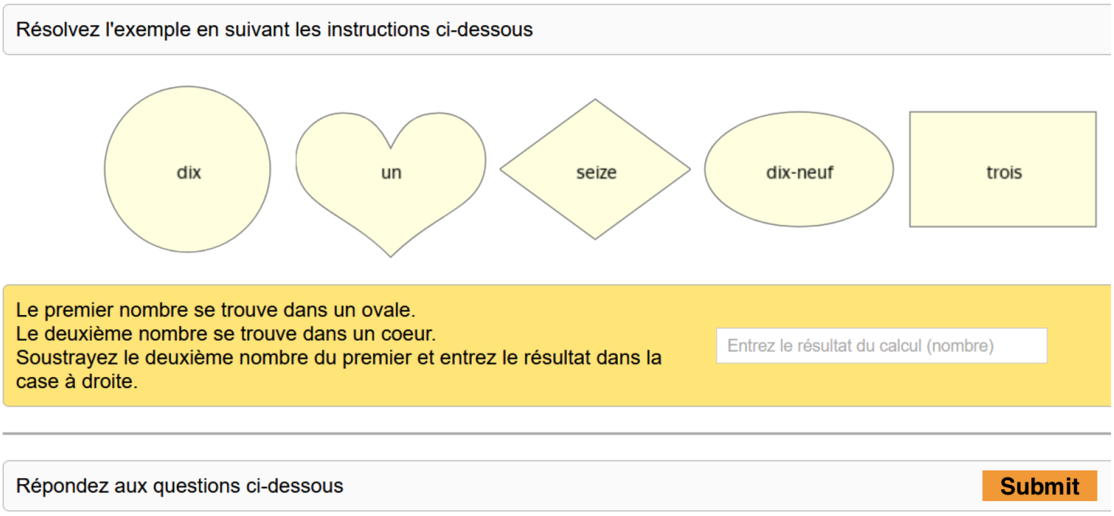


Figure 3.5: Class distributions of our French racial dataset


Résolvez l'exemple en suivant les instructions ci-dessous

Le premier nombre se trouve dans un ovale.
Le deuxième nombre se trouve dans un coeur.
Soustrayez le deuxième nombre du premier et entrez le résultat dans la case à droite.

Entrez le résultat du calcul (nombre)

Répondez aux questions ci-dessous

Figure 3.6: French language test example presented for performers

3.3.5 Mitigating the Challenges of Crowdsourcing Racial Hate Speech Annotations in French

One of the main challenges of crowdsourcing data annotation is the prevalence of malicious data annotators who merely participate in the annotation task to gain financial rewards (Öhman, 2020; **Ayele**, Dinter, et al., 2022). In order to prevent potential malicious performers from engaging in the annotation task, we prepared a French language test and presented it to each performer as indicated in Figure 3.6. Toloka performers needed to pass the French language test in order to participate in the main French racial hate speech annotation task. We also limited the location of performers and allowed those performers who lived in France or Belgium. The performers who successfully completed the two training task pools, lived in France or Belgium, and passed the French language test were qualified and provided the privilege to access the main annotation task pools. A Fleiss' kappa of 0.3 inter-annotator agreement, which indicated a fair agreement, is

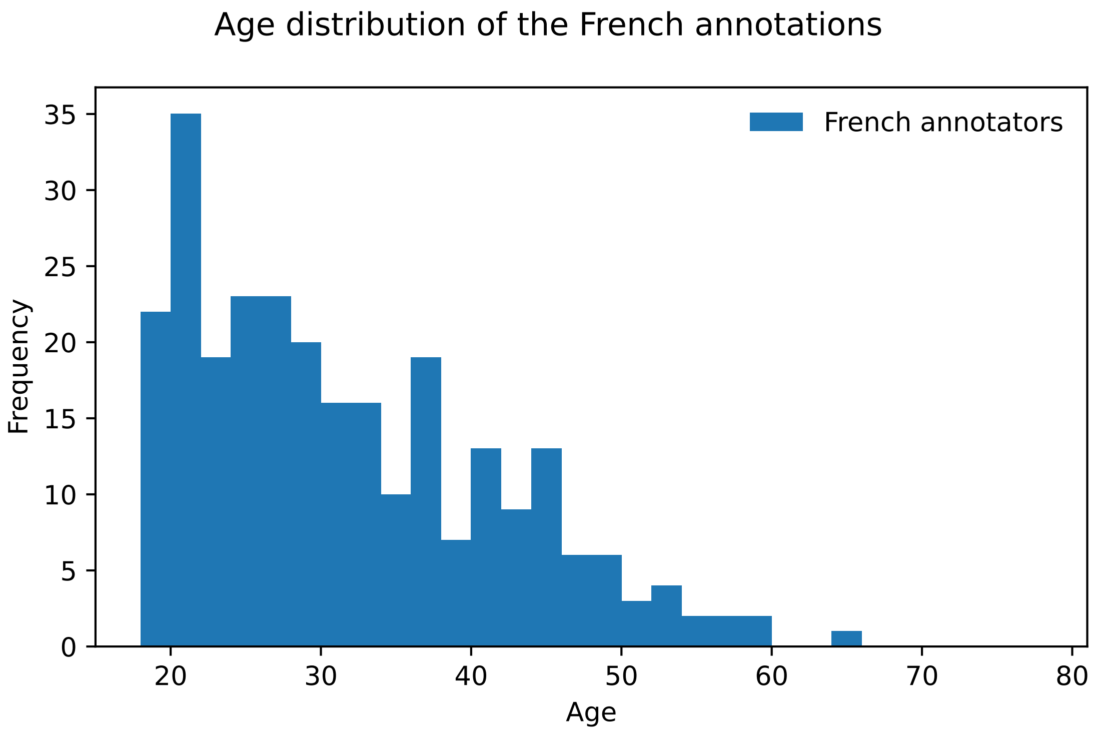
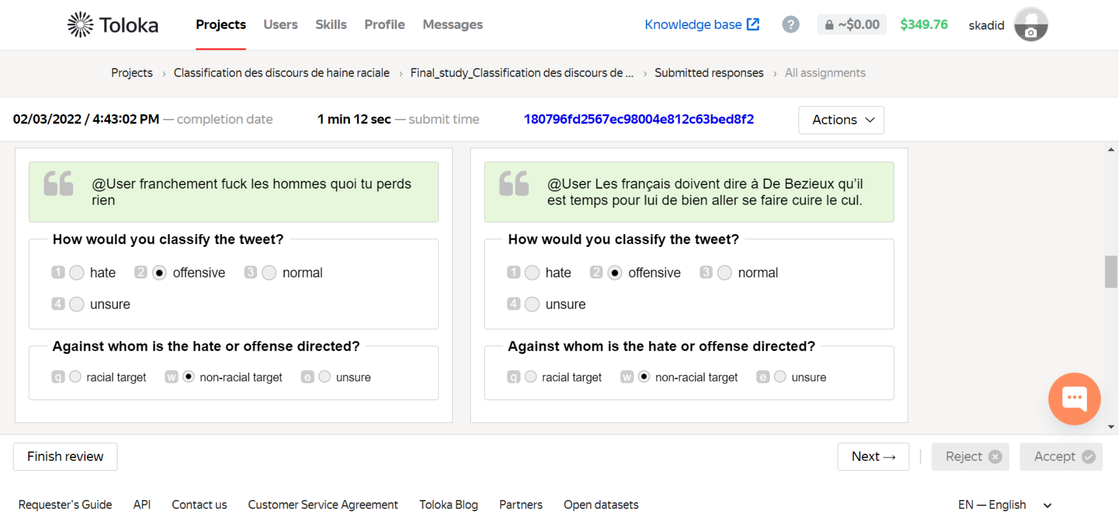


Figure 3.7: The age distribution of the annotators.



02/03/2022 / 4:43:02 PM — completion date

1 min 12 sec — submit time

180796fd2567ec98004e812c63bed8f2

@User franchement fuck les hommes quoi tu perds rien

How would you classify the tweet?

1 ○ hate 2 ● offensive 3 ○ normal

4 ○ unsure

Against whom is the hate or offense directed?

q ○ racial target w ● non-racial target e ○ unsure

@User Les français doivent dire à De Bezieux qu'il est temps pour lui de bien aller se faire cuire le cul.

How would you classify the tweet?

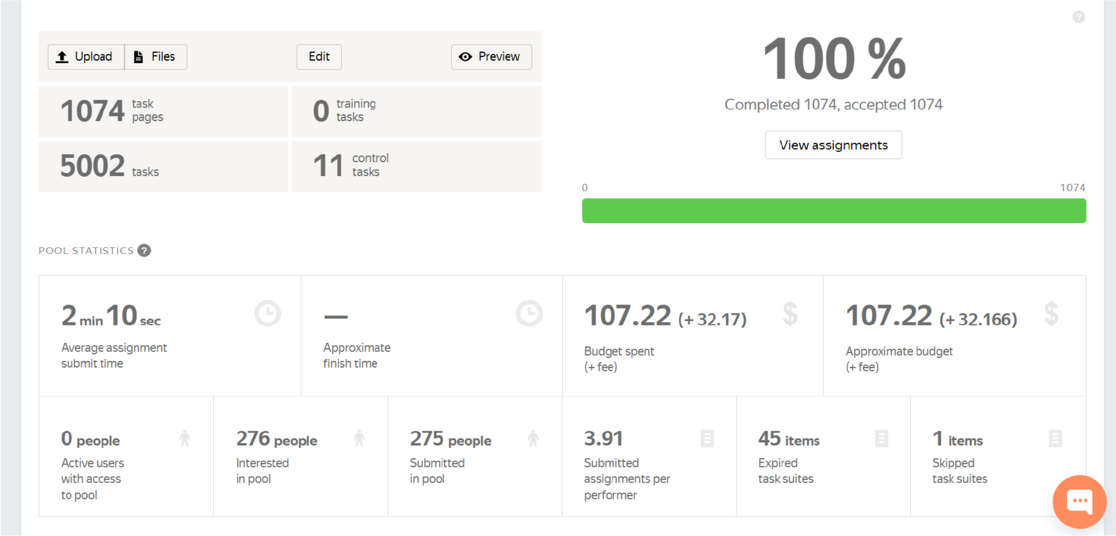
1 ○ hate 2 ● offensive 3 ○ normal

4 ○ unsure

Against whom is the hate or offense directed?

q ○ racial target w ● non-racial target e ○ unsure

Figure 3.8: Example of the French annotation task.



1074 task pages

0 training tasks

5002 tasks

11 control tasks

100 %

Completed 1074, accepted 1074

2 min 10 sec Average assignment submit time

— Approximate finish time

107.22 (+ 32.17) Budget spent (+ fee)

107.22 (+ 32.166) Approximate budget (+ fee)

0 people Active users with access to pool

276 people Interested in pool

275 people Submitted in pool

3.91 Submitted assignments per performer

45 items Expired task suites

1 items Skipped task suites

Figure 3.9: Completed French annotation project.

Fleiss' Kappa score	0.3
Total number of Annotated tweets	5002
Number of annotators participated in the task	275
Mean age of annotators in years	31.11
Country distribution of annotators	265 Fr, 8 Be, 3 O
Accuracy for 50 random tweets	0.24
F1 score for 50 random tweets	0.24
Racial accuracy for 50 random tweets	0.12
Average time for 15 tweets	2 min 10 sec
Number of collected keywords	3473

Table 3.6: Basic annotation information (Fr= French, Be = Belgium, O = Others)

achieved. The gold labels for each instance is determined by aggregating the results from the three annotations with a majority voting scheme.

As indicated in Figure 3.5, 45% of the tweets that are annotated as hate speech contains racial content, while another 11.25% have also ties with racism. Hateful tweets had more probability to contain racial content and ties than offensive tweets. Figure 3.7 shows that the majority of Toloka performers who have participated in the French racial hate speech annotation were young adults below 40 years. Besides, he summary of the overall annotation information is presented in Table 3.6. Out of the 275 participants, 265 performers were from France alone, indicating that a significant majority of the performers, with over 96% participated from France. Moreover, the sample annotation task presented to Toloka performers for annotation is depicted in Figure 3.8, and the completed French racial Toloka project indicating the overview of the French racial hate speech annotation project is also provided in Figure 3.9. Each annotator earned $0.1 per task.

3.3.6 Experiments

Baseline Models

The BERT (Bidirectional Encoder Representations from Transformers) language model has been extensively used in natural language processing tasks. It consists of transformer encoder layers with a self-attention mechanism (Devlin et al., 2019). The model has grown into a family of language models for a wide range of languages. The multilingual BERT and CamenBERT models are examples of such extensions. The works like HateXplain (Mathew et al., 2021), further fine-tuned the models with hate speech dataset collected from posts on Twitter[7] and Gab[8], which were filtered with keyword lists. The dataset was constructed for English and accommodated rationales to better explain the decisions of the crowd workers who annotated the posts. The HateXplain (Mathew et al., 2021) model achieved an accuracy of 70% and an F1 score of 69% on this dataset.

Our task in Section 3.3 utilizes CamemBERT (Martin et al., 2020) and HateXplain (Mathew et al., 2021) to fine-tune and adapt these models to our new dataset.

HateXplain is a BERT-based model for detecting hate speech and providing explanations for its predictions. It identifies hate, offense, and normal speech categories to build models by improving explainability in hate speech detection (Mathew et al., 2021). On the other hand, CamemBERT is a French-specific language model based on the BERT architecture, designed for natural language understanding tasks. It was trained on a large corpus of diverse French texts, enabling it to handle the behavior of the French language effectively.

The HateXplain dataset was used for fine-tuning the BERT models, which are pre-trained for a wide range of language processing tasks. It was further preprocessed and applied for fine-tuning the multilingual BERT model. Additionally, the dataset was translated with Google Translate to French and trained on the French language model camemBERT[9].

We conducted different experiments by fine-tuning the HateXplain model with the multilingual BERT (ML BERT) and CamemBERT models on different datasets and class label generations. As indicated in Table 3.7, the first four experiments focused on the ML BERT and HateXplain model combinations (i.e., 1.0, 1.1, 1.2, and 1.3) while the next four experiments focused on the CamemBERT and HateXplain model combinations (i.e., 2.0, 2.1, 2.2, and 2.3). We analyzed the influence of different kinds of datasets and label aggregations on the performance of the models as shown in Table 3.7. One of them is the automatic aggregation of the three annotations for each tweet based on the Dawid-Skene aggregation method (Toloka, 2024) Opposed to automatic aggregation, some studies were conducted with a custom aggregation method that combines the votes in the following way: the classifications with at least two votes were considered the ground truth for each tweet. When there are three different classifications, the tweet is either removed (Experiment 1.1 and 2.1) or if there is at least one hateful label, it is considered hateful and otherwise offensive (Experiment 1.3 and 2.3) as shown in Table 3.7.

[7] Twitter: https://twitter.com

[8] Gab Social Network: https://gab.com

[9] CamemBERT: https://huggingface.co/camembert-base

Exp.	Pretrained Model	Label generation	Accuracy	F1 score	Ties	Training time
1.0	ML BERT	HateXplain	51.0%	41.0%	-	12m 47s
1.1	ML BERT+ HateXplain	self aggregated	84.0%	77.0%	no ties	3m6s
1.2	ML BERT+ HateXplain	Dawid Skene	78.0%	69.0%	automatically	4m3s
1.3	ML BERT+ HateXplain	self aggregated	65.0%	51.0%	if hate: hate, otherwise offensive	4m9s
2.0	camemBERT	HateXplain	59.2%	57.0%	-	10m45s
2.1	**HateXplain on camemBERT**	**self aggregated**	**88.8%**	**86.0%**	**no ties**	**3m19s**
2.2	HateXplain on camemBERT	Dawid Skene	80.6%	75.0%	automatically	3m54s
2.3	HateXplain on camemBERT	self aggregated	72.6%	67.4%	if 1 hate:hate, otherwise offensive	3m12s

Table 3.7: Studies for building a French hate speech detection model based on different BERT models and datasets. (Exp: Experiment)

Experiment	Accuracy	F1 score	Epochs	Learn. rate
2.1 a)	88.6%	85.9%	3	5e-5
2.1 b)	89.9%	88.2%	2	5e-5
2.1 c)	88.8%	87.6%	1	5e-5
2.1 d)	88.2%	86.9%	4	5e-5
2.1 e)	85.2%	78.4%	3	5e-4
2.1 f)	**89.2%**	**86.9%**	3	**5e-6**
2.1 g)	**89.2%**	**87.4%**	4	**5e-6**

Table 3.8: Further experimental results based on Experiment 2.1 of Table 3.7

Results and Discussion

For both of the BERT-based models, the datasets performed nearly similar results, as shown in Table 3.7. Hence, the model based on the Dawid Skene aggregation gained a better accuracy and F1 score than the aggregation based on the ones with a majority voting for both the multilingual BERT and camemBERT. The removal of the votes with ties has led to the best results for both base models. This implied that adding ties does not lead to better results. Experiments on the multilingual BERT such as Experiment 1.1 in Table 3.7 performed worse than the corresponding camemBERT (Experiment 2.1). This indicated that augmenting target datasets with translated English datasets like the HateXplain can improve the performance of the BERT modes.

The offensive tweets were predicted well but some normal tweets were also classified as offensive. There were remarkable differences between the performance of the models based on the multilingual BERT and the French camemBERT. Whilst the multilingual BERT always predicted *normal* as the class label with nearly the same score for every tweet, the camemBERT labeled the tweets appropriately. The multilingual experiments achieved a lower score than the camemBERT models. A random sample of 50 tweets that were incorrectly classified by the model was analyzed together with the reasons for the incorrect classification.

Even though all three annotators reached a consensus of 100% agreement on the labels for certain tweets, the classification model displayed variability, leading to inaccuracies in some classifications. For example, although all annotators identified specific tweets as hateful, none of these were classified as such in the test set. This discrepancy can be attributed to the inherent class imbalance within the original dataset, where certain classes, such as hate speech, may be underrepresented compared to others. Addressing this imbalance is crucial for improving the model's ability to correctly identify and classify instances of hate speech in future applications.

Through additional fine-tuning, we selected the best-performing model and experimented with varying hyperparameters such as the number of epochs and learning rate, as detailed in Table 3.8. Recognizing the dataset's imbalance across classes, we implemented a stratified splitting approach for both the training and test sets as an additional experiment. This stratified approach aimed to mitigate the effects of imbalance and resulted in noticeable improvements in the models' overall performance. This iterative process underscored the importance of hyperparameter optimization and appropriate data handling techniques in enhancing the effectiveness of the classification model, particularly in contexts with skewed class distributions.

3.3.7 Conclusion

Section 3.3 presented a crowdsourcing-based racial hate speech dataset and fine-tuned the BERT-base model (HateXplain and camemBERT) for French language. A total of 5k tweets are annotated as hate, offensive, normal, and unsure using Toloka. Furthermore, hate and offensive tweets were labeled as racial, non-racial, and unsure classes. This dataset can be used as a benchmark dataset for French racial hate speech research. The BERT model is successfully fine-tuned with the dataset together with the translated HateXplain dataset. Our experiment achieved an accuracy of 88.8% and an F1 score of 86% which are improving over the baseline HateXplain model.

3.4 Summary of Crowdsourcing for Hate Speech Annotations

In Chapter 3, we utilized and tested crowdsourcing as a potential data annotation technique in both low-resource settings, example for Amharic, and high-resource language contexts, such as French. This chapter has explored the challenges and opportunities of crowdsourcing specifically focusing on hate speech data annotations, sourced from social media, X/Twitter. The findings presented in Section 3.2 and Section 3.3 highlighted the challenges and opportunities of crowdsourcing for hate speech annotations in both low-resource and high-resource language settings, respectively.

The challenges of crowdsourcing hate speech dataset annotations presented in Section 3.2 and Section 3.3 include limited or lack of quality control over diverse annotators, limitations in managing biases and subjective interpretations of hate speech annotations, difficulty of properly controlling malicious annotators, and lack of formal training for annotators. Besides, finding proficient annotators on the crowd is very challenging particularly for low-resource languages. Overcoming these challenges is crucial for ensuring the reliability and consistency of annotations in a crowdsourcing setup.

However, the studies presented in this chapter also explored the opportunities offered by crowdsourcing platforms, particularly within the task of hate speech studies in both low-resource and high-resource languages settings. These opportunities incorporate:

- Accessing diverse perspectives on hate speech interpretations.
- Establishing scalability of annotating large scale datasets.
- Ensuring efficiency and effectiveness of hate speech annotations with reasonable cost and time.

Properly addressing the challenges and leveraging the opportunities of crowdsourcing annotations provide valuable options for collecting datasets of diverse perspectives, which are required to effectively combat hate speech on social media.

"Data is the nutrition of artificial intelligence. When an AI eats junk food, it's not going to perform very well."

— Matthew Emerick

4

Investigating Hate Speech Using In-House Annotations

Contents

4.1 Introduction

Ensuring quality datasets for different NLP tasks requires sufficient time and resources for collecting and annotating the data (Rae et al., 2021). Dataset quality plays a significant role in determining the performance of machine learning models (Bhadauria et al., 2024). While high-quality datasets lead to models that are more accurate, robust, and capable of generalizing well to new, unseen data instances, poor-quality datasets can significantly hinder the ability of models to learn effectively and efficiently (Kern et al., 2023).

In-house annotation can provide high-quality datasets specially for studying hate speech, when compared with crowdsourcing annotation approaches. The task of hate speech annotation is very subjective and context-dependent, as it is influenced by various individual and communal factors such as demographics, social norms, and cultural backgrounds (Leonardelli et al., 2021; Waseem and Hovy, 2016). This approach employs clear and detailed guidelines, face-to-face training of annotators, better supervision and guidance, efficient data quality control measures, and proper ethical considerations. Datasets created in such methodologies ensure the reliability and consistency of the annotation results, which further ensures to train efficient and effective machine learning models for automated hate speech detection (Fišer et al., 2017).

In Chapter 3, we presented hate speech datasets collected through crowdsourcing annotation approaches and analyzed the challenges in such annotation frameworks. We also associated one of the reasons for lower performances of the machine learning models presented in Chapter 3, Section 3.2, with the low quality Amharic hate speech dataset collected through crowdsouring annotation scheme.

In (**Ayele**, Yimam, et al., 2023), we mainly emphasis on collecting hate speech datasets through in-house data annotation approach, which utilizes native Amharic speaker expert annotators who received sufficient training, detailed annotation guidelines and better supervision throughout the annotation processes.

In-house annotation has the capacity to provide high quality and consistent datasets, specially good for highly sensitive tasks, such as hate speech, that involve subjectivity. It utilizes expert level annotators, close supervision and immediate feedback, onsite training of annotator, and detailed annotation guidelines, which allows annotators to capture contextual backgrounds of the messages.

Despite its advantages, in-house annotation requires high cost annotations, longer time, active supervisor and infrastructure facilities for conducting the annotation task (Leonardelli et al., 2021; Fišer et al., 2017). Additionally, the limited availability of human expert annotators and the slower processing speeds make scalability of such annotation approaches challenging.

This chapter discusses the identification of suitable data collection methods and sample selection for the construction of hate speech datasets, along with an analysis of the primary challenges associated with the annotation and detection tasks for Amharic hate speech.

This chapter, presented benchmark hate speech data sampling strategies, a dataset consisting of over 15.1k annotated tweets, and various classification models. This work has the following main contributions:

- **A well-defined hate speech data selection and preprocessing pipeline**: we have developed a systematic pipeline for selecting and preprocessing data for hate speech annotations. This pipeline ensures that the data is representative and free from noise, which is crucial for accurate annotation. The preprocessing steps include filtering irrelevant content such as retweets, near duplicate tweets, deleted tweets, tweets written other than Amharic, mixed scripts, and handling special characters commonly found in tweets.

- **The collection of benchmark hate and offensive speech lexicon entries**: to ensure a dataset with relatively balanced labels that can aid in the detection and annotation of hate speech, we have compiled a lexicon of terms commonly associated with hate and offensive speech in Amharic. The lexicon list includes 65 offensive and 102 hate keywords, serving as a valuable reference for researchers by offering a standardized set of terms to aid in the identification of hate speech in Amharic.

- **The development of hate speech annotation guidelines and strategies**: we have established comprehensive guidelines and strategies to ensure the quality and consistency of hate speech annotations. These guidelines cover various aspects of annotation, including the definition of hate speech, contextual considerations, and the use of clear examples representing each category label, namely, hate, normal, offensive, and unsure labels. We also provide training for annotators to ensure they clearly understand the guidelines in general and the task in particular.

- **Releasing a benchmark dataset and classification models for the Amharic hate speech task**: to facilitate further research in this area, we have released a benchmark dataset consisting of over 15.1k annotated tweets, along with the classification models developed and evaluated as part of this work[1]. This benchmark dataset is one of the first of its kind for the Amharic language, providing a valuable resource for researchers and practitioners working on hate speech detection. The accompanying classification models serve as baseline models that can be used for comparison in future studies.

4.2 Data Collection and Preprocessing

The dataset employed in this task is selected from our database repository, which has been extracted from X/Twitter on daily basis. The detailed descriptions of the data collection procedures and techniques are presented in Chapter 2, Section 2.2.

For the task presented in this chapter, we specifically gather the dataset targeting tweets written between October 1, 2020 and November 30, 2021. The dataset comprised of 3.8 million tweets collected over 14 consecutive months. This period was purposefully chosen to reflect the rapidly changing socio-cultural, and political situations in Ethiopia. The following events collectively contributed to a period of intensified tensions and complexities in Ethiopia, affecting both its internal stability and global relations. The events include, but not limited to the following:

1. **The start of conflict in the north**: The conflict began when Tigray People's Liberation Front (TPLF) rebel groups attacked the northern command of the Ethiopian army. This act marked the official start of hostilities between the TPLF and the federal government.

2. **The conflict escalated**: The situation between the TPLF and the federal government worsened significantly as the conflict reached a critical point. The TPLF rebels managed to gain control of the neighboring Afar and Amhara regions, intensifying the conflict and expanding its geographic impact.

3. **National election**: Ethiopia held its 6th national election. This event was significant in its own right, given the country's political climate and the ongoing conflict in the north, adding further complexity to the national situation.

4. **The GERD dispute**: The Grand Ethiopian Renaissance Dam (GERD) dispute between Ethiopia and Egypt also reached a critical peak during this period. The disagreement over the dam's impact and operations became so intense that it was

[1] Amharic Hate Speech Resources: https://github.com/uhh-lt/AmharicHateSpeech

brought before the UN Security Council. Despite Ethiopia's insistence that the GERD issue was not a matter of international security, the dispute highlighted significant regional tensions and the need for diplomatic resolution.

4.2.1 Data Sampling

Figure 4.1 illustrates the comprehensive data collection, preprocessing, and sampling strategies employed in Chapter 4. Initially, we began with a dataset of 3.8 million tweets. To ensure relevance and quality of the dataset, we first removed retweets and filtered out non-Amharic tweets using the Python language detection tool[2]. This process significantly reduced the dataset to 902k tweets as indicated in Figure 4.1.

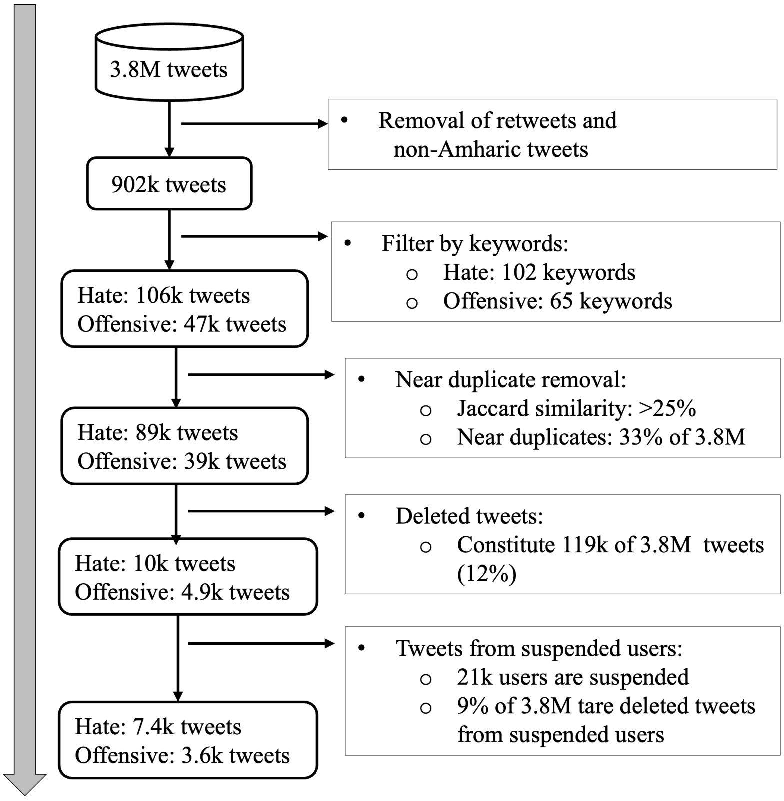

3.8M tweets
• Removal of retweets and non-Amharic tweets
902k tweets
• Filter by keywords:
o Hate: 102 keywords
o Offensive: 65 keywords
Hate: 106k tweets
Offensive: 47k tweets
• Near duplicate removal:
o Jaccard similarity: >25%
o Near duplicates: 33% of 3.8M
Hate: 89k tweets
Offensive: 39k tweets
• Deleted tweets:
o Constitute 119k of 3.8M tweets (12%)
Hate: 10k tweets
Offensive: 4.9k tweets
• Tweets from suspended users:
o 21k users are suspended
o 9% of 3.8M tare deleted tweets from suspended users
Hate: 7.4k tweets
Offensive: 3.6k tweets

Figure 4.1: Data selection and preprocessing pipeline.

Next, we focused on identifying hate speech and offensive content within this refined dataset. By employing a lexicon of hate and offensive terms, we further filtered the tweets, ultimately reducing the target dataset to 153k tweets as indicated in Figure 4.1.. The lexicon entries were essential to ensure a relatively balanced dataset through selecting the specific content of interest for analysis. Figure 4.1 presents a sample of some hate and offensive keywords utilized in the filtering process. Since some of the keywords are extremely offensive, we masked them with the "*" symbol.

The keywords used for this filtering were carefully collected from various sources. We engaged volunteer communities, gathering inputs through Google Forms that were distributed via emails and social media platforms, finally aggregated and curated by human experts. This approach ensured a diverse and comprehensive collection of

[2] https://pypi.org/project/langdetect/

relevant terms/keywords. Furthermore, we employed lexicon entries presented in Yimam et al. (2019) as an initial list of terms/keywords, which provide a robust foundation for the construction of our lexicon collections.

Hate keywords		Offensive Keywords	
Lexicon	Translation	Lexicon	Translation
ቆ*ጣ	Lep*er	አ*ያ	Id*ot
ነፍ*ኛ	Musketeer	ዲ*ላ	Basta*d
ጋ*	Ga**a	መተተኛ	Conjurer
ውሃ*ያ	Wuhabiya	ጅል	Buffoon
ቅማ*ም	Bug**	ገልቱ	Incompetent
ጠባብ	Narrow	ጥ*ብ	carrion
አህ*ሽ	Ah*ash	ደ*ዝ	D*ll
ትምክህተኛ	Sn*b	ደደብ	Id*ot
አ*ሜ	Ag*mie	ሸር*ጣ	S*ut
ወ*ኔ	T*LF	አውሬ	Brutal
ኦ*ግ	O*F	ፈሪ	Runagate
ቡ*	Witch	ጨካኝ	Tyrannical
ጁ*ታ	Jun*a	ሰካራም	Drunker
ኦ*ሙማ	Oro*uma	ባለጌ	Naughty
ሰፋሪ	S*ttler	ጉረኛ	Boaster
መጤ	*mmi*rant	ደንቆሮ	Ignorant

Table 4.1: Sample hate and offensive keywords.

After closely reviewing the filtered tweets, we found that certain tweets, although they had unique IDs, were duplicates or near-duplicates to each other. This phenomenon likely arises from users copying and re-posting tweets with slight modifications.

In response to this issue, we implemented a comprehensive approach to address the presence of duplicate and near-duplicate tweets within the dataset. Initially, we applied shingling techniques in conjunction with the Jaccard similarity index. This method involved assessing pairwise tweet similarities to detect instances of content overlap. As presented in Figure 4.1, we established a rigorous threshold of 25% similarity to differentiate between unique tweets and near-duplicate or closely similar tweets. Tweets that exceeded this threshold were systematically excluded from our dataset to ensure the integrity and uniqueness of the remaining data.

This rigorous filtering process was essential to ensure that our final dataset consisted of predominantly distinct tweets, thereby enhancing the consistency and robustness of our analyses, and mitigating potential biases and inaccuracies that could arise from redundant data samples.

As indicated in Figure 4.1, about 33% of the original tweets in our corpus were identified as near duplicates and thus excluded from our study. This exclusion was crucial for maintaining the quality and validity of our research findings, especially in our examination of hate speech and offensive content within Amharic tweets. Addressing the challenge of near-duplicate tweets early in the data curation process ensures accurate interpretations that capture the diverse and complex dynamics of online discourse.

4.2.2 Dealing with Deleted Tweets

X/Twitter actively removes tweets reported as inappropriate and suspends user accounts for a variety of reasons. As indicated in Figure 4.1, our in-depth analysis of the target dataset revealed that 12% of the tweets originally included in our repository have been deleted and are no longer accessible on X/Twitter. Besides, among these deleted tweets, approximately 9% were contributed by previous X/Twitter users whose accounts have already been suspended.

We manually annotated samples of selected deleted tweets from both active and suspended users as part of our initial pilot investigations. The objective was to assess whether these deleted tweets contained a higher prevalence of hateful content compared to tweets that remain accessible.

4.3 Data Annotation

Previous studies on Amharic hate speech classification such as Mossie and Wang (2018, 2020), Getaneh (2020), Tesfaye and Kakeba (2020), and Abebaw et al. (2022b, 2022a) identified binary categories (i.e. hate or non-hate). Studying hate speech as a binary class problem might have overlooked its complex and multi-faced characteristics. However, studies for other languages like English utilized more categories such as *hateful*, *offensive*, and *normal* class labels (Davidson et al., 2017; Mulki et al., 2019).

The study conducted by Mathew et al. (2021) introduced additional category, called *unsure*, resulting in a four class label category, which include *hate*, *offensive*, *normal*, and *unsure*. We have used the *WebAnno*, a web-based annotation framework, which is designed for diverse types of NLP annotations (Yimam et al., 2013).

4.3.1 Pilot Annotation

In the initial round of pilot annotation, we carefully annotated over 3k tweets that contained hate and offensive keywords. As presented in Table 4.2, this pilot data annotation focused primarily on tweets from three distinct categories: accessible tweets, deleted tweets from suspended users, and deleted tweets from active users. The accessible tweets were those still available on the platform, while the deleted tweets included those removed either because the users were suspended or by the users who remained active. Considering datasets from these three distinct categories of tweets helps us to clearly understand the nature and distribution of hate and offensive content across different states of tweets.

Each tweet is annotated by three native speakers. While the first two annotators labeled each tweet independently, the third annotator who served as a curator or an adjudicator made the decisions on the final gold labels and resolve disputes. A total of 5 annotators were involved in the pilot annotation task and each annotator earned 0.5 ETB or $0.01 cents per tweet. The annotators can label 150 tweets per hour and earn 75 ETB or $1.5, which is nearly equivalent to the hourly wage of B.Sc. degree holders in Ethiopia.

Category	Number of samples
Accessible tweets containing keywords	956
Deleted tweets from suspended users containing keywords	1002
Deleted tweets from active users containing keywords	1054
Total number of annotated tweets in pilot study	**3012**

Table 4.2: Pilot Annotated Tweets by Category.

We prepare training manuals and annotation guidelines and deliver intensive training to make the task clear for the annotators and the curator. We developed comprehensive training manuals and detailed annotation guidelines, and delivered intensive training sessions to ensure whether the task was clearly understood by both the annotators and the curator. These resources were designed to provide an exhaustive understanding of the annotation process and procedures, and enabling the annotation team to accurately identify and categorize the tweets in to the corresponding labels. The training focused on describing the annotation guidelines covering various aspects of the annotation task, including basic definitions, examples, and best practices, to ensure consistency and precision in the annotations task.

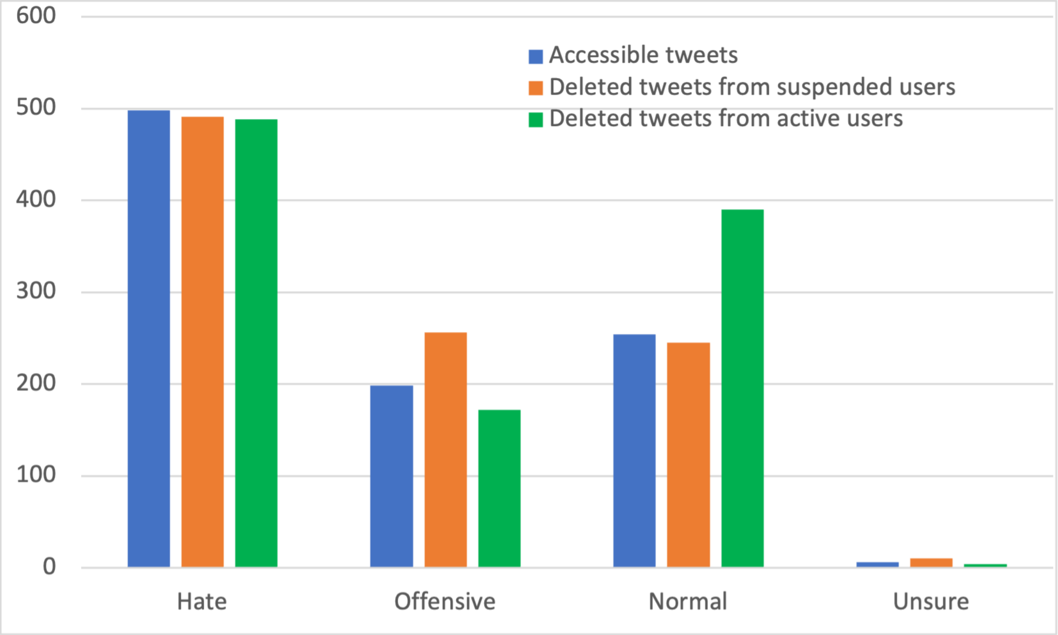


Figure 4.2: Distribution of labels across categories in the pilot study.

The pilot annotation result consisted of 1,476, 625, 883, and 28 tweets labeled as hate, offensive, normal, and unsure class labels, respectively. We employed Cohen's kappa coefficient to compute the inter-annotator agreement (IAA) and achieved a 0.44 agreement score for the pilot annotation. Other related studies, for example, the work by Del Vigna et al. (2017) reported a 0.26 inter-annotator agreement score on the Italian dataset while Ousidhoum et al. (2019) reported 0.153, 0.202, and 0.244 IAA scores of kappa coefficient on English, Arabic, and French datasets respectively. Besides, Mathew et al. (2021) reported a 0.46 inter-annotator agreement score on the English data set, which indicated a moderate agreement among annotators. Therefore, our 0.44 inter-annotator agreement score fell under the moderate category which encouraged us to pursue the main annotation task.

As shown in Table 4.2, hateful tweets appeared to be more prevalent in the dataset. This dominance was evident across all three categories in the pilot annotations: accessible tweets, deleted tweets from suspended users, and deleted tweets from active users. The likely reason for this trend is that the tweets were selected exclusively based on the presence of hate and offensive keywords. This keyword-based selection method naturally led to a higher proportion of hateful content being included in the dataset, reflecting the characteristics of our sampling approach.

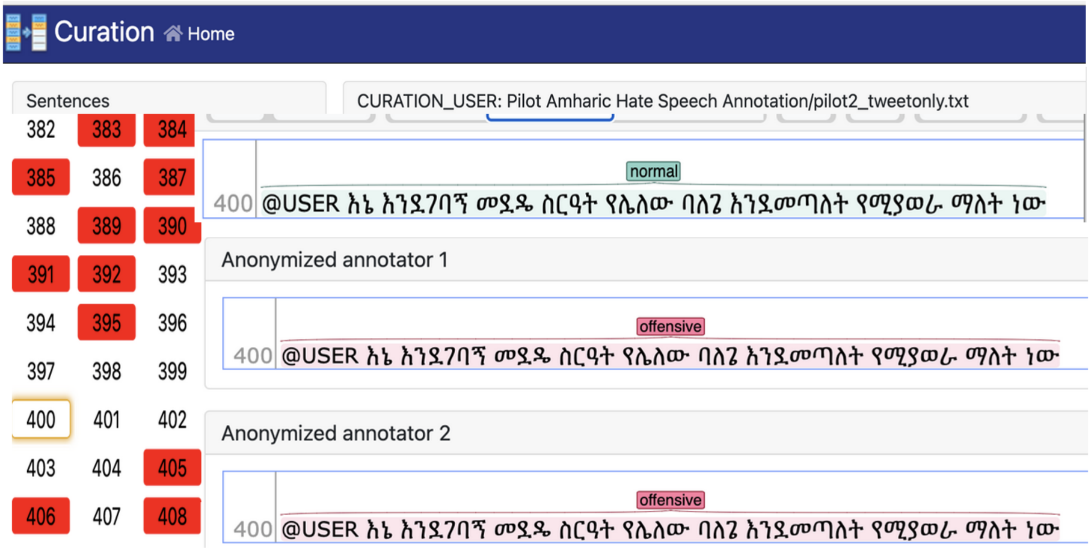
The deleted tweets were examined and compared with the accessible tweets if they contained more hateful content. No significant differences were found in the distributions of hateful tweets across the three categories (accessible tweets, deleted tweets from suspended users, and deleted tweets from active users). The deleted tweets are excluded from being sampled in the final dataset since they are no more available on Twitter.

Despite our preliminary assumption that deleted tweets generally contain more hateful content compared to accessible tweets, our analysis of the pilot annotation results disproved the preliminary assumption. Contrary to our initial expectations, we found that the prevalence of hateful content in deleted tweets were not significantly different from those in accessible tweets. presented in Figure 4.2. Furthermore, the result suggested that deleted tweets can be excluded from being sampled in the final dataset since they are no more available on X/Twitter and they do not have any special interesting features that are distinct from accessible tweets. Therefore, the main annotation task is constructed solely from accessible tweets.

We have finally created two large pools of unlabelled tweets, one containing keywords and the other without keywords. The keyword-based unlabelled pool consisted of around 113k accessible tweets containing hate and offensive keywords. The second unlabelled pool, which is without keywords, is comprised of 757k accessible tweets that do not contain hate and offensive keywords. The tweets are anonymized by replacing usernames with <USER> tokens and removing URLs from the tweets.

4.3.2 Error Analysis of Pilot Annotations

Hate speech annotation is highly subjective and challenging even for human annotators (Fortuna et al., 2022; **Ayele**, Belay, et al., 2022). During the pilot study, we observed disagreement between annotators on their annotation labels due to the complex and subjective nature of hate speech annotation. In some cases, the curator also deviated from both annotators who sometimes selected a different annotation label. Such annotation errors were analyzed with examples as presented in Figure 4.3. Despite hate speech annotation is highly subjective task, we tried to understand the different views of annotators through exploring various expert judgments. Three experts, a lawyer (Assistant professor in Law), a political science expert (Ph.D. student), and a journalism expert (Associate professor of media and communications) were engaged in a focus group discussion to analyze the potential sources of annotation disagreement between the annotators as well as the adjudicator. The experts evaluate the annotation deviations and suggest possible justifications for the source of the disagreement on the labels of those tweets. In general, we observed that hate speech annotation is a highly context-sensitive and challenging task (**Ayele**, Belay, et al., 2022), which usually resulted in lower inter-annotator agreement.



CURATION_USER: Pilot Amharic Hate Speech Annotation/pilot2_tweetonly.txt

400 @USER እኔ እንደገባኝ መደዴ ስርዓት የሌለው ባለጌ እንደመጣለት የሚያወራ ማለት ነው

Anonymized annotator 1

400 @USER እኔ እንደገባኝ መደዴ ስርዓት የሌለው ባለጌ እንደመጣለት የሚያወራ ማለት ነው

Anonymized annotator 2

400 @USER እኔ እንደገባኝ መደዴ ስርዓት የሌለው ባለጌ እንደመጣለት የሚያወራ ማለት ነው

Figure 4.3: Sample deviations between annotators and the adjudicator taken from WebAnno (Yimam et al., 2013).

As shown in Figure 4.3, a tweet that has been labeled by two annotators (on the bottom side) and a curator (the top side) is presented within the WebAnno annotation tool's graphical user interface. On the one hand, the two annotators agreed that this tweet (translated in English here) "*as I understood it, 'Medede' means a crazy, naughty and disrespectful person who speaks randomly*" is an *offensive* instance. The reason for labeling decisions of the two annotators could be attributed to their perception that the tweet is directed at the user mentioned with '@USER'. They likely interpreted the content as containing offensive targeted towards that specific individual due to the '@USER' mention.

On the other hand, the curator classified the tweet as *normal*, because the curator interpreted the author's intent as explaining the meaning of the word 'Medede' rather than directing any form of hostility or offense towards an individual. The decision of the curator/adjudicator is likely influenced by the understanding of the context and linguistic variations within the tweet, focusing more on the informative or descriptive aspect rather than perceiving it as containing offensive or derogatory content towards a specific person.

The red colored numbers (on the left side) in Figure 4.3 showed that the two annotators disagreed on that item label while the tweets shaded with light red and light cyan colors (right side) represented the annotator's and curator's decisions, respectively. In cases where annotators encountered tweets written in a combination of languages other than Amharic, or tweets that were complex and difficult to understand for a variety of reasons, they generally labeled these tweets as "*Unsure*". The "Unsure" label served as a placeholder for tweets that require further review or additional context to be properly understood by annotators.

4.3.3 Main Annotation

The pilot annotation indicated that the selection from the lexicon-based unlabelled pool suffered from data imbalance problems. Therefore, we mixed the lexicon-based unlabelled pool with the non-lexicon-based pool on a 70/30 proportion. Each batch of annotations comprised 70% from the keyword-based unlabelled pool and the remaining 30% from the other pool, with no keywords, respectively. The overall annotation process of the dataset including the pilot study took over a period of a year. We performed the pilot annotations in 6 batches and the main annotations in 22 batches, where we analyzed each batch before pursuing the subsequent batch.

The annotators were nominated from different cultural, religious, gender, and age categories, and each user annotated from 3,800-4500 tweets. A Cohen's kappa score of 0.48 is achieved on a dataset of over 15.1k tweets on the main annotation task which is better than the pilot task. As indicated in Figure 4.4, hateful tweets are dominant in the dataset, which accounts about 44% of the total annotated instances. The 86 tweets annotated as "unsure" were further examined with expert consultations to explore the sources of annotation decisions. Since the majority of the tweets labeled "unsure" contained mixed languages of non-Amharic words that confused annotators, they were excluded from being used in the experiment.

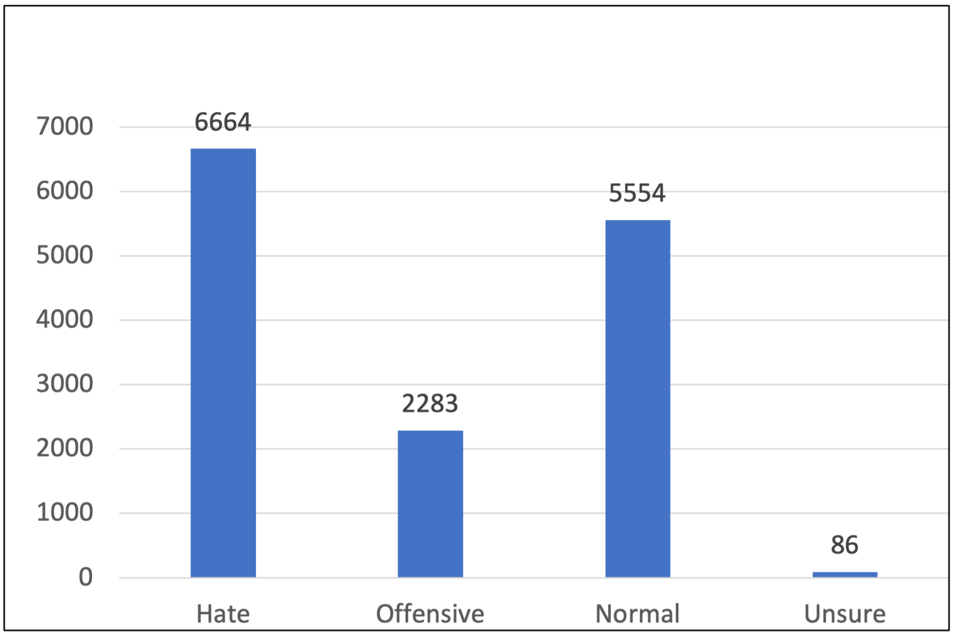


Figure 4.4: Distribution of class labels in the overall annotated dataset.

4.4 Experimentation

We employed the 80:10:10 data split mechanism for creating the train, development, and test instances. We have used the development dataset to optimize the learning algorithms. All the results reported in the remaining sections are based on this test dataset instances. Deep learning algorithms are computed using the following hyper-parameters, *embedding dimension = 100, epochs = 10, batch_size = 64, activation = softmax, and optimizer = adam.*

We conducted experiments employing the classical machine learning models such as LR, LSVM, and NB through utilizing BOW, TF-IDF, and n-gram feature extraction methods. We also explored the deep learning models like RNN, LSTM, BiLSTM; and CNN, and the fine-tuned Amharic transformer models such as AmFLAIR and AmRoBERTa.

Model performance evaluation metrics such as F1 score (F1), Precision (P), Recall (R), and Accuracy (Acc) are utilized to assess and compare the effectiveness of the models implemented in the study.

4.5 Results and Discussion

As presented in Table 4.3, logistic regression (LR) achieved 67% F1 score and 68% performance for precision, recall, and accuracy. LSVM achieved a 68% precision score, and 67% recall, accuracy, and F1 scores. Naïve Bayes obtained the least F1 score which is 63% from all classical methods. LR and LSVM outperformed Naïve Bayes in all measures except for precision. LSTM, BiLSTM, RNN, and CNN achieved lower and nearly similar results in all measures of precision, recall, accuracy, and F1 scores. We attribute this to the size of the dataset; while it is common sense that deep learning approaches can achieve higher results by better modeling the properties of large training data, it seems that our dataset was not large enough to leverage their power. The Am-FLAIR contextual embedding model achieved 72% scores for all measures such as precision, recall, accuracy, and F1 scores, which is the overall best result in our experiments. AmRoBERTa also achieved 70% precision, recall, accuracy, and F1 scores, which are the second-best scores. In general, the contextual embedding models such as AmFLAIR and AmRoBERTa outperformed both the deep learning and the classical machine learning methods in all performance measures on the dataset. This confirms the general trend of well-performing transformer-based language models also for the case of Amharic.

Classifier	Prec	Rec	Acc	F1
LR	68%	68%	68%	67%
SVM	68%	67%	67%	67%
NB	68%	65%	65%	63%
RNN	61%	62%	62%	62%
LSTM	61%	62%	62%	61%
BiLSTM	61%	62%	62%	61%
CNN	62%	63%	63%	62%
Am-FLAIR	**72%**	**72%**	**72%**	**72%**
Am-RoBERTa	70%	70%	70%	70%

Table 4.3: Performance of the models.

4.6 Error Analysis from Model Outputs

We examined model-predicted tweets against their corresponding gold labels to observe discrepancies within the result. As indicated in Table 4.4, the model correctly classified 1,034 tweets out of 1,501 test examples.

As shown in Table 4.4, the model could correctly classify *hate* class label better than the other class labels, achieving over 76% accuracy score. On the other hand, the model faced sever difficulties to accurately classify the *offensive* class label, which might be due to its smaller representations in the dataset. The accuracy of the model to correctly classify offensive tweets is around 50% accuracy, as it can be computed in the confusion matrix presented in Table 4.4.

		PREDICTION			
		Hate	Offensive	Normal	Total
GOLD	Hate	**516**	85	101	702
	Offensive	63	**154**	47	264
	Normal	104	67	**364**	535
	Total	683	306	512	**1501**

Table 4.4: Confusion matrix from FLAIR.

We randomly choose 25% of the incorrectly classified instances and conducted extensive investigations in a focus group discussion with three domain experts to explore the potential reasons for the classifications errors on model outputs. The descriptions presented in Figure 4.5, 4.6 and Table 4.5 are based on these expert evaluations.

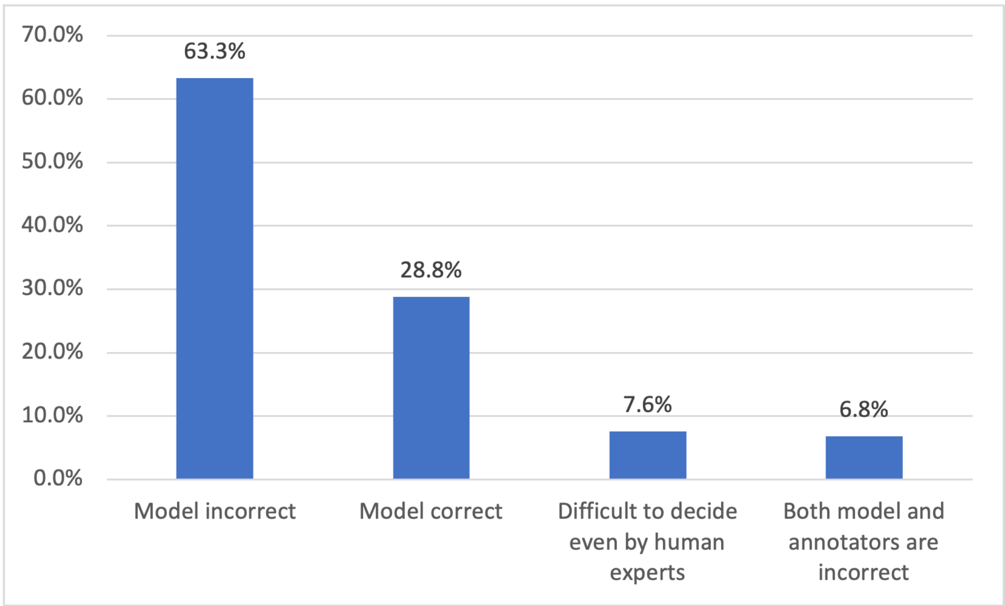


Figure 4.5: Statistics for incorrectly classified instances.

As indicated in Figure 4.5, 63.6% of the errors are mistakes by the model while 28.8% of errors are due to annotator mistakes. Besides, the experts found that the remaining 7.6% errors are difficult to judge due to a lack of background contexts.

As dipicted in Figure 4.6, we have conducted an exhaustive analysis to explore the potential sources of errors within the misclassified instances. We found out that the main reasons for the errors are annotation bias, association with some keywords, lack of background contexts, informal writing styles in social media, mixed language use, the presence of sarcasm, and idiomatic expressions. Annotation bias, presence of sarcasm, association with some keywords, and the lack of background contexts constituted 29.7%, 13.6%, 11%, and 8.5% of the causes for the errors, respectively.

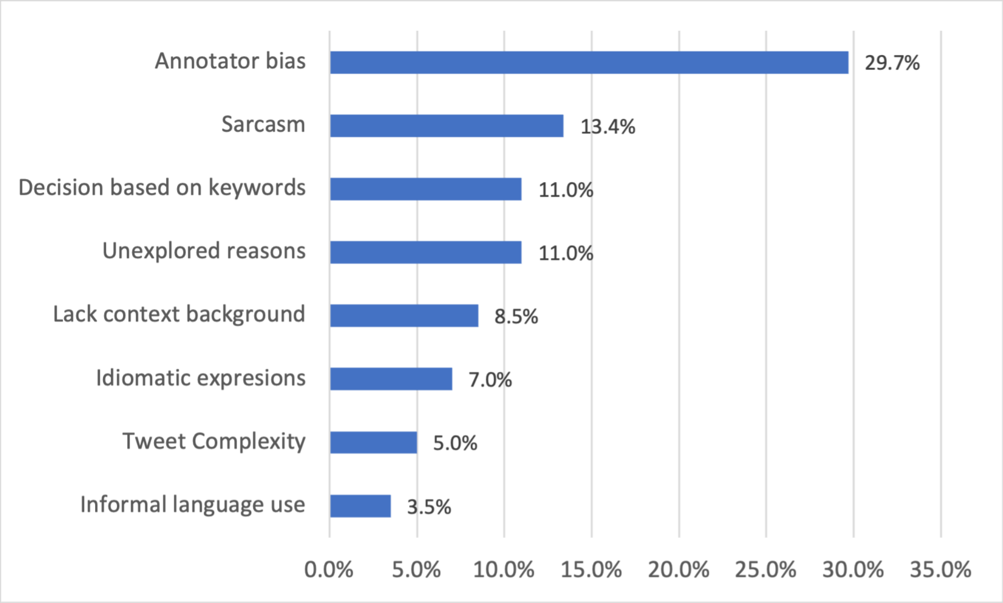


Figure 4.6: Sources of errors for misclassifications.

Besides, this analysis further explored some special difficult scenarios that challenged even for human experts. As indicated in Figure 4.6, human expert evaluators could not come up with justifications for the cause of some errors due lack of background contexts to evaluate such tweets against the annotators and model outputs.

In order to showcase the potential possible justifications for the source of errors, we to took 5 tweets as presented in Table 4.5, with subject expert consultations. This table presents annonymized original twetets Tweets with ironic/sarcastic expressions can even confuse human annotators. For example, *Tweet 1* in Table 4.5 with the gold label 'offensive', targeted an individual with sarcasm expression and is wrongly predicted as 'normal' by the model. *Tweet 2* annotated as 'hate' is wrongly predicted as 'normal' by the model. This is due to typographic errors in the tweet such as missing characters and unnecessary spaces between characters that we indicated with the '-' symbol. The '*' symbols are used to hide sensitive words from the tweets.

	Tweet	English Translation	Gold	Predicted
1	እውነትም በእድሜ ትንሹ መሪ?	Oh, really the youngest leader?	offensive	normal
2	የ**ስ ዜ*ዬ አያት በቀኝ በኩል የቆመው ባንዳ	M*l*s Z*n*wi's grandfather, the batrayer, standing on the right.	hate	normal
3	በኦሮሚያ ክልል የተደረገው የድጋፍ ሰልፍ የኦሮሞ ብልፅግና እና ኦነግ ሸኔን አንድነት ያሳየ ነው ተባለ።	The rally in Oromia showed the unity of prosperity party and OLF-Shenie.	hate	normal
4	@USER ከወራሪ ጋር ሽምግልና የለም። አምሽክ ነው።	@USER No mediation with the invaders, just destroy them.	hate	normal

Table 4.5: Model errors: wrongly predicted tweets against the gold labels.

Despite *Tweet 3* looking positive news, it contained ironic expressions that the model did not predict correctly. But annotators knew the additional background contexts to understand and label the tweet. *Tweet 4* with gold label 'hate' is wrongly predicted as 'normal' by the model due to the inclusion of informal terms that are not used in the standard Amharic writing system that could confuse the model.

4.7 Conclusion

This chapter presented typical data selection and annotation strategies for the Amharic Twitter dataset. A total of 15.1k tweets were annotated into hate, offensive, normal, and unsure classes. We proposed data selection and sampling strategies, a list of hate and offensive lexicon entries, and a relatively large amounts of annotated dataset for Amharic hate speech research. We also presented both classical and deep learning models trained on a new dataset.

The study explored hate speech annotation challenges and revealed that annotation of social media texts for hate speech classification is highly subjective and complex. Hate speech annotation requires diverse contextual background information about a text collected from social media and the author of the text as well at that particular time and situation when the text is written.

Models that have used contextual embedding architectures such as Am-FLAIR and Am-RoBERTa outperformed all the models, with Am-FLAIR achieved the best scores in all performance measures.

"There is a fine line between free speech and hate speech. Free speech encourages debate whereas hate speech incites violence."

Newton Lee (2015)

"Free speech/hate speech–all depends upon one's perspective. but all would agree that it is metaphorical, formulaic, not an actual call for immediate action or slaughter. And we want to think that of those who chant such phrases there may be some/ many who do not entirely understand what the phrases entail & how they strike the hearts of some listeners."

Joyce Carol Oates (May 6, 2024 on X)

5 Exploring Hate Speech Beyond Binary Categories: Uncovering the Complex Intensities and Targets

Contents

5.1 Introduction

In previous chapters, we have explained that a substantial body of research studies has been conducted to develop automatic hate speech detection systems over the last decades, despite most of these attempts focused on classical hate speech classification approaches. For instance, the works conducted by Davidson et al. (2017), Waseem and Hovy (2016), Founta et al. (2018), Mathew et al. (2021), Plaza-del-arco et al. (2023), Caselli and Veen (2023), Fortuna et al. (2020), and Clarke et al. (2023) and many other studies considered hate speech as a binary problem. Davidson et al. (2017) and Mathew et al. (2021) identified hate, offensive, and normal class categories while Waseem and Hovy (2016) differentiate each tweet as racist hate speech, sexist hate speech, or neither. Founta et al. (2018) utilized hateful, abusive and offensive labels and Plaza-del-arco et al. (2023) employed hate or non-hate class labels. Despite the studies by Fortuna et al. (2020) and Clarke et al. (2023) explored diverse previous datasets and identified the complexity of the topic to aggregate the labels and propose common class labels, they highlighted that such attempts emphasized hate speech as a discrete binary concept.

Nevertheless, this binary viewpoint lacks the capacity to capture the diverse and context-dependent features of hate speech, focusing on the traditional detection and classification approaches.

In contemporary studies, there has been a recognition of these limitations through promoting a shift towards adopting multifaceted methodologies to gain a better understanding of the nature, dimension, and intensity of hate speech (Beyhan et al., 2022; Sachdeva et al., 2022). This further enhances hate speech detection capabilities and employs more effective mitigation strategies to tackle its propagation on social media and its impact on the physical world.

Similarly, most studies in low-resource languages such as Amharic, predominantly concentrated on detecting and classifying hate speech as a desecrate binary concept, overlooking its varying levels of intensities. For instance, the studies conducted by Abebaw et al. (2022a, 2022b), Mossie and Wang (2018), and Tesfaye and Kakeba (2020), predominantly concentrated on the detection and classification of hate speech as a simple binary concept, hate or non-hate, overlooking its varying levels of intensities and subtle differences. In this regard, our previous studies, such as **Ayele**, Dinter, et al. (2022) and **Ayele**, Yimam, et al. (2023), are similar, with the exception of their contributions of new datasets and the incorporation of offensive and unsure class categories.

Recent studies indicated that hate and offensive speeches are not simple binary concepts, rather they exist on a continuum, with varying degrees of intensity, harm, and offensiveness (Bahador, 2023; Sachdeva et al., 2022). In practical scenarios, hate speech exhibits a wide spectrum, encompassing mild stereotyping on one end and explicit calls for violence against a specific group on the other (Beyhan et al., 2022). Demus et al. (2022) explored hate speech categories, targets, and sentiments in two or three discrete categories while analyzing the toxicity of the message using the Likert scale ratings of 1-5 to show the potential of a message to ”poison” a conversation.

The study by Chandra et al. (2020) investigated the intensity of online abuse by classifying it into three separate discrete labels, namely 1) biased attitude, 2) act of bias and discrimination, and 3) violence and Genocide. The annotators chose among these labels and employed the majority voting scheme for the gold labels. This online abuse intensity study employed the classical categorical approach which is a binary perspective and failed to represent the diverse fine-grained context in a spectrum of continuum values.

In (**Ayele**, Jalew, et al., 2024), we hypothesize that hate speech is not a simple desecrate concept, but a complex and subjective task demonstrated in a spectrum of continuity (Bahador, 2023). The focus of the study presented in this chapter is extended beyond the binary approach to include the diverse intensities of hatefulness and offensiveness in tweets. The study also explored the portion of the community who are targeted within hateful tweets on X/ the former Twitter.

This chapter employed similar data collection and data processing procedures and strategies that are presented in Chapter 2. We collected over 3.9 million tweets from X and annotated a total of 8.3k tweets, with each tweet being assessed by 5 native speakers. Our annotations in this chapter covered three distinct types of tasks, namely *category*, *target*, and *intensity level* annotations.

In the *category* type of annotations, we requested annotators to classify each tweet into specific hate speech categories: *hate*, *offensive*, *normal* and *indeterminate*. The descriptions and definitions of these categories are presented in Chapter 1 and Chapter 2.

The *target* annotation type involves identifying the specific groups, individuals, or communities who are targeted by hate speech within the tweet. This process aids in understanding the intended targets of the harmful content, providing insights into the context and potential impact.

Lastly, the *intensity level* annotation type is a valuable measure for assessing the intensities of hate and offensive speech. It provides a means to measure where a tweet falls along the spectrum of harm, from milder instances to more severe cases. This type of annotation aids in understanding the varying degrees of harm and evaluating the subtle nature of such content.

In addition to identifying and investigating the targets of hate speech in the current dataset, this chapter aims to explore the extent of offensiveness and hatefulness in tweets on a rating scale of 1 to 5, where 0 represents normal tweets and 5 indicates very offensive tweets.

The primary contributions of this chapter include the following:

1. benchmark dataset for hate speech category and target detection tasks, supplemented with intensity level ratings,

2. comprehensive annotation guidelines for hate speech categories, targets, and approaches to measure the intensity of offensiveness and hatefulness, and

3. developed classification and regression models for predicting hate intensity levels and detecting hate speech categories and their targets.

Despite our study mainly focuses on Amharic on Ethiopian context, the outlined approaches can be further extended to other languages, social norms and cultural contexts.

5.2 Data Collection and Annotation

This section outlines the data collection and annotation strategies utilized in this chapter, which also provides a brief description of the annotated dataset.

5.2.1 Data Collection

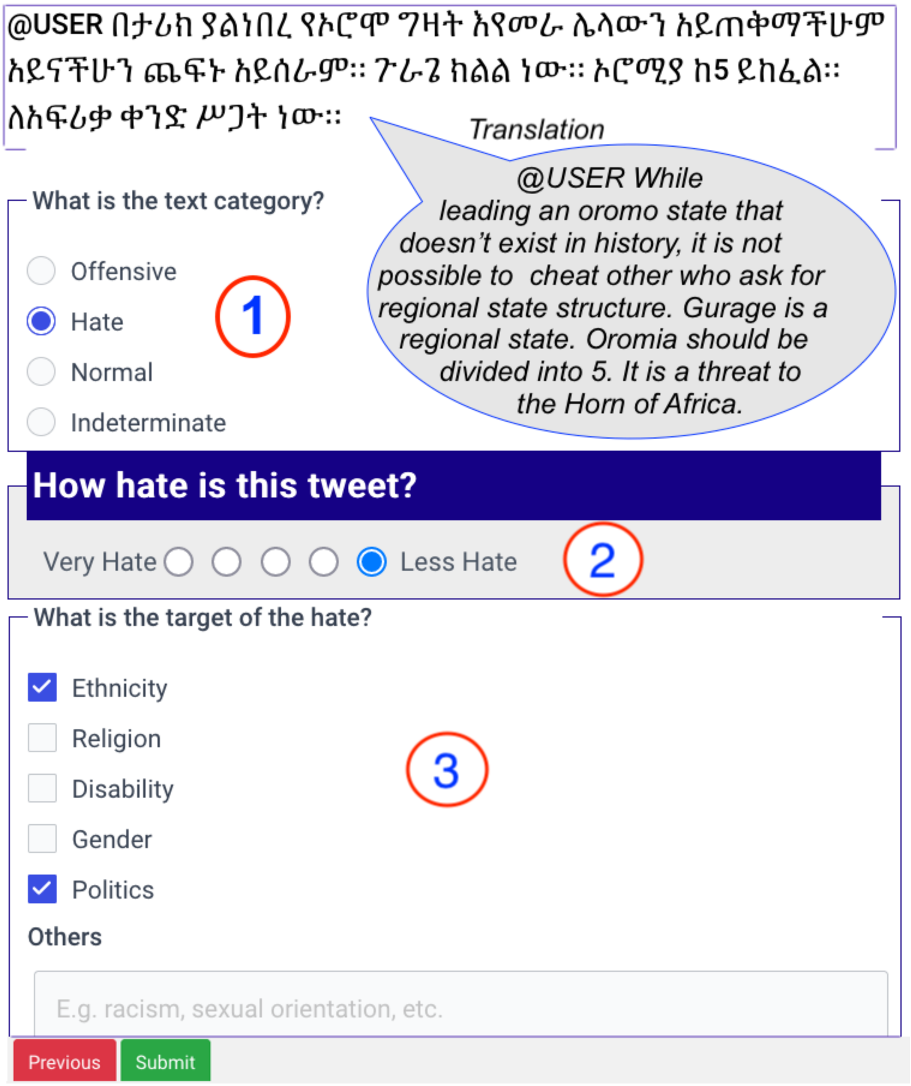
This dataset was collected from Twitter/X over a period of 15 months starting from January 1, 2022, over 3.9 million tweets written in Amharic. The details of data collection strategies and procedures are presented in Chapter 2.

5.2.2 Data Annotation

In this part, we provided a brief analysis of the data annotation strategies and procedures, including pilot and main task annotations descriptions.

Overall Annotation Procedures

We customized and employed the Potato-POrtable Text Annotation TOol (Pei et al., 2022) for the data annotation. Annotators were provided annotation guidelines, took hands-on practical training, completed independent sample test tasks, and participated in group evaluation of independent sample tests they completed. To further ensure better annotation quality, we conducted a pilot annotation of 300 tweets, achieving a Fleiss' kappa score of 0.46 agreement across the five annotators. A total of 8.3k tweets are annotated into *hate*, *offensive*, *normal*, and *indeterminate* classes. Besides, annotators were requested to identify the targets of hateful tweets and also indicate their ratings of the extent of hatefulness and offensiveness intensities of tweets on a 5-point Likert scale as indicated in Figure 7.2. The entire annotation process consists of a pilot round and five subsequent batches of main task annotations.



@USER በታሪክ ያልነበረ የኦሮሞ ግዛት እየመራ ሌላውን አይጠቅማችሁም አይናችሁን ጨፍኑ አይሰራም። ጉራጌ ክልል ነው። ኦሮሚያ ከ5 ይከፈል። ለአፍሪቃ ቀንድ ሥጋት ነው።

Translation

@USER While leading an oromo state that doesn't exist in history, it is not possible to cheat other who ask for regional state structure. Gurage is a regional state. Oromia should be divided into 5. It is a threat to the Horn of Africa.

What is the text category?

○ Offensive
● Hate
○ Normal
○ Indeterminate

How hate is this tweet?

Very Hate ○ ○ ○ ○ ● Less Hate

What is the target of the hate?

☑ Ethnicity
□ Religion
□ Disability
□ Gender
☑ Politics

Others

E.g. racism, sexual orientation, etc.

Figure 5.1: Potato GUI for the three types (1 - category, 2 - intensity, and 3 - target) of annotation tasks.

Each tweet is annotated by 5 independent annotators, and the gold labels are determined with a majority voting scheme. A Fleiss' kappa score of 0.49 is achieved

Tweet	Category					Hatred Targets					Offensiveness Intensity					Hatefulness Intensity				
አንች ሸርሙጣ ከማያገባሽ አትግቢ ይሄ ጭፈራ ቤት አይደለም ግም You a whore, don’t interfer in matters that doesn’t concern you. This is not night club.	off	off	off	off	off	--	--	--	--	--	3	4	5	5	4	--	--	--	--	--
አሸባሪው የኦሮሙማ መንግስት The terrorist Oromo-led government	hat	hat	hat	hat	hat	['eth', 'pol']	['eth']	['eth']	['eth', 'pol']	['pol']	--	--	--	--	--	4	4	4	4	4
@USER አንተ ደንቆሮ ነህ ስለ አርቶዶክስ አታቅም. You are ignorant, you don’t know about Orthodox.	off	off	off	off	hat	--	--	--	--	['rel', 'dis']	4	3	4	4	--	--	--	--	--	3
ቀይ መስቀል ለተፈናቃዮች 5 ሚሊየን ብር ግምት ያለው የዓይነት ድጋፍ አደረገ The Red Cross provided 5 million birr in-kind support to the displaced.	nor	nor	nor	nor	nor	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 5.1: Dataset examples across 5 independent annotators for category, hatred target and intensity (hatefulness and offensiveness) annotations.
Keys: off = offensive, hat = hate, nor = normal, eth = ethnicity, pol = politics, rel = religion, dis = disability.

among the five annotators in the main task. We compensated annotators with a payment of $0.03 per tweet, roughly 180 ETB per hour on average, nearly the same as the hourly wage of a Master’s degree holder in Ethiopia. On average, annotators can annotate 100 tweets in an hour, as we have already evaluated it during the pilot study.

Backgrounds of Annotators

A total of 11 Amharic native speakers, 5 female and 6 male annotators, were engaged in the annotation task, representing a diverse range of ethnic, religious, gender, and social backgrounds. This annotation project consisted of a team of experts who participated in the task over the course of a month. It comprised of 6 Master of Science graduates as well as 5 Master of Science students. These annotators were selected from different institutions and locations across Ethiopia, and represented a range of academic backgrounds encompassing both Natural and Social Science disciplines.

Table 5.1 presents examples, which showed the structure of the annotated dataset for the three types of annotations; namely category, hatred target and intensity (hatefulness and offensiveness) annotations.

Tweet Category Annotation

As indicated in Figure 5.2, the category annotation consisted of 4,149 hate, 2,164 offensive, 1,945 normal and 42 indeterminate labeled tweets. The dataset is predominantly composed of tweets that have been labeled with the "hate" class. The 5 annotators unanimously agreed on the class labels of 3.2k tweets out of a total of 8.3k, which accounts for approximately 39% of the entire dataset. The absolute agreements on each category label among the annotators were 38% and 31% for hateful and offensive tweets, respectively. We achieved an absolute annotation agreement on 49% of the tweets labeled as *normal* in the category annotation task, indicating that almost half of the tweets received perfect labeling consensus among all annotators. The indeterminate class, consisting of only 42 tweets, demonstrated exceptionally infrequent occurrence and is excluded from our experiments. While examining the tweets labeled with the "indeterminate" class, we found that such tweets constituted language content written in languages other than Amharic or unintelligible collections of text, failing to convey clear messages to the annotators or human experts in general. When determining majority-voted tweets for two labels with equal frequency of 2, we handle ambiguities by giving priority to the *hate*, *offensive*, and *indeterminate* labels, respectively.

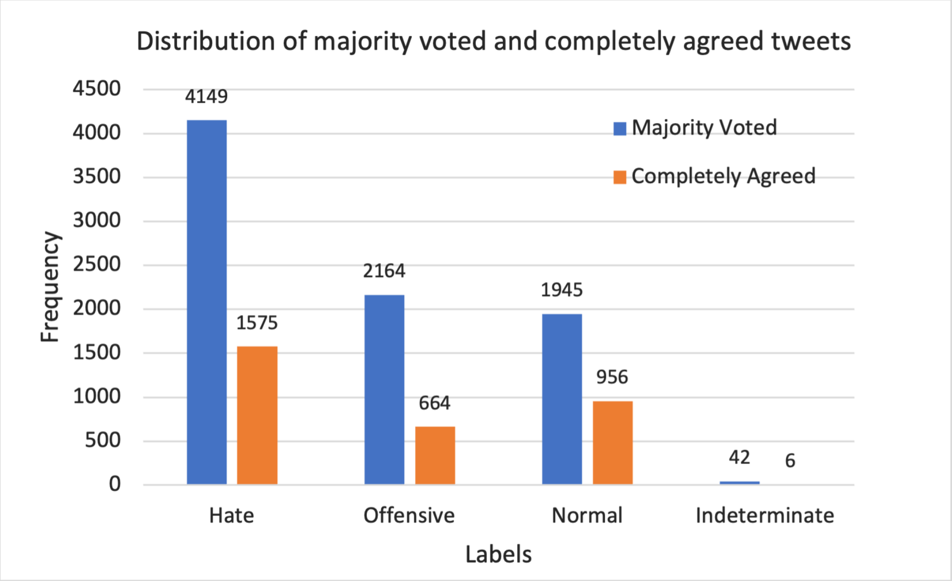


Figure 5.2: Distribution of category labels in the dataset.

Target Annotation

As indicated in Table 5.2, a significant majority of the target dataset, totaling 3,249 tweets (53.4%), comprised of instances expressing hatred and hostility towards *political* targets. Political hatred tweets primarily centered on individuals based on their political ideologies, affiliations, or support for specific occasions. Hateful tweets in the dataset mainly targeting ethnic-based identities contributed the second largest majority, approximately 38.8% of hateful tweets. The remaining portion of the hateful tweets targeting identities such as religion, gender, and other related identities exhibited smaller proportions in the target dataset.

While examining the proportion of tweets that achieved absolute consensus among the annotators on their respective labeling decisions, we found that *ethnic*, *political*, and *religious* hatred targets received complete consensus more frequently in the dataset. All five annotators entirely agreed on their labeling decisions to identify hatred targets in 867 instances of hateful tweets, which is approximately 14.3% of the hateful tweets. However, targets like *gender* and others such as *disability* are scarcely represented in this dataset. The *none_hate* instances in the target dataset indicated in Table 5.2, which

target	majority voted	entirely agreed	entirely agreed tweets in (%)
ethnic	**2,357**	**326**	14%
politics	**3,249**	**487**	15%
religion	359	54	15%
gender	42	0	0%
other	33	0	0%
none_hate	2,220	1,620	73%
total	**8,300**	**2,487**	**30%**

Table 5.2: Distribution of hatred targets across majority voted and fully agreed tweets.

comprises of 2,220 tweets, represent tweets that do not contain any hateful content, whether normal, offensive, or indeterminate.

Figure 5.3 demonstrated the number of times different distinct targets appeared simultaneously across the 5 annotators within the annotated dataset. It provided a detailed overview of the collective perspectives of these annotators regarding the simultaneous presence of distinct targets. The majority of overlapping occurrences that happened between *ethnic* and *political* targets in the dataset showed how *ethnic and political hatred targets frequently intersect and overlap with one another*, emphasizing the complex relationship between these two targets.

This overlap between ethnic and political targets is likely a manifestation of Ethiopia's political landscape, which is predominantly structured across ethnic divisions (Mostafa and Meysam, 2023). In Ethiopia, most political parties are established based on ethnic affiliations. This underscores the complex interconnection between ethnicity and political tensions in the nation's sociopolitical context. These two dominant and interlinked hatred targets, ethnicity and politics, also overlap with religion.

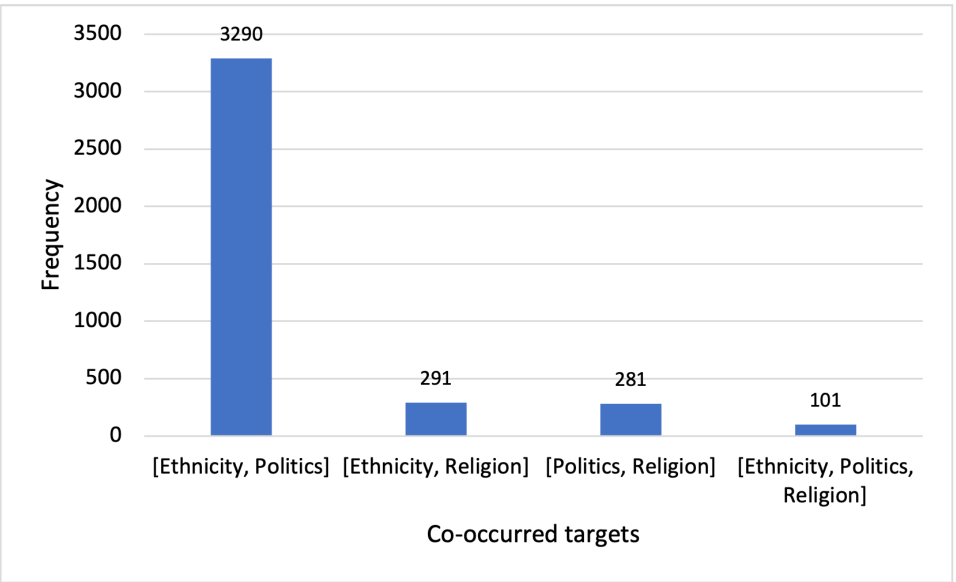


Figure 5.3: Major overlapping hatred target occurrences across hateful tweets in the dataset.

	majority voted		entirely agreed	
label	**intensity range**	**grand average**	**intensity range**	**grand average**
hate	0.4-5.0	2.48	1.4-5.0	3.56
offensive	0.4-4.8	2.34	1.6-4.8	3.66

Table 5.3: Hatefulness and offensiveness intensities. The "intensity range" indicates the intensity spans per each tweet while grand average presents the overall aggregated intensity across the dataset.

Intensity Level Annotation

The perceptions of hate and offensiveness are measurably subjective, indicating that predicting individual judgments is a hard task (Wojatzki et al., 2018). It is important to incorporate the perceptions of multiple annotators and explore measurement options such as rating the intensity of hate and offensiveness. Thus, we have organized our intensity level annotation task into three distinct segments. *Normal* texts are assigned a score of *0*, waiving the need for intensity level annotations. The offensiveness scale spans from *less offensive (1)* to *very offensive (5)*, utilizing a 5-point Likert scale for intensity level annotation. Similarly, the intensity of hatefulness is also rated on a 5-point Likert scale, ranging from *less hate (1)* to *very hate (5)*.

Table 5.3 presented the offensiveness and hatefulness intensities of tweets that appeared at least 2 times as offensive and hateful across the 5 annotators, respectively. Average offensiveness and hatefulness intensities on majority-voted tweets are lower than the entirely agreed tweets. The majority voted tweets exhibit wider ranges of intensities for both offensiveness and hatefulness, spanning from 0.40-4.80 and 0.40-5.0, respectively. This indicated that hate and offensive annotated tweets in the dataset are represented in a spectrum of wider ranges of intensities. Therefore, hatefulness and offensiveness *are not simple binary measures*, rather they exist on *a continuum with varying degrees of intensity*.

In the category of completely agreed tweets, the range of offensiveness intensity spans from a minimum average intensity of 1.60 to a maximum average intensity of 4.80 per tweet. Meanwhile, in the case of hateful tweets, their hatefulness intensity encompasses intensities ranging from a minimum of 1.40 to a maximum of 5.0 across the subset of entirely agreed tweets. The wider intensity ranges and the cumulative average intensity values for offensiveness and hatefulness on the completely agreed tweets highlight the presence of varying degrees of intensity, even among tweets that have complete agreement across all annotators.

5.2.3 Mapping Hate and Offensive Intensities

Bahador (2023) categorized hate speech into three major stages, namely 1) early warning, 2) dehumanization and demonization, and 3) violence and incitement, including genocide. The *early warning* category starts with targeting the *out-groups*[1] to different types of negative speech that have less intensity such as slight insults.

The second category, *dehumanization and demonization* involve dehumanizing and demonizing the out-groups and their members, associating with subhuman such as

[1] Out-groups are anyone who does not belong in the group but belongs to another group

label	average range	intensity stage	# tweets	percentage
offensive	[0.2 - 3.0)	mild	**2,008**	**69%**
	[3.0 - 4.0)	moderate	676	23%
	[4.0 - 5.0]	severe	245	8%
subtotal			2,929	100%
hate	[0.2 - 3.0)	early warning	**3,489**	**72%**
	[3.0 - 4.0)	dehumanization	808	17%
	[4.0 - 5.0]	violence & incitement	528	11%
subtotal			4,825	100%

Table 5.4: Hatefulness and offensiveness intensity ranges, and distribution of tweets across intensity stages.

"Rat", "Monkey", "Donkey" or superhuman negative characters like "Monster", "Cancer" (Bahador, 2023).

The last category, *violence and incitement* starts from the conceptual to the physical attacks and can result in more severe consequences such as incitement to violence and or even death against the out-groups under target, including genocidal incitement (Bahador, 2023).

These categories which classify hate speech into different phases, showcase the need for multifaceted mitigation strategies among stakeholders such as researchers, practitioners, policy makers, and social media organizations.

Similarly, Chandra et al. (2020) classified online abuse into three labels; 1) *biased attitude*, 2) *acts of bias and discrimination*, and 3) *violence and genocide*; to showcase the mild, moderate, and severe categories of abuse intensity.

We employed the different classification strategies of Bahador (2023) and Chandra et al. (2020) to represent the hatefulness and offensiveness intensities of tweets, respectively, as indicated in Table 5.4. In this chapter, we utilized the revised rating scale described in Section 5.2.2 and represent offensiveness into three stage categories (Chandra et al., 2020), mild, moderate, and severe, indicated by 1-3, 4, and 5 rating scales, respectively. We employed a similar strategy to map and represent the three stage categories of hatefulness on the 5-point rating scale as shown in Figure 5.4.

As shown in Table 5.4, we carefully selected tweets which are labeled offensive with the majority voting scheme and further explored their offensiveness intensities. We also examined hateful tweets to uncover hatefulness intensities inherent within tweets. A separate analysis of hatefulness and offensiveness intensities of tweets in the dataset are presented utilizing the strategies of Chandra et al. (2020) and Bahador (2023), respectively.

Offensive tweets that fall under the mild category start from a minimum average intensity of 0.2 when only one of the annotators chooses to label them as offensive, rating their offensiveness as 1, and extend to a maximum average intensity value of 3. Tweets in this category comprised 69% of all offensive tweets and are assumed to be less offensive compared to the other categories. Highly offensive tweets, constituting 8% of all offensive tweets, present incitement or threats of violence against an individual, while the moderate category accounts for 23% of the tweets that dehumanize or demonize individuals.

The majority of hateful tweets, comprising 72% of all tweets, fall under the 'less hate, early warning' category. The remaining 17% and 11% of hateful tweets fall under the second and third categories of hatefulness intensity ranges, respectively. These two higher categories of hatefulness, 'dehumanization' and 'violence and incitement', require serious attention from various stakeholders such as the government, social media platforms, researchers, and non-governmental organizations (national and international).

The mild category of offensiveness and the early warning stage of hatefulness intensities can be seen as a demarcation point to enforce mitigation strategies by content moderators or other stakeholders. The battleground for tackling hate and offensive speech on social media should focus primarily on these first stages of early warning and mild, respectively.

In order to have a comprehensive and unified observation of abusive speech, we transformed the original dataset, which is annotated for hatefulness and offensiveness intensities on a 5-point Likert rating scale, into a range of 0 to 10, effectively creating an *11-point Likert scale*. In this revised scale, a score of 0 represents *normal* tweets while *offensive* and *hate* categories are scaled from 1 to 5 and 6-10 intensity ranges, respectively. The score of 1 and 5 denotes *less offensive* and *highly offensive* tweets, respectively. Similarly, 6 signifies *less hate*, and 10 represents a tweet characterized by *intense hate*. Figure 5.4 shows the transformed dataset on an 11-point Likert rating scale.

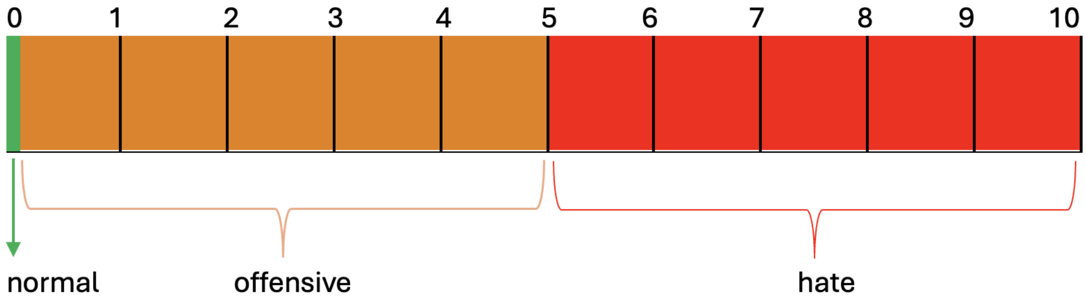


Figure 5.4: Mapping the dataset in an 11-point Likert scale.

5.2.4 Dataset Summary

A total of 8.3k instances were utilized for building classification and regression models, excluding the 42 indeterminate labeled instances. We presented the distributions of the dataset labels for the category, target, and intensity level classification and regression experiments in Figure 5.2, Table 5.2, and Figure 5.5, respectively.

We convert the average values calculated from the input of five annotators into whole numbers, resulting in a set of 11 labels spanning from 0 to 10. In this context, a label of 0 represents tweets labeled as *normal* while a label of 10 indicates tweets characterized as *extremely hateful*. Figure 5.5 illustrates that scale labels 1 and 10 are associated with a relatively smaller number of instances in comparison to the other labels, as these values correspond to the two extremes of the spectrum.

5.3 Experimental Setup

We employed a 70:15:15 data-splitting approach to create the training, development, and test sets. This dataset remained consistent across all experiments, including *category classification*, *target classification*, and *intensity scale regression*. The development dataset

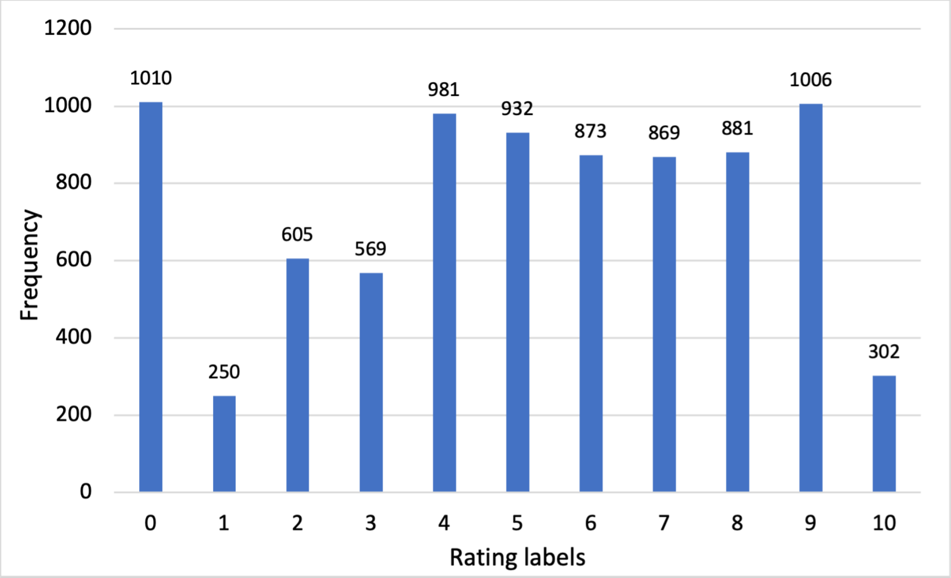


Figure 5.5: Distribution of rating labels.

was instrumental in refining the learning algorithms, and all the results reported in this study are based on data from the test set.

We utilized the transformer models such as *Am-RoBERTa* , *XLMR-Large-fintuned*, *AfroXLMR-large*, and *AfriBERTa* variants (small, base, large), and *AfroLM-Large (w/ AL)* for all experiments. Am-RoBERTa is a RoBERTa-based language model that has been fine-tuned specifically with the Amharic language dataset, making it well-suited for downstream tasks and applications involving Amharic text (Yimam et al., 2021). We also utilized Afro-XLMR-large (Alabi et al., 2022), a multilingual language model tailored for African languages, including Amharic. This model demonstrated exceptional performance in various natural language processing tasks for African languages. Moreover, we fine-tuned the XLMR-Large (Conneau et al., 2020) model using the same corpus that was utilized to train Am-RoBERTa. We also employed the small, base, and large *AfriBERTa* variants (Ogueji et al., 2021), and *AfroLM-Large (w/ AL)*, Pretrained multilingual models on many African languages including Amharic (Dossou et al., 2022). AfroLM Large (w/AL) is a special type of AfroLM Large which is designed with self active learning setups.

5.4 Result and Discussion

As shown in Table 5.5, the Afro-XLMR-large model outperforms the other 6 models on both tweet category and hatred target classification tasks with 75.30% and 70.59% F1 scores, respectively. In comparison to their performance on target classifications, all models exhibit a pronounced increase in all performance indicators such as precision, recall and F1 scores when undertaking the category classification task. Table 5.6 indicates the spectrum of F1 score variations across diverse models. The performance variations observed in these two tasks extends from 4.71% for Afro-XLMR-large to 8.80% for Am-RoBERTa. This disparity might be due to the class representation variations in the target classification task.

Tweet category classification results (in %)			
Classifier	**Precision**	**Recall**	**F1 Score**
Am-RoBERTa	75.01	75.06	74.82
XLMR-large-finetuned	73.60	73.45	73.50
Afro-XLMR-large	**75.37**	**75.30**	**75.30**
Afriberta-large	72.48	72.40	72.43
Afriberta-base	73.46	73.20	73.30
Afriberta-small	73.05	73.12	73.06
AfroLM-Large (w/ AL)	72.02	71.99	71.98
Hate target classification results (in %)			
Am-RoBERTa	66.74	66.42	66.02
XLMR_large_fintuned	65.57	66.18	65.85
Afro_XLMR_large	**70.34**	**70.94**	**70.59**
Afriberta_large	66.94	67.47	67.14
Afriberta_base	66.04	66.42	66.11
Afriberta_small	65.38	66.02	65.68
AfroLM-Large (w/ AL)	64.26	64.57	64.23

Table 5.5: Performance of models for category and hatred targets classification of tweets.

F1 score variations across tasks: category vs target (in %)			
Classifier	**Category**	**Target**	**Difference**
Am-RoBERTa	74.82	66.02	**8.80**
XLMR-large-finetuned	73.50	65.85	7.65
Afro-XLMR-large	**75.30**	**70.59**	**4.71**
Afriberta-large	72.43	67.14	5.29
Afriberta-base	73.30	66.11	7.19
Afriberta-small	73.06	65.68	7.38
AfroLM-Large (w/ AL)	71.98	64.23	7.75
Average	73.48	66.52	6.97

Table 5.6: F1 score Performance variations across models for category and hatred target classification tasks.

Regression results (in %)	
Classifier	**Pearson Corr. Coeff. (r)**
Am-RoBERTa	77.23
XLMR-large-fintuned	76.17
Afro-XLMR-large	**80.22**
Afriberta_large	75.38
Afriberta_base	76.57
Afriberta_small	74.94
AfroLM-Large (w/ AL)	**80.22**

Table 5.7: Performance of models on the regression tasks with Likert's 11-scale data.

We conducted *regression* experiments on the dataset collected through the utilization of an 11-point Likert scale, which was employed to measure intensity levels across a broad spectrum of ratings. In these experiments, real-valued scores spanning from 0 to 10 were utilized, and various models were applied for analysis. As part of our methodology, we focused on enhancing the visualization of the regression results for better interpretation. To achieve this goal, we rounded the results and illustrated them with visual representations presented in Figure 5.6.

Regression experiments were also performed on the 11-point Likert scale data with various models, and their performance was assessed using Pearson's r correlation coefficients. As suggested by Fieller and Pearson (1961) and Schober et al. (2018), correlation coefficients falling between 0.70 and 0.89 are considered to indicate a strong correlation. Hence, the Pearson's r correlation coefficients achieved in this study, ranging from 74.94% to 80.22% demonstrated strong correlations. These findings denote a robust relationship between the predicted values and the actual observations, underscoring promising performance outcomes across all the models. The Afro-XLMR-large and AfroLM-Large (w/ AL) models presented the best results in the intensity scaling regression tasks, which is 80.22%.

5.5 Error Analysis

Figure 5.6 reveals that the majority of misclassified instances are clustered along the diagonal within the dark-colored boxes. This suggests that the true labels and their predicted counterparts are closely aligned. For instance, the true label 9 is frequently predicted as 7, 8, or 10, but seldom as 0, 1, 2, 3, or 4, which are considerably distant from 9. Conversely, there are only a few cases where extremely low true labels, such as 0, 1, 2, and 3, are predicted as higher extreme values, such as 7, 8, 9, or 10, and vice versa. In general, the regression model consistently displayed superior and more dependable performance as evidenced by the distribution of predictions in the confusion matrix.

As presented in Figure 5.7, the majority of errors, 47.84%, within the predicted intensities showed only 1 scale variation with the actual annotation scores. The second majority presented a 2 scale differences between the actual and predicted intensities, which accounts 28.36% of the errors. Over 76% of the predictions are closer to the actual values, with 1 or 2 intensity scale differences. Such small variations are also common experiences among human experts due to subjectivity.

Table 5.8 showcases some examples of incorrectly predicted intensities within the test set from the Afro-XLMR-large model. The first example shows a tweet labeled as 0, which is a normal category, but it was predicted as 8, an early warning hate speech category. The model might be confused when predicting the tweet due to the demonyms that named people with the name of their city/district, such as *Janamoras'* or *Debarks'*. The second example, which is predicted as 7 (an early warning hate speech category), is actually labeled as 0, indicating a normal category. The model fails to capture the local context, as annotators can associate the tweet only with the local socio-political situations including the current ethnic based struggles in Ethiopia. In the third example, the model predicts 6 instead of 0 and fails to recognize sarcastic expressions that require contextual knowledge about the *leader* which human annotators can handle effectively. The fourth and fifth examples show slight variations in intensity between the actual and predicted scores, with a one-point difference from 4 to 5 and from 8 to 7, respectively. The model's predictions are closer to the actual scores, where such slight variation is also common among human expert annotators. Despite there is a

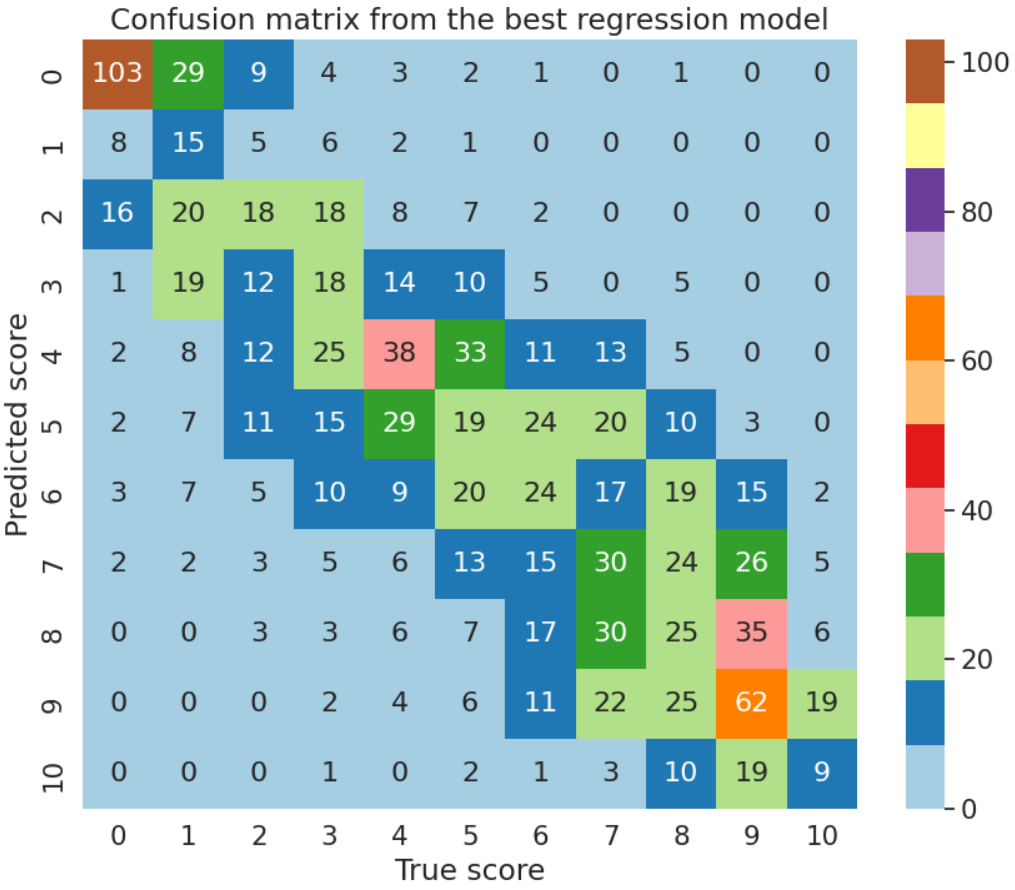


Figure 5.6: Confusion Matrix from Afro-XLMR-large.

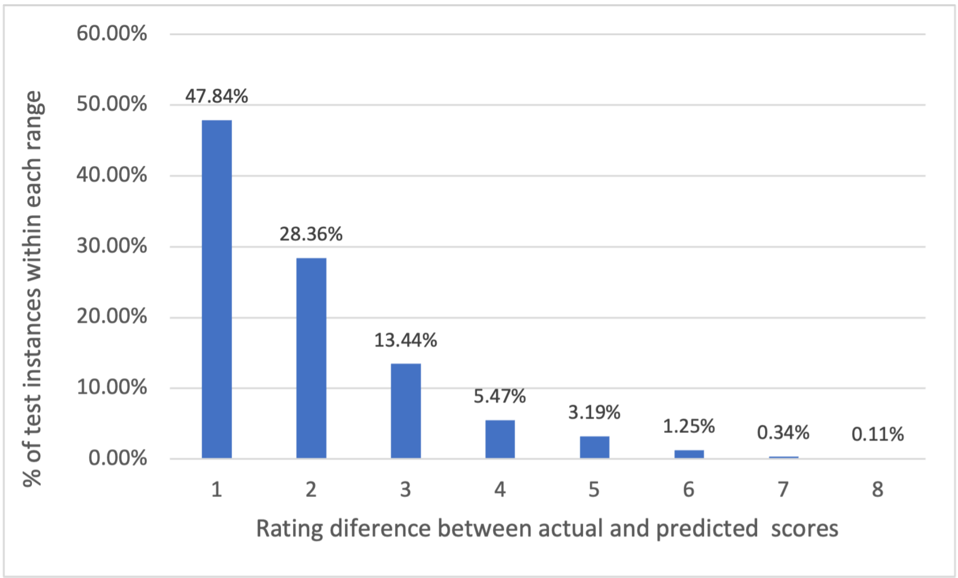


Figure 5.7: Variations within actual and predicted intensity ratings of tweets.

one-point deviation in the fifth example, it is possible to confirm that the model can capture idiomatic expressions in this case.

	Tweet	Actual	Predicted	Remark
1	ለማንኛውም ጃናሞራዎች ደባርቆች እኛ ስንናገር ቅር የተሰኛቹህ እውነታው ይህ ነው	0	8	The model confused due to Demonym naming
	In any case, You Janamoras' and Debarks', those who are disappointed when we speak, this is the fact.			
2	@USER የጠላቶቻችንን ሴራ የምንታገለው አንድ ሆነን ስከን ብለን ስንራመድ ነው ይሄም ያልፋል	7	0	The model fails due lack of context
	@USER We can only fight our enemies' conspiracy when we work together wisely; this will pass.			
3	ተመሰገንንንንንንንንንንንንንንን ቸር ደግ ታማኝ ጀግና ኩሩ ከሰውች በልጦ የተገኘ መሪ ተሰጠንን	6	0	The model fails due to Sarcasim
	Thanks, we are given a leader who is generous, kind, honest, brave, and proud.			
4	@USER አሁን ይሄ ደደብ ምንስ ቢል እንዲህ አለ ተብሎ ይለጠፋል?? ኤሬዲያ!	4	5	Slight subjectivity
	@USER is that worth to post, whatever this idiot says, shit.			
5	ይሄ ጭር ሲል አሎድም የሆነ መንግስታችን ግን ሰሞኑን ደህና ነው??	8	7	The model detects idioms
	Is this government, which do not need peace at all, fine these days?			

Table 5.8: Incorrectly predicted examples from Afro-XLMR-large within the test set.

The findings indicate that considering hate speech as a continuous variable, rather than adopting a binary classification, is a more suitable approach. Regression-based methods excel at capturing the intricate and evolving characteristics of hate speech, recognizing the subtle variations and intensities within this complex and sensitive domain.

This approach aligns with the dynamic and multifaceted nature of hate speech in the real-world situations, where it often exists on a spectrum of varying intensities, defying the usual simple binary categorization approaches.

5.6 Conclusions

This chapter introduced extensive benchmark datasets encompassing 8.3 tweets annotated for three tasks. These tasks included 1) *categorizing* hate speech into labels such as hate, offensive, and normal, 2) identifying the *targets* of hate speech, such as ethnicity, politics, and religion etc, and 3) assigning hate and offensive speech *intensity levels* using *Likert rating scales* to indicate offensiveness and hatefulness. To ensure robust annotation, each tweet is annotated by five annotators, resulting in a Fleiss' kappa score of 0.49. Our contribution extended beyond the dataset itself; we provided comprehensive annotation guidelines tailored to each task and offered illustrative examples that effectively outlined the scope and application of these guidelines.

After a comprehensive analysis of the dataset, a clear pattern emerged, highlighting the prominence of *political* and *ethnic* targets, which mirrors the complex sociopolitical dynamics of Ethiopia. Notably, these two targets often co-occur in hateful tweets, underscoring the intricate nature of Ethiopia's sociopolitical dynamics, especially within ethnic context, addressing the second research question: *To what extent do hate speech disproportionately target specific vulnerable communities?* Furthermore, our findings have demonstrated variations in the intensity of hate speech, emphasizing the necessity to develop regression models capable of measuring the level of intensity in tweets, which addressed the third research question: *How can hate and offensive speech be understood: as distinct categories or as values on a spectrum of varying intensities?*

We conducted a comprehensive exploration of various models for the detection of hate speech *categories*, their associated *targets*, and the diverse *intensity levels* inherent in it. Afro-XLMR-large demonstrated superior performance across all tasks *category classification*, *target classification* and *intensity prediction*. The findings in this chapter illustrated that offensiveness and hatefulness cannot be identified as simple discrete concepts; instead, they manifest as continuous variables that assume diverse values along the continuum of ratings.

One Picture is Worth a Thousand Words

— The San Antonio Light's Pictorial Magazine of the War
(1918)

6
Multimodal Hate Speech Detection and Analysis in Amharic Social Media Memes

Contents

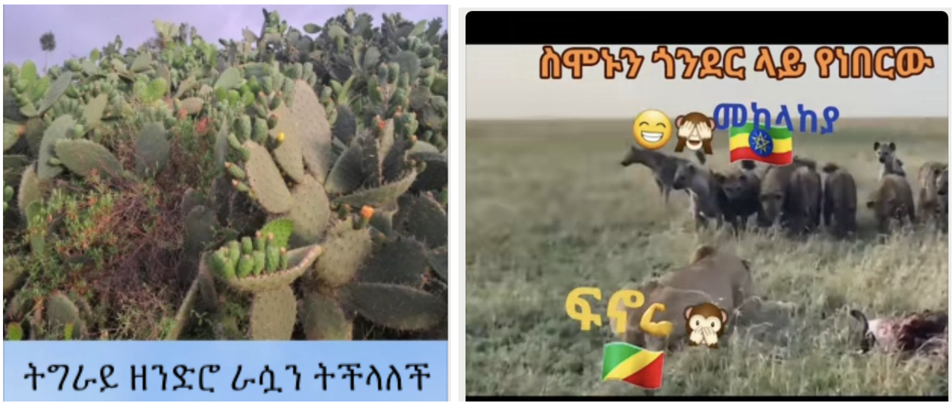
6.1 Introduction

In the previous chapters, we have presented hate speech datasets that were collected using diverse strategies. Besides, we explored various machine learning models to detect, classify, and rate the intensity of hatefulness and offensiveness present within textual content scraped from social media.

Most of hate speech studies in low-resource African languages such as Amharic mainly focus on detecting hate within textual content extracted from social media (Abebaw et al., 2022a, 2022b; Mossie and Wang, 2018; Tesfaye and Kakeba, 2020; **Ayele**, Dinter, et al., 2022; **Ayele**, Yimam, et al., 2023).

In our work (Jigar et al., 2024), we emphasize on detecting and analyzing hate speech in a multimodal setup to utilize Amharic memes that are extracted from various social media platforms.

Identifying hate speech through employing multimodal analysis of social media memes leads to specific challenges associated with its multiple and complex inputs,



(a) Example-1.

(b) Example-2.

Figure 6.1: Text and image fusion examples.

and detection processes (Schmidt and Wiegand, 2017; Kiela et al., 2020; Ahsan et al., 2024). The following are some of these particular challenges facing multimodal hate speech detection:

Image and text synergy: memes often simultaneously utilize visual and textual fusion through combining text and image elements to describe a certain situation. The text may contain some normal phrases, while the image may provide an additional context that give the meme a new message when combined (H Lin et al., 2023). Both the image and the text may not explain any hatred message separately, but may convey hate speech when combined and used together, which makes the task difficult due to the complex relationship between the two modalities (Hossain et al., 2024). In Figure 6.1 (a) & (b), both the texts and the images do not contain any hate speech separately while conveying explicit hatred messages targeting ethnicity (in example 1) and politics (in example 2) when combined together. The text in example 1, which is translated as *"Tigray will be self-sufficient this year"* combined with an image of *Cactus fruits* to form a meme. The text in example 2, translated as *"Defense forces at Gondar" and "Fano"* combined with pictures of *Heynas* and *Lions*, respectively, created hateful memes when combined together.

Indirectness and contextual ambiguity: memes often convey indirect messages that relay on social and cultural references, humor, irony, and sarcasm (Pramanick et al., 2021). It poses challenges to understand the subtle messages explicitly within text and image fusions which require meticulous examination of messages from diverse contextual perspectives (Huang and Bai, 2021). This necessitates the hate speech detection models to understand social and cultural views, which poses challenges for automated systems.

Cultural dynamics: memes evolve diversified cultural representations and spread rapidly across various social media platforms (Agarwal et al., 2024). Memes are often generated by online users and appeared in various forms, which often become viral on social media platforms. Memes increase the subjectivity and difficulty of hate speech to be detected with machine learning models due to its inherent characteristics such as multimodality, contextual variability, evolving diversified cultural, sarcastic and humorous behaviors (De la Peña Sarracén et al., 2020).

Addressing the challenges of multimodal hate speech detection from user-generated memes requires collaborations that combine insights from disciplines such as linguistics, sociology, digital media, natural language processing and computer vision (F Wu et al., 2024; Thapa et al., 2024; Agarwal et al., 2024).

This chapter mainly focuses on detecting hate speech through employing multimodal analysis of memes extracted from various social media platforms such as Facebook, Twitter, and Telegram. We also further explored the performance of multimodal models in detecting Amharic memes and investigate relevant features in multimodal hate speech detection task.

6.2 Related Works

The concept of hate speech is highly subjective due to the prevailing societal norms, individual perspectives, contextual factors, and collective viewpoints (Madukwe et al., 2020; Yimam et al., 2019). Dealing with detecting online hate speech becomes more challenging and complex when the content is presented in the form of memes, which require implementing multimodal hate speech detection model (Thapa et al., 2024).

Over the last decade and a half, a lot of research has been conducted to mitigate the widespread dissemination of online hate speech across social media platforms. Most of these research attempts have mainly concentrated on detecting hate speech using unimodal detection approaches, which employ features from only a single input, such as text, image, or audio (Thapa et al., 2022; Suryawanshi et al., 2020). However, online hate speech in social media often comes in the form of memes that are typically humorous, sarcastic or irony employing an artistic combination of texts and images to reflect the contemporary social, cultural and political contextual variations in a multimodal environment (Pramanick et al., 2021).

Recently, there has been an increasing interest among researchers in exploring multimodal hate speech studies, especially analysis of social media memes (Thapa et al., 2022; Pramanick et al., 2021; F Wu et al., 2024; Ahsan et al., 2024; Schmidt and Wiegand, 2017; Kiela et al., 2020; Velioglu and Rose, 2020; Bhat et al., 2023; Gomez et al., 2020; Cao et al., 2022).

The work by Thapa et al. (2022) collected 5,680 memes from X/ Twitter with text-image pairs, which focuses on Russia-Ukraine war and annotated with a binary labels *hate* or *no-hate*. The authors explored models such as unimodal-text only, unimodal-image only, and multimodal experiments for both text-image fusion features, which indicated that the multimodal experiments outperformed both the textual and visual models. The multimodal modals achieved 5% and 10% F1 score performance improvements, excelling over the image-only and text-only models, respectively.

Pramanick et al. (2021) also introduced *Momenta*, a multimodal fusion model, which utilized memes extracted from Reddit, Facebook and Instagram focusing on the US politics and COVIR-19 pandemic. Pramanick et al. (2021) proved that multimodal models outperformed text-only and image-only unimodal approaches, achieving better performance, over 10% and 8% F1 score on hateful speech classification and target detection tasks, respectively.

Many other studies such as F Wu et al. (2024), Ahsan et al. (2024), Schmidt and Wiegand (2017), Kiela et al. (2020), Velioglu and Rose (2020), Bhat et al. (2023), Gomez et al. (2020), and Cao et al. (2022), also clearly indicated that employing multimodal approaches for the task of hate speech detection has shown better results than implementing text-only and image-only unimodal models separately, achieving up to 17%

performance increment in the multimodal setup. For instance, the multimodal models in the works such as F Wu et al. (2024), Ahsan et al. (2024), and Schmidt and Wiegand (2017), achieved 17.5%, 14% and 17% better results than image-only models, respectively. Each of these works also reported at least a 10% performance improvement on multimodal experiments when compared with text-only models.

Similarly, multimodal hate speech research has attracted researchers from low-resource African Languages such as Amharic. The work by Degu et al. (2023) tried to extract texts from Amharic memes through the application of Abyssinia-OCR, MetaAppz, and Amharic-OCR techniques. The authors utilized the fastText word embedding approach (Joulin et al., 2017) to detect hate speech from the extracted texts by employing unimodal detection approaches (Joulin et al., 2017). The approach utilized by Degu et al. (2023) solely relies on the extracted text from memes, neglecting modeling and analysis of the image component, which potentially resulted in an incomplete interpretation of the memes' intended messages.

On the other hand, the work conducted by Debele and Woldeyohannis (2022) presented a multimodal Amharic hate speech detection from audio and textual features on a dataset of 1,459 audio samples extracted from YouTube videos. The authors employed Word2Vec, and Mel-frequency cepstral coefficients (MFCC) which is a representation of the short-term power spectrum of audio signals, to extract textual and audio features, respectively. The authors applied the Google Speech-to-Text API to transcribe the audio speech signals into textual scripts. The best performing model in Debele and Woldeyohannis (2022), BiLSTM, achieved 78.23%, 83.97% and 88.15% accuracy scores on text-only, audio-only and multimodal models, respectively, which indicates that the multimodal approach outperformed the text-only and audio-only models by 9.92% and 4.18% accuracy scores, respectively.

In this chapter, we mainly aim to tackle the challenges of detecting hateful memes, through employing multimodal approaches that utilize concatenated features arising from image and text inputs, creating fusion models that are capable of recognizing hate speech from memes.

6.3 Data Collection and Annotation

The datasets are collected from three widely used social media platforms in Ethiopia, namely Telegram, Twitter, and Facebook. We have created a Telegram group called *Hate Speech Dataset Collectors*, consisting of 74 members, who are employed as data collectors from social media platforms. The members were trained about the data collection process and provided data collection guidelines. The 74 data contributors collected 10k memes to our Telegram group repository[1]. The memes are mainly collected by employing a variety of keywords, from the selected group accounts that have more than 100k followers. Depending on the number of members, the language used, and the frequency of news or trending discussions pertaining to politics, ethnicity, religion, and gender, we also considered several public pages for the data collection. We exclude memes that have only images or texts and contain only mere humorous content. Following the filtering process, we obtained a final dataset consisting of 2k memes out of 10k collected.

The datasets collected from each social media source are presented in Table 6.1.

[1] https://t.me/hateSpeech_image_data_c

Social Media	Total Number of Memes
Facebook	940
Twitter	261
Telegram	806
Total	2,007

Table 6.1: Distribution of collected memes from different social media.

We employed three native Amharic speakers to annotate the memes into binary categories, *hate* or *non-hate*. Annotators received live training sessions with detailed explanations of the annotation guidelines prior to their involvement in the main annotation task. The dataset comprised of 2k memes, annotated in four separate batches, each containing 500 memes. Each meme underwent annotation by three independent data annotators, achieving a Fleiss' kappa score of 0.50 inter-annotator agreement. A majority voting scheme was utilized to determine the final gold labels, resulting in relatively balanced number of labels, 919 *hate* and 1,088 *non-hate* instances.

6.4 Experimentation

In this section, we present a brief overview of the data processing methods and classification techniques that have been utilized within this chapter. We specifically cover the data preparation tasks such as optical character recognition from meme images and feature extraction tasks from texts and meme images. The models utilized to detect and classify hate speech within Amharic memes are also briefly explained.

6.4.1 Optical Character Recognition

We employed Tesseract[2], an open-source OCR engine utilizing advanced deep-learning algorithms, notably the Pytesseract Python library, to extract text from Amharic memes, as outlined in Ignat et al. (2022). Preceding the input of memes into Tesseract, we applied preprocessing techniques such as *grayscale conversion* and *noise reduction* to enhance meme quality. We extract texts from the preprocessed memes utilizing Tesseract OCR text extraction tool.

We maintain Amharic sentences with mixed English content to include messages from users who frequently switch between languages, Amharic and English, in their message compositions. This approach is mainly used to avoid spontaneous changes in the meaning of messages that might occur when removing mixed scripts such as English content presented within Amharic messages. We employed Python language detection and translation libraries to identify and translate mixed English terms within Amharic messages into their corresponding Amharic equivalents.

The text extracted from the memes passes through several preprocessing steps such as cleaning, normalization, translating specific English words into their Amharic counterparts, expanding abbreviations, eliminating stop words, and tokenizing inputs

[2] https://github.com/tesseract-ocr/tesseract

sentences while we standardized the meme images into uniform dimensions, and rescaled the pixel values into a range of 0 to 1.

6.4.2 Feature Extraction

Word embedding techniques are utilized to process and extract the textual features while the pre-trained *VGG16* and *ResNet* are employed to extract image features from memes as indicated in Figure 6.2. The VGG16, a convolutional neural network architecture, has been extensively trained on a substantial image dataset, providing it with the capability to extract significant image features effectively (Karim et al., 2023). Subsequently, we concatenated the output features from the word embedding process with those derived from VGG16's image feature extraction, combining and feeding the concatenated features to the multimodal models. For the transformer-based models, ResNet architecture is utilized for extracting meme-image features and performing classification tasks (He et al., 2016).

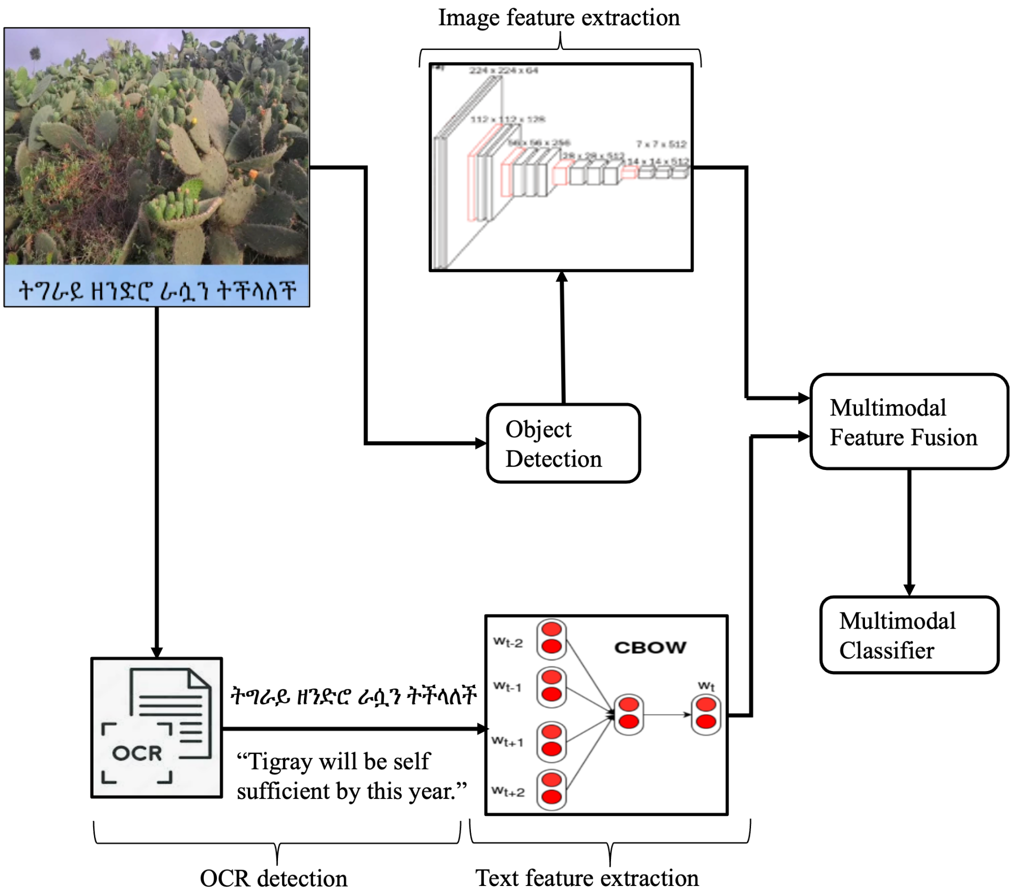


Figure 6.2: Text and image features concatenation.

6.4.3 Classification Models

We leveraged deep-learning algorithms, including LSTM, BiLSTM, CNN and various transformer models, which are fine-tuned for African languages including Amharic such as Am-Roberta (Yimam et al., 2021), Bert-base-uncased[3], Afro-XLMR-base (Alabi et al., 2022), Afro-XLMR-large (Alabi et al., 2022), Rasyosef_Bert-medium-amharic[4] and Rasyosef_Bert-small-amharic[5]. These models have shown proven efficacy in accurately classifying hate speech within meme datasets, as evidenced by prior research studies (Gomez et al., 2020; Debele and Woldeyohannis, 2022; Karim et al., 2023). Transformer models that have been pre-trained on Amharic datasets, such as Am-Roberta and Afro-XLMR-large, have demonstrated their effectiveness in detecting hate speech from textual data, particularly in the context of Amharic datasets (**Ayele**, Jalew, et al., 2024; **Ayele**, Yimam, et al., 2023). The performance capabilities of transformer models in both unimodal and multimodal settings can be attributed to the fine-tuning of these models after they have been pre-trained on high-resource languages such as English and subsequently on low-resource languages. This process helps the models transfer their knowledge of linguistic structures, semantic representations, and contextual understandings from high-resource languages to low-resource languages (Pires et al., 2019; Bao et al., 2022).

6.5 Results and Discussion

In this section, we provide a comprehensive overview of the model results obtained from our experiments, which encompass both unimodal and multimodal approaches. These experiments were designed to address the challenge of hate speech detection in the Amharic meme dataset.

The experiments were structured into three distinct categories, each focusing on a specific modality: *Text-Only*, *Image-Only* and *Multimodal* models. The primary objective of these experiments are to evaluate the effectiveness of these deep learning algorithms in identifying hate speech within the Amharic meme dataset. We systematically examined the performance of models under each modalities and explored insights into the strengths and weaknesses in handling the unique challenges posed by hate speech detection from memes.

We fine-tuned several transformer-based and deep learning models with manually labeled Amharic meme dataset. We employed a range of performance evaluation metrics, including Precision, Recall, F1 scores, and accuracy.

As depicted in Table 6.2, our experimental results revealed that Am_Roberta outpermed all the models in *Text-Only* models, with an F1 score of 74.04%. In general, the transformer models showed better results in *Text-Only* models, ranging between 68.35%-70.42% F1 score except for Bert_base_uncased, which achieved the lowest F1 score, 48.22%. The low performance score of Bert_base_uncased model might be associated with insufficient Amharic dataset durine model pre-training. The performance of the transformer models such as Am-Roberta is attributed to their knowledge obtained during their pre-training utilizing large Amharic textual datasets (Yimam et al., 2021).

Most of the errors in the *Text-Only* models are attributed to the models' weakness, OCR errors while extracting texts, and lack of context to identify the text when separated from the meme image. Figure 6.3 indicated that the basic reasons for the lower performance of the Bert_base_uncased model is mainly attributed to its inefficiency

[3] https://huggingface.co/google-bert/bert-base-uncased

[4] https://huggingface.co/rasyosef/bert-medium-amharic

[5] https://huggingface.co/rasyosef/bert-small-amharic

Model	Accuracy (%)	Precision(%)	Recall (%)	F1 (%)
Text-Only Models				
LSTM	62.00	62.00	62.00	62.00
BiLSTM	63.00	63.00	63.00	63.00
CNN	58.00	58.00	58.00	58.00
Am-Roberta	**73.99**	**74.44**	**73.99**	**74.04**
Bert-base-uncased	58.08	65.07	58.08	48.22
Afro-XLMR-base	69.19	69.09	69.19	69.08
Afro-XLMR-large	68.43	68.34	68.43	68.35
Rasyosef_Bert-medium-amharic	69.70	69.94	69.70	69.75
Rasyosef_Bert-small-amharic	70.45	70.40	70.45	70.42
Image-Only Models				
LSTM	65.00	65.00	65.00	65.00
BiLSTM	65.00	65.00	65.00	65.00
CNN	**69.00**	**69.00**	**69.00**	**69.00**
Resnet_am-roberta	66.41	66.34	66.41	66.37
Bert-base-uncased	66.67	66.83	66.67	65.89
Resnet_afro-XMLR-base	65.66	65.51	65.66	65.50
Resnet_afro-XMLR-large	67.41	67.41	67.42	67.42
Resnet_rasyosef_bert-medium-amharic	65.66	65.66	65.66	65.66
Resnet_rasyosef_bert-small-amharic	65.15	65.22	65.15	65.18
Multimodal Models				
LSTM	69.00	69.00	69.00	69.00
BiLSTM	**75.00**	**75.00**	**75.00**	**75.00**
CNN	69.00	69.00	69.00	69.00
Resnet_am-roberta	72.47	72.49	72.47	72.48
Resnet_bert-base-uncased	66.16	66.01	66.16	65.98
Resnet_afro-XMLR-base	**73.23**	**73.33**	**73.23**	**73.26**
Resnet_afro-XMLR-large	70.96	71.00	70.96	70.98
Resnet_rasyosef_bert-medium-amharic	67.68	67.61	67.68	67.30
Resnet_rasyosef_bert-small-amharic	71.46	71.38	71.46	71.35

Table 6.2: Performance of *Text-Only* and *Image-Only* Unimodal and *Multimodal* models. F1 = F1 score.

to identify hateful texts, which the model mostly failed to identify hateful instances than the non-hateful or normal once. The model detected and classified only 11.67% of hateful texts correctly, which showed that it missed 88.33% of hateful texts and wrongly classifying them as a normal text. Am_Roberta and Rasyosef_Bert-medium-amharic models looks more aggressive in detecting hateful classes while the other models seams more tolerant to classify a text as hate speech.

In *Image-Only* models, CNN outperformed all approaches including transformer-based models due to its inherent strength in extracting and learning features from two-dimensional image datasets and its capacity to learn from small datasets.

In the multimodal settings, the results presented in Table 6.2 show promising achievements compared to the *Text-Only* and *Image-Only* models. With the exception of BiLSTM, which outperformed all models with 75.00% F1 score, most of the transformer models such as Resnet_am-roberta, Resnet_afro-XMLR-base, Resnet_afro-XMLR-large and Resnet_rasyosef_bert-small-amharic achieved better results in the multimodal

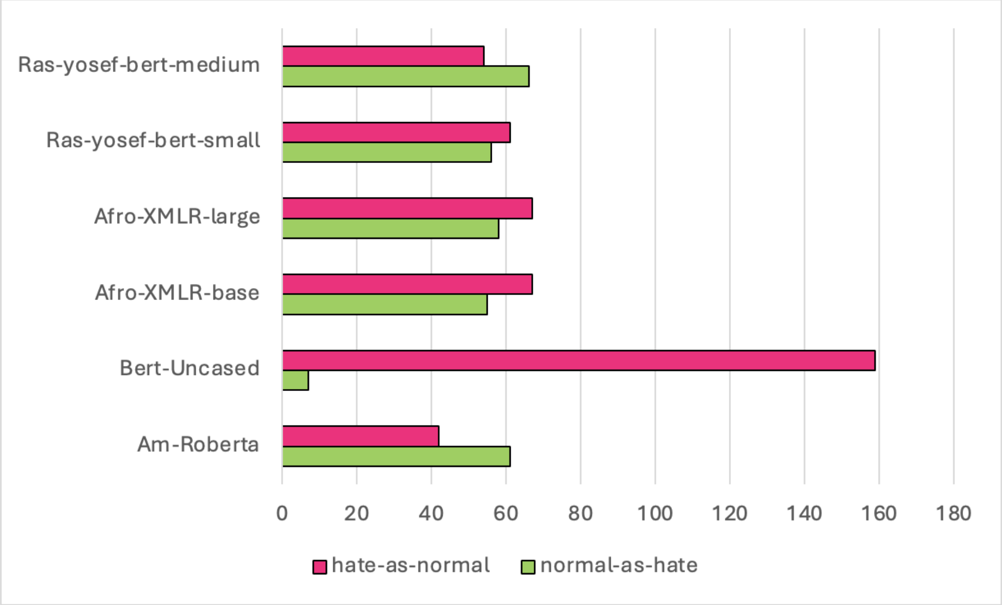


Figure 6.3: Texual model errors: wrong classification among the models per class instance, actual hate wrongly predicated as normal, or normal as hate speech within the *Text-Only* transformer models.

settings, ranging from 70.96% by Resnet_afro-XMLR-large to 73.26% F1 score with Resnet_afro-XMLR-base. Resnet_bert-base-uncased achieved only 65.98% F1 score, which is the lowest score within the multimodal settings.

As shown in Figure 6.4, Resnet_rasyosef_bert-medium-amhari and Resnet_bert-base-uncased faced more challenges to identify hateful memes effectively within the multimodal approach compared to other transformer based models. Figure 6.4 indicated these tow models have committed more errors in detecting hateful memes, specifically classifying hateful memes as normal. Most of the errors within the multimodal approach originated from the mistakes committed by the models themselves and other factors such Tesseract OCR extraction errors and lack of context for the model to identify the meme.

As illustrated in Table 6.3 and Figure 6.5 (B), it is evident that the meaning of the word written on the image is inconsistently used and varies in connotation across different geographic locations. For instance, in *Gojjam*[6] and *Wollo*[7], it represents *slave* or *servant* for men, whereas in *Gondar*[8], it signifies a *Young boy or girl.*

The third row of Table 6.3 indicated that the Tesseract OCR failed to extract the texts presented along the meme in Figure 6.5 (C). The failure of the model to classify the meme correctly might have been associated with the failure of the Tesseract OCR to accurately extract the text from the meme image. The Tesseract OCR encountered difficulties in extracting the text, which might be due to the non-straight line structure of the text arrangements within the meme images. The text on the image is *distorted* and *curved.* This structural distortion of the standard linear text presentation posed challenges for the Tesseract OCR text extractor. In Figure 6.5 (A), the model classified the meme as hate speech, likely because of the use of the derogatory word *donkey*, a term used to denote a belittling expressions towards someone in Ethiopia. Lastly,

[6] https://en.wikipedia.org/wiki/Gojjam
[7] https://en.wikipedia.org/wiki/Wollo_Province
[8] https://en.wikipedia.org/wiki/Gondar

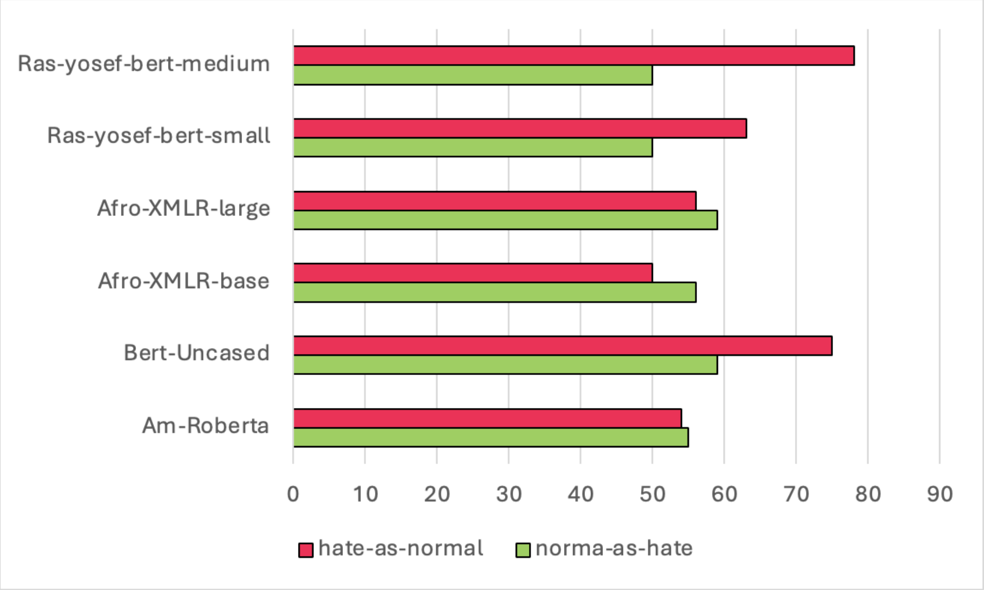


Figure 6.4: Multimodal model errors: wrong classification among the models per class instance, actual hate wrongly predicated as normal, or normal as hate speach within the Multimodal transformer models.

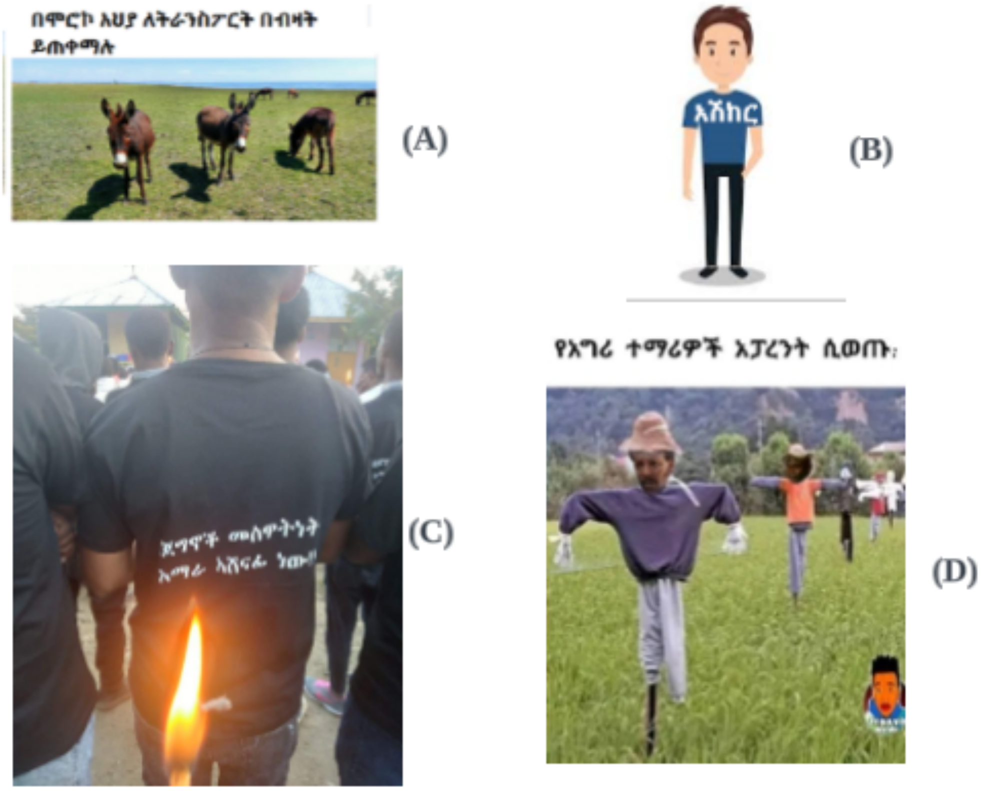


Figure 6.5: Examples for wrongly predicted memes against the gold labels: the extracted texts from each meme and their *English* translations are presented in Table 6.3.

the hateful meme presented in Figure 6.5 (D) is wrongly predicated as "normal". This might be associated with to the sarcastic nature of the particular meme image, which is specifically directed at students who study an agriculture discipline.

Meme	Tesseract OCR	Correct Texts on Memes	English Translation	Gold	Predicted
Figure 4(A)	በሞሮኮ አህያ ለትራንስፖርት በብዛት ይጠቀማሉ	በሞሮኮ አህያ ለትራንስፖርት በብዛት ይጠቀማሉ	Mostly in Morocco, donkeys are used for transportation	Normal	Hate
Figure 4(B)	አሽከር	አሽከር	manservant	Normal	Hate
Figure 4(C)	No text extracted.	በጀግኖች መስዋትነት አማራ አሸናፊ ነው	Amhara is the winner with the sacrifice of its heroes.	Hate	Normal
Figure 4(D)	የአግሪ ተማሪዎች አፓረንት ሲወጡ}	የአግሪ ተማሪዎች አፓረንት ሲወጡ	Agriculture students on apprenticeship	Hate	Normal

Table 6.3: Examples of some extracted texts from meme images and their corresponding *English* translations for the memes (A, B, C and D) presented in Figure 6.5.

6.6 Conclusion

This chapter introduced an Amharic meme dataset for multimodal detection and classification of social media memes, constituting of 2k meme images. The datasets were carefully collected from three prominent social media platforms in Ethiopia such as Facebook, X/ the former Twitter, and Telegram. We employed keywords to scrap memes from social media platforms that might have contained hateful content. Each meme has been meticulously annotated by three dedicated native Amharic speakers into *hate* or *non-hate* classes. Our meme dataset achieved a Fleiss' kappa score of 0.50 inter-annotator agreement.

We trained various models including transformers which have been pre-trained on African languages including Amharic and fine-tuned with *Text-Only*, *Image-Only* and *Multimodal* approaches. BiLSTM exceptionally outperformed all the models including transformers within the multimodal experiments which might be due to the size of the smaller meme dataset and model overfitting. BiLSTMs might generalize better than transformers, which might overfit due to their capacity to learn complex representations even from smaller datasets. In general, the models in the multimodal setup outperformed the unimodal experiments for most of the cases in which most of the transformer models showed better performance in all modalities. These insights, addressed our forth research question: *To what extent do multimodality enhance the detection of hate speech compared to unimodal approaches?*

Extending the size of the dataset and including memes from various social media data sources can be a future work. Besides, incorporating audio and video documents to build a complete multimodal hate speech datasets can advance multimodal hate speech studies. Including more hate speech categories and multiple languages in diverse cultures can also be another future work. We have released the dataset, guidelines, and models with our GitHub repository under a permissive license[9].

[9] https://github.com/uhh-lt/AmharicHateSpeech

"Pay no attention to toxic words. What people say is often a reflection of themselves, not you."

— Christian Baloga

7

Analyzing Amharic Text Detoxification Using Pre-trained Large Language Models

Contents

7.1 Introduction

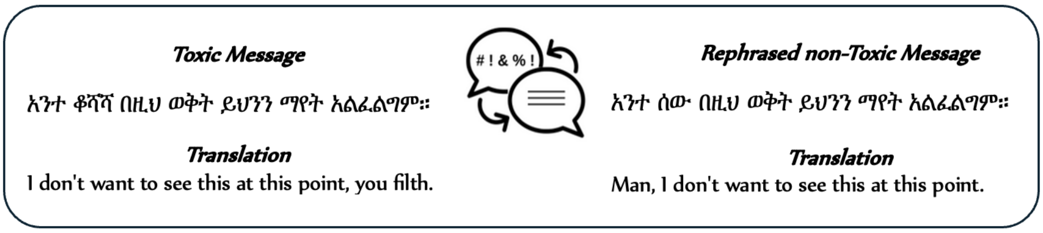
We discussed various strategies employed for collection and annotation of hate speech datasets, and explored detection approaches, the communities targeted with hate speech in Chapter 3, 4, and 5. Besides, we investigated the intensities of hatefulness and offensiveness within textual content in Chapter 5 and studied hate speech from social media memes in multimodal settings in Chapter 6.

This chapter is organized based on our work in (**Ayele**, Babakov, et al., 2024) and (Dementieva et al., 2025), which focuses on employing several text detoxification approaches to re-write and detoxify a given toxic content by keeping the general intent of the message intact.

As the spread of digital violence across online platforms continues to pollute cyberspace, many research studies involving diverse hate speech mitigation strategies have been conducted to address the issue. However, only a few studies have focused on hate speech in the context of low-resource languages, particularly those involving text detoxification strategies that utilize generative models. Most of the attempts in low-resource languages primarily focused on the detection and classification of offensive and hatred content, creating a research gap that highlights the need for more comprehensive methods. This requires the implementation of effective mitigation approaches to manage harmful interactions in the context of low-resource languages.

Hate speech mitigation strategies developed so far assist content moderators only to identify toxic messages, granting the authority to remove the toxic message entirely from such platform or block the user who created or posted the content (Yimam et al., 2024; Dementieva et al., 2024).

Recently, the emergence of generative large language models (LLMs) introduced novel mechanisms for the detection and classification of toxic messages (OpenAI, 2024; Das et al., 2024). These models have changed the way content moderators enforce mitigation measures on toxic social media messages, such as deleting the entire message instead of remove the toxic part (Demus et al., 2022; Floto et al., 2023; Logacheva et al., 2022). Rather, generative models offer text detoxification capabilities by rewriting toxic messages in a non-toxic way, while maintaining the original intent of the content, as presented in Figure 7.1. This approach offer content moderators multiple options to employ appropriate actions, re-writing messages that can be detoxified or removing messages that cannot be detoxified.



Toxic Message

አንተ ቆሻሻ በዚህ ወቅት ይህንን ማየት አልፈልግም።

Translation

I don't want to see this at this point, you filth.

Rephrased non-Toxic Message

አንተ ሰው በዚህ ወቅት ይህንን ማየት አልፈልግም።

Translation

Man, I don't want to see this at this point.

Figure 7.1: Toxic text re-writing into a non-toxic neutral way.

In (**Ayele**, Babakov, et al., 2024), multilingual text detoxification is one of the four shared tasks that we organized at the Conference and Labs of the Evaluation Forum (CLEF-2024), which presented *Paradetox*, a parallel text detoxification dataset for nine languages, including Amharic.

In this chapter, we further extend (**Ayele**, Babakov, et al., 2024) and assess the challenges of annotating text detoxification datasets for low-resource languages such as Amharic and investigate the performance of LLMs in automatically identifying toxic terms and detoxifying the entire message content. We have annotated a total of 3,120 tweets to create a detoxification dataset and explore models for re-writing textual documents in order to detoxify, without losing the original intent of such messages. We further explore several models in (Dementieva et al., 2025) and analyze results for nine languages including Amharic. For this chapter, we include Amharic language related stuff from (Dementieva et al., 2025), which accounts our contribution in the paper.

7.2 Related Work

These days, text detoxification has gained a raising interest among social NLP researchers working on hate and offensive speech detection and social media content moderators. The study conducted by X Wu et al. (2019) introduced *Cond-BERT*, a novel conditional BERT contextual data augmentation method for labeled sentences, which utilizes a masked language model (MLM) to replace toxic words or phrases found in sentences with their non-toxic alternatives.

Another study conducted by Dale et al. (2021) further explored *Cond-BERT* and introduced a new model called *ParaGeDi* which paraphrases toxic text inputs to neutralize explicitly offensive content while preserving the meaning of the original message. Dale et al. (2021) presented the first large-scale paraphrasing models and comparative study of re-writing and style transfer models on the task of toxicity removal.

The work presented by Hallinan et al. (2023) proposed *MARCO*, a text detoxification architecture that integrates controllable generation and text rewriting methods utilizing a combination of experts and auto-encoder language models. *MARCO* is implemented based on the probabilities of likelihood between a non-toxic language model (the expert), fine-tuned on data with desirable attributes, and a toxic language model (the anti-expert), fine-tuned on data with undesirable attributes, which aims to detect candidate toxic words to mask and replace with non-toxic synonyms.

Floto et al. (2023) launched an integrated model called *DiffuDetox*, which consisted of two intertwined components. The first part is the conditional component that takes offensive textual inputs as a condition and tries to reduce toxicity by producing a diverse set of detoxified sentences. The second component, the unconditional component, is mainly aimed at ensuring the fluency of the rephrased input message while neutralizing its toxicity.

In an attempt to address text detoxification task, Dementieva, Ustyantsev, et al. (2021) has manually prepared paraphrases for over 1.2k toxic sentences. The sentences are collected from Reddit, Twitter, and Wikipedia discussion pages and produced parallel corpora of toxic and non-toxic sentences utilizing Yandex Toloka crowd sourcing platform. Dementieva, Ustyantsev, et al. (2021), explored the most toxic regions of a sentence based on multiple detoxification attempts provided by different annotations. The work proposed by Mukherjee et al. (2023), attempted to automatically transform toxic sentences into non-toxic counterpart while preserving content and maintaining fluency, extending the study presented by Dementieva, Ustyantsev, et al. (2021).

Dementieva et al. (2023) have proposed a system that can perform text translation and detoxification simultaneously. Besides, the authors have tried to explore the potential methods for cross-lingual text style transfer (TST), specifically focusing on text detoxification tasks, and introduced an automatic text detoxification evaluation metrics which achieved higher correlations with human expert judgments. Dementieva et al. (2023) presented state of the art results in text detoxification across models such as *BART* and *T5*, fine-tuned for text detoxification.

A recent study conducted by Dementieva et al. (2024) introduced *MultiParaDetox*, a multilingual parallel text detoxification datasets for non-English languages such as Russian, Ukrainian and Spanish. Despite the authors achieved the highest style transfer accuracy (STA) results with *LLaMa-7b* model across all languages, the content of the outputs are just random due to hallucination. Dementieva et al. (2024) showed that the baseline models achieved higher results with a specific metrics and lower values in others. However, *mBART* model fine-tuned with the new parallel detoxification corpus works well in general and do not fail in any of the evaluation metrics such as style transfer accuracy, content similarity, fluency and the joint score.

Most of the attempts conducted in low-resource languages, including Amharic, mainly focused on only detecting or classifying hate speech, which can only offer limited assistance for content moderators in managing abusive social media posts and comments. These studies often overlook detoxification approaches that provide new strategies to mitigate harmful online communications (Pavlopoulos et al., 2017). For instance, the studies conducted by Abebaw et al. (2022a, 2022b), Mossie and Wang (2018), Tesfaye and Kakeba (2020), **Ayele**, Dinter, et al. (2022), and **Ayele**, Yimam, et al. (2023) primary focused on detecting and classifying hate speech, despite **Ayele**, Jalew, et al. (2024) further examined varying levels of hate and offensive speech intensities, and the subtle targets of hatred messages.

Text detoxification, which involves changing the style of toxic language to a more neutral or positive form, is an important task in combating polarized communications across on online platforms (Logacheva et al., 2022; Dementieva, Moskovskiy, et al., 2021; M Tran et al., 2020). Particularly, text detoxification plays a crucial role in protecting vulnerable groups, such as children and women from direct exposure to harmful content. By converting toxic messages into neutral language, text detoxification not only mitigates immediate harm but also contributes to creating a safer and more favorable online environment. This proactive strategy shifts the focus from merely detecting and removing harmful content to adopting a more efficient and sustainable approach to addressing online toxicity (Yimam et al., 2024).

In this chapter, we explored the specific challenges in annotating toxic content while creating a parallel detoxification corpus as part of **Ayele**, Babakov, et al. (2024) and (Dementieva et al., 2025). Besides, we examine how large language models (LLMs) are effectively utilized in identifying toxic content and detoxify the text a neutral manner in a low-resource language settings such as Amharic.

7.3 Data Collection and Annotation

The following subsections present the overview of data collection strategies and annotation procedures utilized in this chapter.

7.3.1 Data Collection

We compiled a new detoxification dataset by merging two existing Amharic hate speech datasets introduced in (**Ayele**, Yimam, et al., 2023; **Ayele**, Dinter, et al., 2022). The original datasets are collected from Twitter employing diverse data collection and sampling strategies. The datasets are publicly accessible in GitHub repository[1].

The first dataset, presented in **Ayele**, Dinter, et al. (2022), was annotated as *hate*, *offensive*, *normal*, and *unsure* class labels with three native speakers using Yandex Toloka crowdsourcing platform. This dataset achieved a relatively low inter-annotator agreement with a Fleiss' kappa score of 0.34. The gold labels were determined through a majority voting scheme.

[1] https://github.com/uhh-lt/AmharicHateSpeech

The second dataset, presented by **Ayele**, Yimam, et al. (2023), employed an expert-guided annotation process in which the annotators received better training and close supervision from the task organizers and annotation experts. Each tweet was annotated by two native speakers and curated by an adjudicator to decide the gold labels on the disputed annotations, achieving a Fleiss' kappa score of 0.48.

We selected portions of both datasets with the label class *offensive*. A total of 3.1k tweets are collected and utilized to create a new parallel text detoxification dataset for Amharic language. Each tweet is re-annotated and paraphrased by well trained annotators.

7.3.2 Data Annotation

We customized and utilized the *Potato-POrtable Text Annotation TOol* for annotating a parallel text detoxification dataset (Pei et al., 2022). The annotation process consisted of two steps: classification and detoxification, as illustrated in Figure 7.2.

In the first step, annotators were requested to read and classify each tweet into one of the four categories: *detoxifiable*, *hate*, *normal*, and *indeterminate*. In the second step, annotators rephrase tweets, which they labeled as *detoxifiable* in the first step to make the toxic content more neutral as shown in Figure 7.2.

We conducted a pilot annotation consisting of 125 posts with three native speakers and achieved a 0.35 inter-annotator agreement score on the classification task. In the pilot annotation, at least two of the annotators agreed on 54% of the tweets as detoxifiable and provided a non-toxic rephrased versions.

In the main annotation task, 3k tweets were carefully annotated by well trained and expert-level native speakers. In the first annotation step, annotators identified 1.5k tweets as detoxifiable instances and provided respective paraphrases. The remaining 1.5k tweets were classified as non-detoxifiable, presenting considerable challenges for the annotators to rephrase the texts.

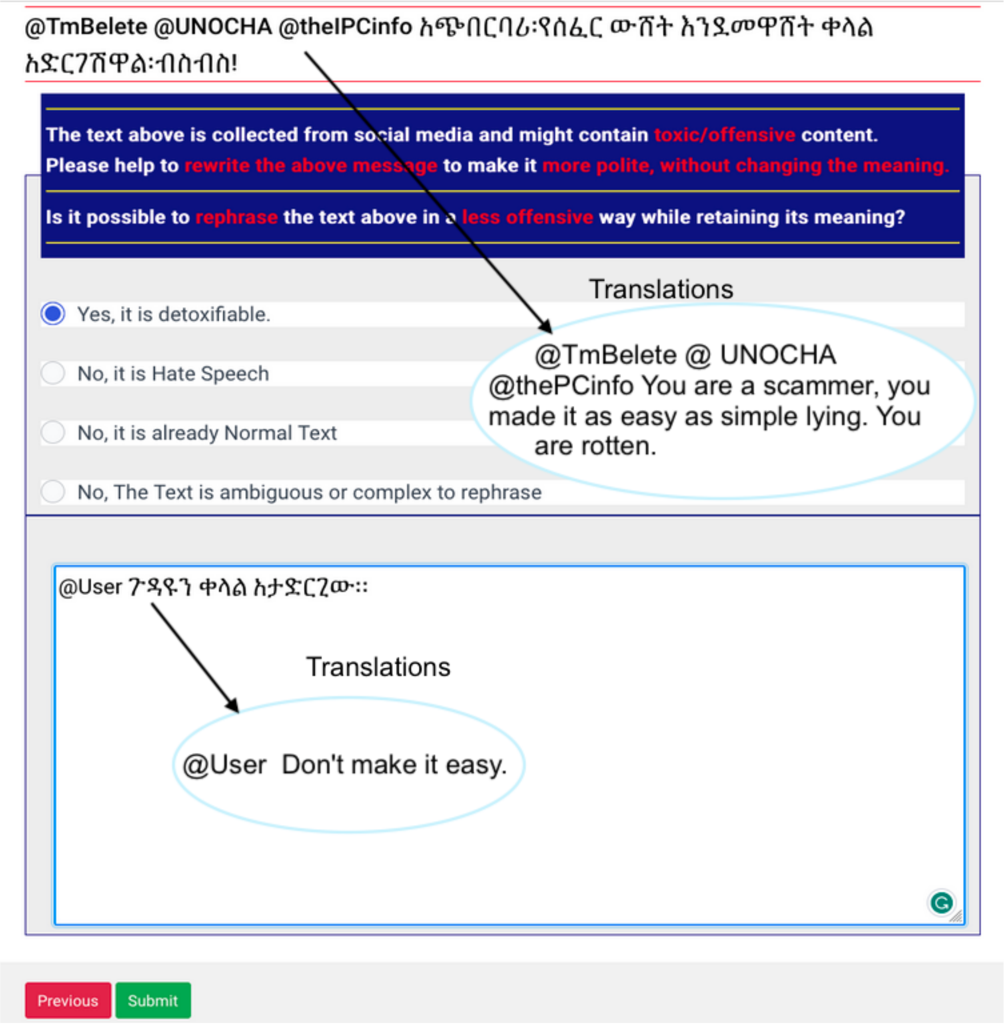
7.4 Results and Discussion

In this section, we present the evaluation of various tasks including classification, explainability analysis and rephrasing for detoxification. The classification task determines whether a tweet can be detoxified or not while explainability provide insights into why a given tweet is considered toxic. The detoxification results demonstrate how LLMs perform in rewriting toxic messages, particularly in low-resource languages.

7.4.1 Toxic Keyword and Detoxification Analysis

We observe that the toxicity of messages within the rephrased texts is reduced from high toxicity to medium or low levels, confirming that the paraphrases have been effectively detoxified. We have also extracted the most frequent negative and toxic phrases from the toxic texts as presented in Figure 7.4

In addition to extracting toxic words, we also analyzed the way how text detoxification was performed as indicated in Table 7.1. We searched for lemmas that reflect various editorial actions such as *delete*, *rephrase*, *replace*, or *insert* using the following prompt template: Answer shortly, how this text: {toxic text} was rephrased into this: {detoxified text}.



@TmBelete @UNOCHA @theIPCinfo አጭበርባሪ፡የሰፈር ውሸት እንዷመዋሸት ቀላል አድርገሽዋል፡ብስብስ!

The text above is collected from social media and might contain toxic/offensive content.
Please help to rewrite the above message to make it more polite, without changing the meaning.

Is it possible to rephrase the text above in a less offensive way while retaining its meaning?

● Yes, it is detoxifiable.

○ No, it is Hate Speech

○ No, it is already Normal Text

○ No, The Text is ambiguous or complex to rephrase

Translations

@TmBelete @ UNOCHA @thePCinfo You are a scammer, you made it as easy as simple lying. You are rotten.

@User ጉዳዩን ቀላል አታድርገው።

Translations

@User Don't make it easy.

Figure 7.2: Potato GUI for two-level detox annotation: Toxic classification and toxic rewriting tasks.

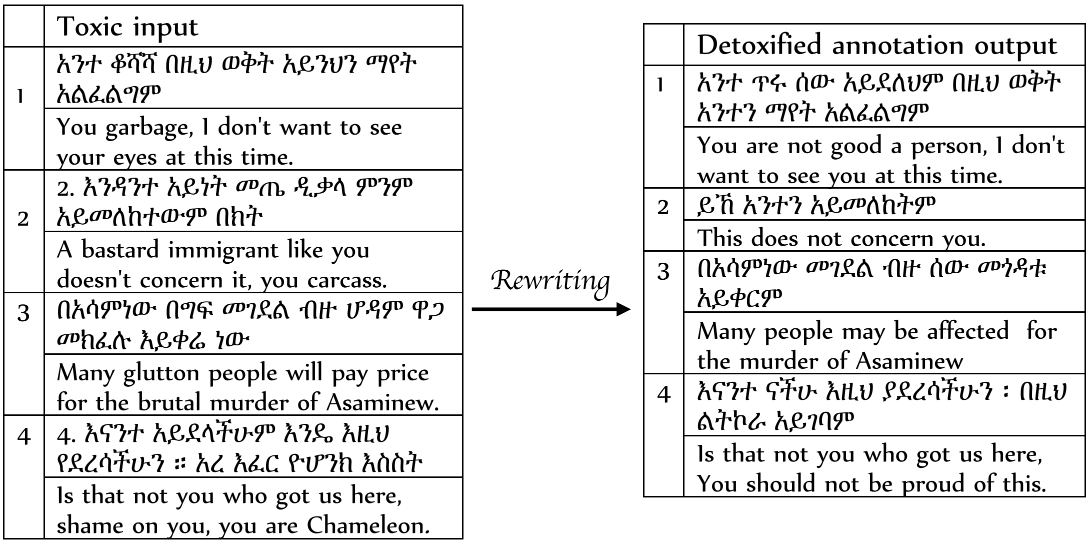
The addition of new phrases is quite rare, where the main modifications typically involve the removal or rephrasing of toxic phrases. As a result, localized editing techniques that incorporate suitable and fluent substitutions are generally effective for successful text detoxification.

Operation	Deleted	Rephrased	Inserted
Performance	45%	44%	11%

Table 7.1: Percentage of toxic phrases "Deleted", "Rephrased", or new non-toxic parts "Inserted" to achieve text detoxification.

7.4.2 Classification

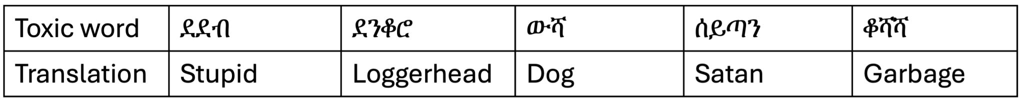
In the first place, we built classification models to classify tweets as detoxifiable or non-detoxifiable. Non-detoxifiable tweets represent posts that cannot be rephrased easily for various reasons. Most of these non-detoxifiable tweets contain harmful content



	Toxic input
1	አንተ ቆሻሻ በዚህ ወቅት አይንህን ማየት አልፈልግም
	You garbage, I don't want to see your eyes at this time.
2	2. እንዳንተ አይነት መጤ ዲቃላ ምንም አይመለከተውም በክት
	A bastard immigrant like you doesn't concern it, you carcass.
3	በአሳምነው በግፍ መገደል ብዙ ሆዳም ዋጋ መክፈሉ አይቀሬ ነው
	Many glutton people will pay price for the brutal murder of Asaminew.
4	4. እናንተ አይደላችሁም እንዴ እዚህ የደረሳችሁን ። አረ እፈር የሆንክ እስስት
	Is that not you who got us here, shame on you, you are Chameleon.

	Detoxified annotation output
1	አንተ ጥሩ ሰው አይደለህም በዚህ ወቅት አንተን ማየት አልፈልግም
	You are not good a person, I don't want to see you at this time.
2	ይኸ አንተን አይመለከትም
	This does not concern you.
3	በአሳምነው መገደል ብዙ ሰው መጎዳቱ አይቀርም
	Many people may be affected for the murder of Asaminew
4	እናንተ ናችሁ እዚህ ያደረሳችሁን ፡ በዚህ ልትኮራ አይገባም
	Is that not you who got us here, You should not be proud of this.

Figure 7.3: Sample toxic input texts and their corresponding detoxified outputs after annotators rewrite to remove the toxic content.



Toxic word	ደደብ	ደንቆሮ	ውሻ	ሰይጣን	ቆሻሻ
Translation	Stupid	Loggerhead	Dog	Satan	Garbage

Figure 7.4: Top-5 extracted toxic keywords from toxic parts.

that remains dangerous to readers even after being rephrased. As illustrated in Table 7.2, *AmRoBERTa* outperformed the other models in classification tasks. Notably, *AmRoBERTa* exhibited superior performance across most metrics, indicating its robustness in identifying content that can be safely detoxified.

Classifier	Precision	Recall	F1 score
Am-RoBERTa	**69.57%**	**73.37%**	**70.01%**
Afro-xlmr-large	69.07%	73.53%	67.75%
Bert-medium-amharic	66.13%	69.90%	64.62%

Table 7.2: Model performance in identifying detoxifiable tweets from non-detoxifiable ones.

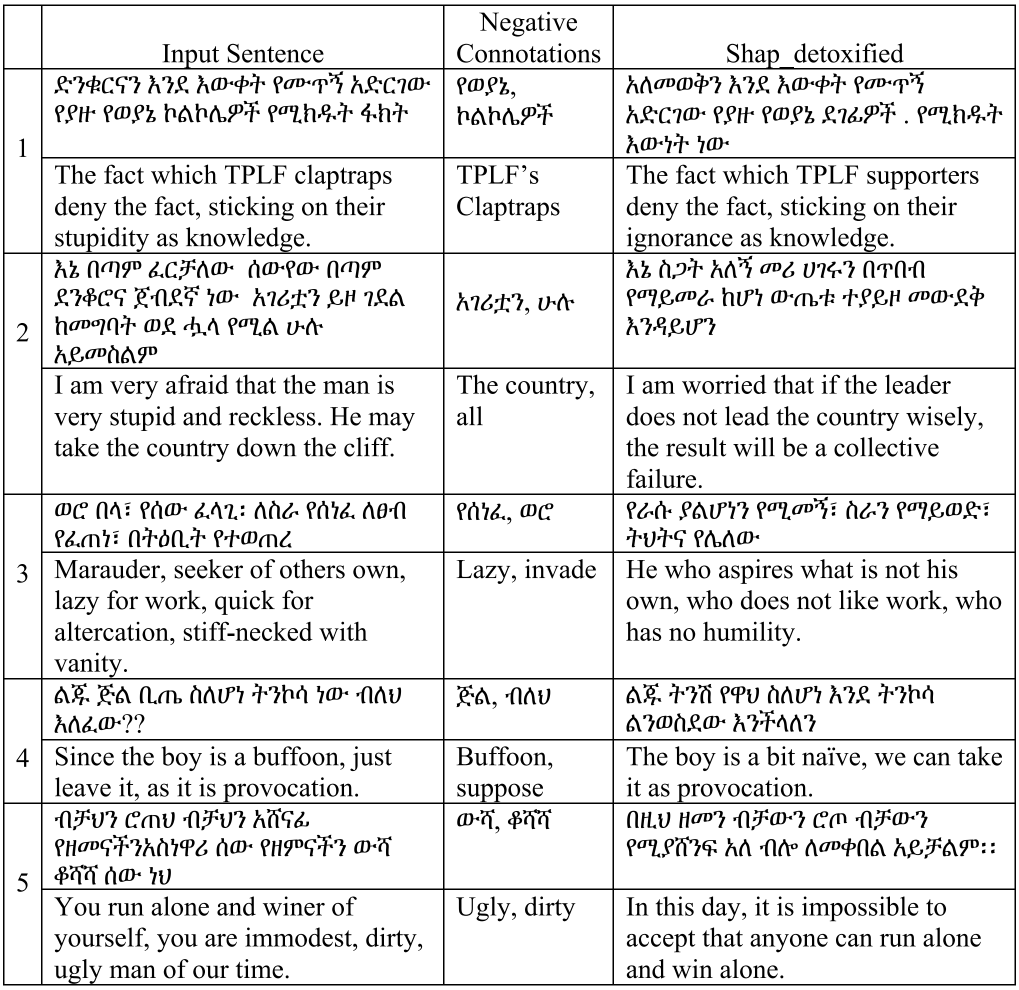
7.4.3 Explainability

Through utilizing SHAP (Lundberg and Lee, 2017) and GPT-4 models, we attempted to identify the parts of tweets that contain toxicity within the entire message content.

GPT-4o mini	SHAP
76.33%	54.83%

Table 7.3: Performance of models in detecting toxic words for explainable analysis.

Both models aimed to identify words or phrases that potentially contribute to toxic nature of the message. We requested the SHAP and GPT-4 models to pick two most toxic words while rephrasing toxic input tweets as presented in Table 7.4 and Table 7.5, respectively.



	Input Sentence	Negative Connotations	Shap_detoxified
1	ድንቁርናን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ኮልኮሌዎች የሚክዱት ፋክት	የወያኔ, ኮልኮሌዎች	አለመወቅን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ደገፊዎች . የሚክዱት እውነት ነው
	The fact which TPLF claptraps deny the fact, sticking on their stupidity as knowledge.	TPLF's Claptraps	The fact which TPLF supporters deny the fact, sticking on their ignorance as knowledge.
2	እኔ በጣም ፈርቻለው ሰውየው በጣም ደንቆሮና ጀብደኛ ነው አገሪቷን ይዞ ገደል ከመግባት ወደ ኋላ የሚል ሁሉ አይመስልም	አገሪቷን, ሁሉ	እኔ ስጋት አለኝ መሪ ሀገሩን በጥበብ የማይመራ ከሆነ ውጤቱ ተያይዞ መውደቅ እንዳይሆን
	I am very afraid that the man is very stupid and reckless. He may take the country down the cliff.	The country, all	I am worried that if the leader does not lead the country wisely, the result will be a collective failure.
3	ወሮ በላ፣ የሰው ፈላጊ፡ ለስራ የሰነፈ ለፀብ የፈጠነ፣ በትዕቢት የተወጠረ	የሰነፈ, ወሮ	የራሱ ያልሆነን የሚመኝ፣ ስራን የማይወድ፣ ትህትና የሌለው
	Marauder, seeker of others own, lazy for work, quick for altercation, stiff-necked with vanity.	Lazy, invade	He who aspires what is not his own, who does not like work, who has no humility.
4	ልጁ ጅል ቢጤ ስለሆነ ትንኮሳ ነው ብለህ እለፈው??	ጅል, ብለህ	ልጁ ትንሽ የዋህ ስለሆነ እንደ ትንኮሳ ልንወስደው እንችላለን
	Since the boy is a buffoon, just leave it, as it is provocation.	Buffoon, suppose	The boy is a bit naïve, we can take it as provocation.
5	ብቻህን ሮጠህ ብቻህን አሸናፊ የዘመናችንአስነዋሪ ሰው የዘምናችን ውሻ ቆሻሻ ሰው ነህ	ውሻ, ቆሻሻ	በዚህ ዘመን ብቻውን ሮጦ ብቻውን የሚያሸንፍ አለ ብሎ ለመቀበል አይቻልም፡፡
	You run alone and winer of yourself, you are immodest, dirty, ugly man of our time.	Ugly, dirty	In this day, it is impossible to accept that anyone can run alone and win alone.

Table 7.4: SHAP detecting toxic words.

Figure 7.5 illustrates how the SHAP model, based on the hate speech classification model from **Ayele**, Yimam, et al. (2023), highlights toxic terms within a given message. As shown in Table 7.3, GPT-4 was able to identify the majority of toxic terms or phrases. We then compared and computed the accuracy by analyzing whether these terms were removed within the human-paraphrased tweet column.

7.4.4 Detoxification with Automatic Evaluation

In (**Ayele**, Babakov, et al., 2024) and (Dementieva et al., 2025), we implement an automatic evaluation setup to evaluate the outputs based on three parameters, namely style of text, content preservation, and conformity to human references, which all combined into a more cumulative final metrics called, the *Joint* score. The details are:

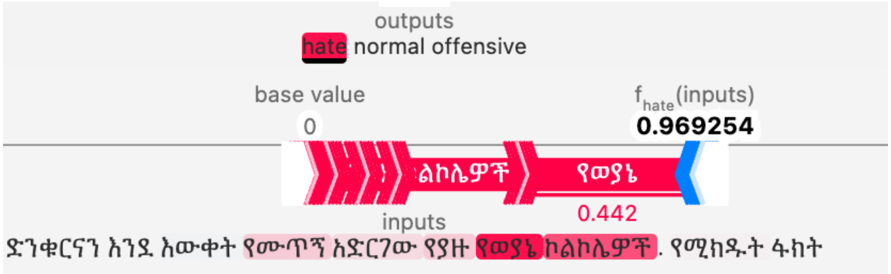


Figure 7.5: Output examples for explainable toxicity detection.

	Input Sentence	Negative Connotations	Gpt4_detoxified
1	ድንቁርናን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ኮልኮሌዎች የሚክዱት ፋክት	ድንቁርናን, ኮልኮሌዎች	አለመወቅን እንደ እውቀት የሙጥኝ አድርገው የያዙ የወያኔ ደገፊዎች . የሚክዱት እውነት ነው
	The fact which TPLF claptraps deny the fact, sticking on their stupidity as knowledge.	stupidity, claptraps	The fact which TPLF supporters deny the fact, sticking on their ignorance as knowledge
2	እኔ በጣም ፈርቻለው ሰውየው በጣም ደንቆሮና ጀብደኛ ነው አገሪቷን ይዞ ገደል ከመግባት ወደ ኋላ የሚል ሁሉ አይመስልም	ደንቆሮና, ጀብደኛ	እኔ ስጋት አለኝ መሪ ሀገሩን በጥበብ የማይመራ ከሆነ ውጤቱ ተያይዞ መውደቅ እንዳይሆን
	I am very afraid that the man is very stupid and reckless. He may take the country down the cliff.	Stupid, reckless	I am worried that if the leader does not lead the country wisely, the result will be a collective failure.
3	ወሮ በላ፣ የሰው ፈላጊ፡ ለስራ የሰነፈ ለፀብ የፈጠነ፣ በትዕቢት የተወጠረ	ፈላጊ, የተወጠረ	የራሱ ያልሆነን የሚመኝ፣ ስራን የማይወድ፣ ትህትና የሌለው
	Marauder, seeker of others own, lazy for work, quick for altercation, stiff-necked with vanity.	seeker, Stiff-necked	He who aspires what is not his own, who does not like work, who has no humility.
4	ልጁ ጅል ቢጤ ስለሆነ ትንኮሳ ነው ብለህ እለፈው??	ጅል, ትንኮሳ	ልጁ ትንሽ የዋህ ስለሆነ እንደ ትንኮሳ ልንወስደው እንችላለን
	Since the boy is a buffoon, just leave it, as it is provocation.	Buffoon, Provocation	The boy is a bit naïve, we can take it as provocation.
5	ብቻህን ሮጠህ ብቻህን አሸናፊ የዘመናችንአስነዋሪ ሰው የዘምናችን ውሻ ቆሻሻ ሰው ነህ	አሸናፊ, ቆሻሻ	በዚህ ዘመን ብቻውን ሮጦ ብቻውን የሚያሸንፍ አለ ብሎ ለመቀበል አይቻልም፡፡
	You run alone and winer of yourself, you are immodest, dirty, ugly man of our time.	Winner, Dirty	In this day, it is impossible to accept that anyone can run alone and win alone.

Table 7.5: GPT4 detecting toxic words.

- **Style Transfer Accuracy (STA)** a measure that ensures that the generated text is indeed more non-toxic. We have utilized 5k samples, 2.5k toxic and 2.5k neutral instances from Amharic hate speech datasets (**Ayele**, Yimam, et al., 2023; **Ayele**, Dinter, et al., 2022) and fine-tuned XLM-R (Conneau et al., 2020) for the binary toxicity classification task, which the model evaluates the degree of non-toxicity in the texts.

- **Content Similarity (SIM)** evaluates the cosine similarity between LaBSE[2] embeddings (Feng et al., 2022) of the source texts and the generated texts.
- **Fluency (ChrF1)** is a metrics specifically utilized to estimate the proximity of the detoxified texts with respect to human references and their fluency (Post, 2018). The ChrF1 score from `sacrebleu` library is implemented to evaluate the fluency of the detoxified texts.
- **Joint score (J)** is a measure that presents the aggregate evaluation of the three above metrics:

  $$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{STA}(y_i) \cdot \mathbf{SIM}(x_i, y_i) \cdot \mathbf{ChrF1}(x_i, y_i),$$

  where :
  $\mathbf{STA}(y_i)$, $\mathbf{SIM}(x_i, y_i)$, $\mathbf{ChrF1}(x_i, y_i) \in [0, 1]$ for each text detoxification output y_i.

Baseline Models for Automatic Evaluation

We explored several unsupervised and supervised text detoxification approaches together with a baseline models for comparison.

Duplicate. Duplicate assumes that the output text is a copy-paste of the input sentence. It is a trivial baseline with a default similarity (SIM) score of 1.0 (or 100%).

Delete. The delete method removes offensive terms utilizing the manually compiled list of vulgar words based on the previous studies in (**Ayele**, Yimam, et al., 2023; **Ayele**, Dinter, et al., 2022).

Backtranslation. We perform translation of non-English texts into English with NLLB (Costa-jussà et al., 2022) instances and then compute detoxification with the BART model which is fine-tuned on English ParaDetox train dataset (Logacheva et al., 2022) in order to set a more sophisticated unsupervised baseline model.

CondBERT. CondBERT adapts one of the MLM-based unsupervised methods proposed in (Dale et al., 2021) and mBERT in (Devlin et al., 2019) as a base models, and utilizes these methods to generate list of substitutes selecting non-toxic ones.

Fine-tuned LM on Translated Data. We also tried to obtain synthetic parallel corpora by translating selected 400 English ParaDetox samples to Amharic languages and utilized mBART model (Liu et al., 2020) for the translation step. We tuned the mBART (Tang et al., 2021) on the obtained data.

Fine-tuning on the parallel data. We have utilized our parallel detoxification corpus, which we have manually created for the Amharic language and fine-tuned with the multilingual text-to-text generation model, mBART-Large.

[2] https://huggingface.co/sentence-transformers/LaBSE

GPT-4 few-shot prompting. We have conducted experiments with GPT-4 in few-shot prompting approach, providing instructions to the models and requesting to detoxify the input sentences without changing the original meaning of the messages.

Metrics	STA	SIM	ChrF	J-score
Human Reference	89.30%	68.30%	100.00%	60.10%
Unsupervised Approaches				
Duplicate	42.60%	100.00%	48.50%	21.60%
Delete	53.90%	97.90%	48.60%	26.90%
Backtranslation	81.90%	61.80%	13.50%	7.50%
condBERT	99.80%	16.90%	0.70%	0.30%
Supervised Approaches				
mBART-Translated	50.10%	87.50%	39.10%	17.80%
mBART-mParaDetox	50.60%	91.50%	41.20%	20.40%
LLM-based Approaches				
GPT-4 few-shot	46.70%	94.60%	45.30%	20.50%

Table 7.6: Automatic evaluation results of text detoxification experiments.

Table 7.6, presented the text detoxification experimental results in metrics such as style transfer accuracy, content similarity, fluency and the joint score. Surprisingly, the Delete method outperformed other approaches. This may be due to the nature of the language as presented in Table 7.1, where detoxification relies heavily on paraphrasing. Since the proposed methods still struggled with appropriate paraphrasing, the Delete method, which removes toxic content without rephrasing presented the best performance results. For the supervised approaches, the mBART model, specifically mBART-mParaDetox, which is fine-tuned on human-curated data, outperformed the other model, mBART-Translated, which is fine-tuned on translated data. Toxic keywords that reflect cultural nuances contribute significantly to the performance of models trained on manually curated data.

GPT-4, with few-shot prompting, shows better results in achieving content similarity, fluency and joint-score despite the detoxified models outputs yielded inaccurate and incoherent generations as depicted in Table 7.7.

7.5 Error Analysis of Models Results

We have randomly selected 10% of the test set instances, similar across tasks, for assessing the performance of models in toxic word detection and the paraphrasing quality. An independent expert who has not been participated in *paradetox* annotation task was employed to evaluate the quality of human annotations (detoxification), and the model outputs in both toxic word detection and sentence paraphrasing tasks.

As indicated in Figure 7.6, the SHAD model has fully detected toxic words in 33% and partially detected such words in 25% of the test instances that have negative connections in the original input text. GPT-4 has shown better performance in toxic word detection, which fully and partially detected toxic words in 50.8% and 32.8% of the cases, respectively. Table 7.5 and Table 7.4, presented examples that consist of original

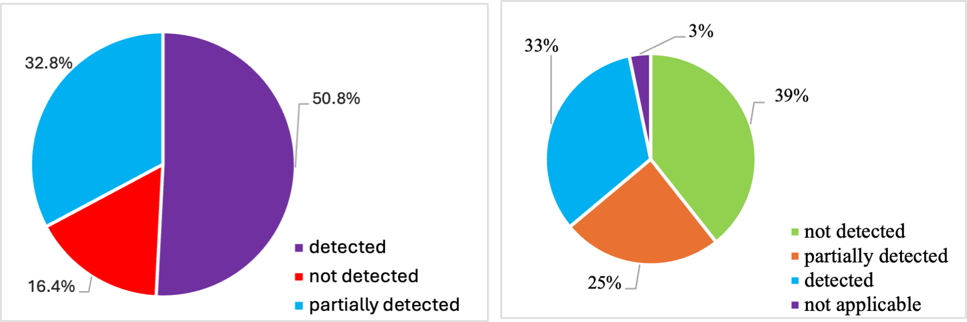
	Iput Sentence	GPT-4 Detoxified	Human Detoxified	Remark
1	ፈሳም (ፈሪ) ሰው አይመቸኝም ለሀገርም አይጠቅምም	ሁሉም ሰው ለሀገር ጥሩ ነው የሚሆን	ፈሪ የሆነ ሰው አልወድም ለሀገርም ይጠቅማል ብዬ አላስብም	The model generates inaccurate outputs & Expert do not detoxify well.
	A cowardly person is not comfortable, and it is not good for the country.	Everyone is good for the country.	I don't like a coward and I don't think it is good for the country.	
2	የሆንሽ ወሸላ ምን ዶለሽ እዚህ ውስጥ . አፉፉቱ ቱልቱላ	እባክህ ያለውን እንደ ማስታወሻ ይቀበሉ	አይ አንቺ ምን እዚህ ውስጥ ከተተሽ	The model generates inaccurate and incoherent outputs.
	You are uncircumcised, Why you are here, talkative	Please accept it as a note.	What are you doing here?	
3	ወሮ በላ፣ የሰው ፈላጊ፡ ለስራ የሰነፈ ለፀብ የፈጠነ፣ በትዕቢት የተወጠረ	ወሮ በላ፣ እውነተኛ የሰው ፈላጊ ስለ ስራ የተሰነፈ እና የፀብ የፈጠነ ነው።	የራሱ ያልሆነን የሚመኝ፣ ስራን የማይወድ፣ ትህትና የሌለው	The model don’t detoxify compound words well and also generates incoherent outputs.
	Marauder, seeker of others own, lazy for work, quick for altercation, stiff-necked with vanity.	Marauder , a true man, seeking others own, lazy and quick to anger.	He who desires what is not his own, who does not like work, who has no humility	
4	አንተ ነጫጭባ ሽማግሌ የግልገሎችህ መሸነፍ አላስተኛ አለህ አይደል? ቆይ ጠብቀኝ	እቅፍ ይቅርታ፣ እንደ ግልገሎች አስተናግድ በዚህ ጊዜ ላይ ይቀጥል	አንተ አዛውንት ሰውዬ የልጆቻህን መሸነፍ ማመን አልቻልክም? ቆይ ታለህ	The model generates inaccurate and incoherent outputs.
	You whitish man, do the defeat of your colts disturbed you? Wait me.	Hugs sorry, treat as your colts, continue at this time.	You old man, don’t you belive the defeat of your children? You will see	
5	እኔኮ ደና ሰው መስላኝ ነው ጅል መሆኗን መች አውቄ ፈያሜታዬ	እኔ ደና ሰው ነኝ እና እንደ አንድ ሰው እንደ ምን እንደ ተወዳጅ ነኝ	እኔኮ ደህና ሰው መስላኝ ነው ነገር ቶሎ የማይገባት መሆኗን መች አውቄ ፈያሜታዬ	The model generates inaccurate and incoherent outputs
	I thought she is a good person, I didn’t know that she is a fool.	I am a friendly person and I am as a person, as lovely person.	I thought she is a good person, I didn’t know that she can not understnd easily.	

Table 7.7: GPT-4 detoxified samples.

input sentences, toxic words that show negative connections, and the detoxified outputs for both GPT-4 and SHAD models, respectively.

The independent human evaluator examines both model outputs and human annotations to explore whether the input texts are properly rephrased and detoxified, maintaining the meaning of the original content. The model was able to rewrite and detoxify properly for only 11.7% of the test instances while 5% of the model outputs showed unintelligible rephrases that do not convey a clear meaning. As shown in Figure 7.7-a), the model faced significant difficulties in generating accurate and coherent outputs on the majority of the input texts, which account for 83.3% of the model outputs.

Our evaluation on human expert detoxification outputs revealed that, about 20% of the instances are not properly rephrased and detoxified even by human annotators, which might have effects on the performance of the models. As indicated in 7.7-b), 6.7% of texts detoxified by human annotators are not entirely detoxified while 1.7% are only partly detoxified. Additionally, it is also indicated that 6.7% of human rephrased



(a) GPT-4 toxic word detection.

(b) SHAP toxic word detection.

Figure 7.6: Error analysis of GPT-4 and SHAP toxic word detection performance in explaining the negative connections.

instances are unclear and confusing while 5% of the remainder have changed the meaning of the original input content.

Figure 7.7-c), presents the potential reasons for the inaccurate and incoherent generations of model outputs, which include sarcasm, idiomatic expressions, incomplete sentences, usage of informal languages, compound words. Our evaluation revealed that major reason for the inaccurate generations, which accounts for 55%, is not known. As shown in Figure 7.7-d), the GPT-4 is effective in removing toxic terms while rephrasing toxic inputs despite its inaccurate and incoherent generations, which accounts for 73.3% of the model outputs. The model outputs contain toxic content only for 13.3% of the inputs.

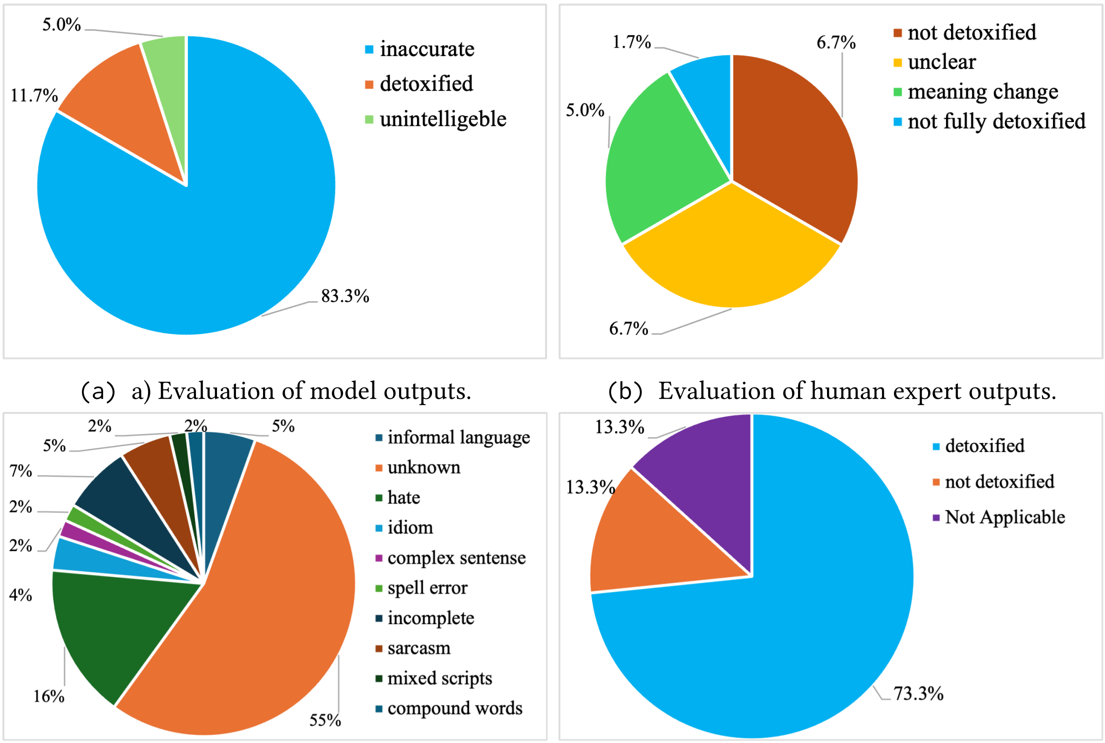
Table 7.7, presents sample outputs comparing GPT-4 detoxification results with human expert annotations. The findings show that many of the detoxification outputs generated by GPT-4 significantly differ from human expert annotations, largely due to notable issues with inaccuracies and incoherence in model generated outputs, particularly in low-resource languages like Amharic.

7.6 Conclusion

This chapter mainly tackles the challenges of text detoxification in low-resource languages, specifically in Amharic. We created a new detoxification parallel dataset from existing hate speech datasets by conducting comprehensive data selection and annotation procedures. In this chapter, three key tasks were evaluated, namely classification of texts into detoxifiable or not-detoxifaible, toxic word detection to justify why a message is toxic (explainability), and paraphrasing toxic textual inputs in a more neutral way (detoxification).

Am-RoBERTa outperformed the other models in most classification metrics. Using SHAP and GPT-4, we identified toxic terms, with GPT-4 being more effective but still facing significant issues in detoxifying messages due to frequent inaccurate and incoherent model generated outputs, addressing our fifth research questions: *What challenges do large language models (LLMs) face in Amharic text detoxification task?*

The baseline Delete method in the unsupervised approach outperformed all models utilizing available keywords as toxicity in social media mainly employ common toxic phrases in Amharic.



(a) a) Evaluation of model outputs.

(b) Evaluation of human expert outputs.

(c) Potential reasons for model errors.

(d) Inclusion of toxic content on model outputs.

Figure 7.7: Error analysis of GPT-4 outputs: evaluation of model outputs, human expert annotations, potential reasons for inaccurate model outputs, and extent of toxicity on detoxified outputs.

The current LLMs still struggle with rephrasing toxic text in a non-toxic way in text detoxification, specially in low-resource language such as Amharic. Our experiments demonstrated that fine-tuning LLMs with sufficient data in low-resource languages, such as Amharic can help to leverage the challenges of text detoxification task.

This study highlights the need for advanced methods to improve detoxification processes which can add a contribution in ensuring safe environment for online users.

Conducting additional experiments and exploring various LLMs such as LLaMA3 and Mistral may present more insights. Besides, extending the task to other low-resource language can be a future work for researchers in the area.

8
Conclusion and Future Directions

Contents

8.1 Summary

In this dissertation, we presented several studies that explored hate speech on four different levels: detecting and classifying hate speech categories, identifying communities targeted by hate speech, rating the intensities of offensiveness and hatefulness within tweets, and paraphrasing toxic text into a more neutral, non-toxic form. In the previous seven chapters, each chapter described the unique aspects of the topic under study.

In Chapter 1, we surveyed social NLP research studies, such as hate speech and sentiment analysis, conducted so far in low-resource languages, with a special emphasis on Amharic, a Semitic language widely spoken in Ethiopia. Before we conducted extensive investigations regarding hate speech, we explored sentiment analysis as a preliminary study to understand the nature of social media content, which helped us to motivate the necessity of the entire dissertation. Our preliminary survey study indicated that over 80% of the research endeavors conducted on sentiment analysis and hate speech in Ethiopian languages, including Amharic, Oromo, and Tigrinya, did not release any resources, which hampers the progress of social NLP tasks in Ethiopian languages. As a preliminary task, we also collected sentiment analysis datasets for Amharic, Oromo, and Tigrinya languages, and built several classification models. The XLMR-large demonstrated the best results for the Amharic sentiment analysis task. The content of datasets collected from social media platforms within Ethiopian languages, such as Amharic, Oromo, and Tigrinya, is heavily skewed towards negative sentiment.

In Chapter 2, we presented three broad issues: literature review, dataset construction approaches, and machine learning models that had been utilized throughout the entire study. In the literature review section, we extensively assessed the status of hate speech studies in Amharic, with respect to the social, cultural and political landscape of Ethiopia and also explored the state of the art approaches on the topic under study. We discussed data collection, representative sample selection, annotation and data quality evaluation strategies that had been utilized in the entire dissertation. Finally, the chapter provided an overview of machine learning approaches, including classical methods, deep learning, and transformer models, employed throughout the dissertation.

Chapter 3 introduced crowdsourcing as a data annotation approach, and showed the appropriateness of crowdsourcing for annotating hate speech datasets in both low and high-resource language settings. We employed the *Yandex Toloka* crowdsourcing platform to annotate Amharic hate speech (for the low-resource language) and French (for the high-resource language), presented in two broad consecutive sections. This chapter showed the appropriateness of crowdsourcing for annotating hate speech datasets in both low and high-resource language settings.

Section 3.2 presented a crowdsourced dataset of 5.4k annotated for Amharic hate speech samples, categorized into hate, offensive, normal, and unsure classes. We developed several classification models, with Am-RoBERTa, a transformer-based contextual embedding model, achieving the best performance, reaching an F1score of 50%.

Similarly, Section 3.3 introduced over 5k crowdsourced hate speech datasets in French. We fine-tuned multilingual BERT-base models such as HateXplain and CamemBERT. We achieved an F1 score of 86%, exceeding the baseline models including HateXplain.

The chapter concluded that annotating datasets in a crowdsourcing setting faces challenges such as quality control over diverse annotators, biases and subjective interpretations, difficulty in managing malicious annotators, and a lack of formal training for annotators in both high-resource and low-resource languages. Controlling malicious annotators and finding the right personnel for the task is more difficult in low-resource languages.

In Chapter 4, we presented hate speech datasets consisting of over 15.1k tweets, annotated into categories of hate, offensive, normal, and unsure labels in a controlled lab-based approach. This results in a relatively higher quality dataset as compared to those presented in Chapter 3, achieving a Cohen's Kappa score of 0.48. We introduced well-defined data selection pipelines and sampling strategies, along with a comprehensive list of hate and offensive lexicon entries. Transformer-based models, such as Am-FLAIR and Am-RoBERTa, outperformed all classical and deep learning models, achieving F1 scores of 72% and 70%, respectively. Our findings indicated that hate speech is highly subjective and complex topic, requiring diverse contextual background information about the original text posted on social media and the intentions of the author at that particular moment. The findings presented in Chapter 3 and Chapter 4 addressed the first research question: *What are the main challenges in crowdsourcing and in-house hate speech annotation approaches?*.

In Chapter 5, we introduced extensive benchmark datasets consisting of 8.3k tweets annotated for three tasks: hate speech category annotation, target community identification, and intensity rating for hatefulness and offensiveness. We annotated each tweet with five native speakers and achieved a Fleiss' Kappa score of 0.49. We determined the gold labels based on majority votes. We built several models for each of the three tasks, where Afro-XL-large demonstrated superior performance in category classification and target identification, achieving F1 scores of 75.30% and 70.59%, respectively, and a

Pearson correlation coefficient of 80.22% for the intensity prediction task. The findings highlighted that hate speech in Ethiopia disproportionately targets communities primarily based on their political and ethnic identities, indicating the turbulent sociopolitical situations in Ethiopia. These two group identities, among others, often co-occur in the dataset, suggesting that politics and ethnicity are significantly intertwined in the nation's sociopolitical landscape, which addresses our second research question: *To what extent do hate speech disproportionately target specific vulnerable communities?*. Additionally, the third research question: *How can hate and offensive speech be understood: as distinct categories or as values on a spectrum of varying intensities?* is addressed with our findings in the intensity rating task, which underlined that hate speech is not a simple discrete concept; instead, it manifests itself in a continuum of ratings.

Chapter 6 presented a multimodal Amharic hate speech dataset, consisting of over 2k memes collected from prominent social media platforms in Ethiopia, such as Facebook, Telegram and X and built several classification models utilizing the dataset. BiLSTM exceptionally outperformed all the models, including transformers, in the multimodal setup, achieving F1 score of 75%. This might have been associated with BiLSTM's capability to learn complex representations from smaller datasets. The findings highlighted that most of the models in the *Multimodal* setup outperformed the *Text-Only* and *Image-Only* models, indicating the importance of exploring hate speech in multimodal approach. *Image-Only* models particularly attained worst results, as compared to the multimodal approach. These findings addressed the fourth research question: *To what extent do multimodality enhance the detection of hate speech compared to unimodal approaches?*

In Chapter 7, we addressed the text detoxification task, which utilized the new *Paradetox* parallel dataset, consisting of the original input text and the rephrased detoxified counterpart. We created the dataset through a meticulous selection of offensive instances from our previous datasets presented in Chapter 3 and Chapter 4, producing the detoxified versions for each offensive tweet with trained human experts. We conducted three distinct tasks: text classification, toxic term detection and detoxification, which involves rewriting the original messages into a more neutral manner. Am-RoBERTa achieved the best F1 score of 70.01% on the classification task, while GPT-4 outperformed SHAP on toxic term identification task, with a score of 76.33%. Despite GPT-4 demonstrating better performance in the toxic term detection task, the generated detoxified versions in the text rewriting task significantly suffered from hallucinations, mainly due to changes in the meaning of the original messages as well as producing unintelligible outputs. This finding suggests that large language models (LLMs) like GPT-4 encounter significant difficulties in generating accurate and coherent text when working with low-resource languages such as Amharic. Thus the fifth research question, which was posed as: *What challenges do large language models (LLMs) face in Amharic text detoxification task?* is properly addressed by the findings presented in Chapter 7.

This dissertation extensively investigated the pressing topic of hate speech, particularly focusing on a low-resource language: Amharic. The dissertation addressed hate speech by implementing five major components: hate speech detection, target identification, intensity prediction, multimodal hate speech detection, and text detoxification. We studied hate speech in a relatively broader aspects of analysis as compared to other previous studies conducted within the context of low-resource African languages including Amharic.

8.2 Main Contributions

This dissertation comprised of the following main contributions:

i) Datasets

In this dissertation, we introduced seven different datasets annotated for various tasks and applications, each created in unique designs utilizing specific annotation guidelines and strategies. These datasets are publicly available to enhance the progress of hate speech studies, specially in low-resource languages, which include:

1. **AfriSenti sentiment analysis dataset**, part of the AfriSenti sentiment analysis datasets which include Amharic, Oromo and Tigrinya tweets are contributions emanated from this dissertation as a preliminary social media analysis resource.

2. **Crowdsourced Amharic hate speech dataset**, which consists of 5.3k tweets annotated by three crowd performers into hate, offensive, normal and unsure labels.

3. **Crowdsourced French hate speech dataset**, which is annotated by three Toloka crowd performers into hate, offensive, normal and unsure labels, including racial and non-racial targets for hateful tweets. The dataset comprises of 5k tweets.

4. **Lab-Controlled Amharic hate speech dataset**, consisting of over 15.1k tweets annotated by two independent native speakers and the final label curated by a more experienced expert to achieve better quality datasets, labeled as hate, offensive, normal, and unsure.

5. **Multitask Amharic hate speech dataset**, this dataset incorporates 8.3k tweets annotated by 5 Amharic native speakers for three independent tasks, namely, *category classification* (labeled into hate, offensive, normal, and indeterminate), *hatred target detection* (labeled into ethnicity, religion, disability, gender, politics and etc...), and *intensity rating* which represent hatefulness and offensiveness intensities of tweets in a Likert rating scale ranging from 1 to 5.

6. **Multimodal Amharic Meme dataset**: this dataset consists of 2k Amharic memes annotated by three native speakers into hate and non-hate, which presented both annotated meme images and extracted texts from each meme.

7. **ParaDetox Amharic Dataset**: this dataset introduced over 3k offensive tweets selected from previous datasets and annotated into detoxifiable or non-detoxifiable labels. Paraphrased versions of detoxifiable tweets are produced by human experts to make the tweets more neutral, which consists of over 1.5k parallel textual inputs, creating the *Paradetox* dataset.

ii) Baseline Models

This dissertation presented several models fine-tuned on datasets specifically created for the purpose of this dissertation, which include:

- **Classical machine leaning methods**, which includes logistic regression, support vector machines, and Naïve Bayes.

- **Deep learning techniques** such as LSTM, BiLSTM, CNN, and RNN models.
- **Transformer models**, constitute models that has been pretrained on low-resource languages including Ethiopian languages, which include Amroberta, XLMR-large, Afro-XLMR-large, Afriberta_large, Afriberta_base, Afriberta_small, AfroLM-Large (w/ AL), Rasyosef_Bert-medium-amharic and Rasyosef_Bert-small-amharic.

iii) Annotation Guidelines and Sentiment, Hate and Offensive speech Lexicons

We have designed several annotation guidelines, which have been utilized for the various hate speech studies conducted as part of this dissertation, including the preliminary sentiment analysis task. We created list of several lexicon entries, which include sentiment lexicon (positive and negative), hate speech lexicon (hate, offensive).

8.3 Future Directions

Ethiopia is a multicultural and multilingual country, which serves as a home of over 82 ethnic groups, several religions, many languages, diverse cultures and evolving sociopolitical dynamics. Extending hate speech studies to create multilingual datasets and develop cross-lingual and multilingual models, which can address hate speech in a more comprehensive context could be one of the future research directions.

Incorporating social and cultural information, such as community-specific norms, dialects, slang, and idiomatic expressions, is crucial when creating hate speech datasets and developing models. This inclusion could serve as a valuable research direction for improving hate speech detection. It enables models to capture subtle variations associated with cultural, social, and political diversity.

Another future research direction that several of our findings indicated is model explainability, which increases the transparency and trust of models while deciding hatefulness or offensiveness.

The findings of our studies have shown that hateful social media content are rapidly spreading across platforms, reaching numerous users in a short period of time. Our dataset would be essential components in building models for social media moderation and digital peacebuilding initatives.

The of bulk our work mainly focused on textual hate speech, while hate speech is manifested in multiple modalities. Extending our multimodal dataset to cover more instances in several languages and including additional modalities like audio and video, would pave the way to a more comprehensive approach to combating hate speech online.

Combating online hate speech requires proactive mitigation strategies, while several of our works mainly approached hate speech within the reactive methods. Counter-speech, which promotes positive and corrective responses, utilizing AI-driven generation, context sensitivity, and real-time integration with detection systems, seems to be a promising direction to mitigate online hate speech.

Appendices

A
Additional Material

This guideline is mainly used for Amharic hate speech data annotation tasks that utilized the **Yandex Toloka** crowdsourcing platform in Chapter 3 and **WebAnno** in Chapter 4.

Amharic Version

Bahir Dar University, Bahir Dar, Ethiopia
and
University of Hamburg, Hamburg, Germany

መመሪያ!

የዚህ መጠይቅ ዋና አላማ በማህበራዊ ሚዲያ ላይ የሚለቀቁ የጥላቻ (hate) እና አፀያፊ (offensive) ንግግሮችን ለመለየት የሚያስችል በቴክኖሎጅ ምርምር የሚታገዝ መፍትሄ ለመስጠት ነው። ስለሆነም የአማርኛ ቋንቋን በመጠቀም የሚተላለፉ ጥላቻ እና አፀያፊ ጽሁፎችን የያዙትን በትክክል ለይቶ ለመክፈል ያለመ ስራ ነው። በዚህ መጠይቅ ላይ የቀረበው የጽሁፍ መረጃ ከቲዉተር (Twitter) የተወሰደ ስለሆነ የጥላቻ (hate) እና አፀያፊ (offensive) ጽሁፎችን ይዟል። ስለሆነም እነዚህን ንግግሮች ማየት ካልፈልጉና በዚህ ስራ ለመሳተፈ ካልፈለጉ ጽሁፎቹን ሳይመለከቱ ውጣ የሚለውን መተግበሪያ ተጭነው ይውጡ።

- በመልእክት ላይ ያተኩሩ (በራስዎ አስተያየት ላይ ላለማተኮር ይሞክሩ)።

- *ሥራውን ለመጨረስ አትቸኩሉ እና ሁል ጊዜ ጥርጣሬ በሚፈጠርበት ጊዜ ተመራማሪዎቹን ያግኙ።*

በመጠይቁ ለመሳተፍ ከወሰኑ መመሪያውን በአጽናኦት ያንብቡና የደስራው ይግቡ።

ስለትብብርዎ በቅድሚያ እናመሰግናለን!!!

የቀረቡት ጽሁፎች ከሚከተሉት አራት (4) ክፍፍሎች ውስጥ አንዱን በመመምረጥ መልስዎን ይስጡ።

I. **የጥላቻ ንግግር (Hate Speech)** ማላት በቀረቡት ጽሁፎቻ ውስጥ አንድን ግለሰብ ወይም ተቋም የቡድን ማንነቱን የሚያንቋሽሽ (በብሄር፡ በጎሳ፡ በሃይማኖት፡ በባህል፡ በልምድ፡ በአካል ጉዳተኝነት፡ በጾታ፡ በፖለቲካ አመለካከት ወ.ዘ.ተ.) ንግግር ማለት ነው። ለምሳሌ፤ መድሎ፤ ማስፈራራት፤ ለጥፋት/ጥቃት ማነሳሳት፤ ማንኳሰስ፤ ንቀት፤ አሽሙር፤ ስድብ፤ ዘለፋ፤ የመሳሰሉ ነገሮች ማንነትነ መሰረት አድርግው ሲቀርቡ ነው። አንድ ጽሁፈ ጥላቻና አፀያፊ ቅልቅል ንግግሮቸን ከያዘ፡ ጥላቻ ንግግሮች ውስጥ ይካተታል።

II. **አፀያፊ (Offensive)** ንግግር ማለት አንዲን ግለሰብ ወይም ተቋም እንዲናደድ፥ እንድበሳጭ፥ እንዲቀየም ወይም ሞራሉ እንድጎዳ የሚያደርግ ትገቢ ያልሆነ ንግግር ማለትም ስድብ፡ ስም ማጥፋት፥ የአሽሙር፡ የንቀት፡ ወ.ዘ.ተ. ንግግር የያዘ ሲሆን ነው።

III. **መደበኛ (Normal)** ንግግር ማለት አንድ ጽሁፍ ምንም አይነት የጥላቻም ሆነ አጻያፊ ንግግር ሳይዝ ሲቀርና

በአንባቢው ሰው አስተያየት ምንም ምጥፎ ነገር ያልያዘ ንግግር ማለት ነው።

IV. **ለመወሰን የሚያስቸገር (UnSure)** ንግግር ማለት አንድ ጽሁፍ መደበኛ ወይም ጥላቻና አፀያፊ ይሁን አይሁን ለአንባቢው ሰው መለየት ሲያቅው ለመወሰን የሚያስቸግር ይባላል።

የአእምሮ ጤና አደጋ እና ደህንነት:

ጎጂ ይዘትን መግለጽ ሥነ ልቦናዊ ጭንቀት ሊሆን ይችላል። በሂደቱ ወቅት ጭንቀት ወይም ምቾት የሚሰማው ማንኛውም ገላጭ እረፍት እንዲያደርግ ወይም ስራውን እንዲያቆም እና እርዳታ እንዲፈልግ እንመክራለን። ቅድመ ጣልቃ ገብነትን ለመቋቋም ከሁሉ የተሻለው መንገድ ነው

English Version

Bahir Dar University, Bahir Dar, Ethiopia

and

University of Hamburg, Hamburg, Germany

Annotation Guideline!

The main objective of this research is to identify hate and offensive speech on social media with the help of technology. It will help to tackle hateful or offensive content that are written in Amharic languages and posted on social media. The data is crawled from Twitter using Twitter API and don’t reflect the researchers’ opinions.

If you do not want to see such content and do not want to participate in the work, you can click the logout button and leave without looking at the content.

- Focus on the message conveyed in the tweets and try not to focus on your own opinion on the topic.
- Do not rush to finish the task and always reach out to the researchers with questions when in doubt.

If you want toparticipate in the task, please read the guideline carefully.

Thank you for your participation in advance

Please read each tweet/comment carefully and choose its approperate label.

Category:

- **Offensive speech** is any form of bad language expressions including rude, impolite, insulting, or belittling utterance intended to offend or harm an individual.
- **Hate speech** is language content that expresses hatred towards a particular group or individual based on their group identities such as race, ethnicity, religion, gender, disability, political affiliation, or other characteristics. It also includes threats of violence associating group identities.

- **Normal** is any form of expression that does not contain any bad language belonging to any of the above classifications.
- **Unsure** any tweet that is difficult to give a specific label, or if you are not quite sure about the content's label, label it as unsure.

Mental health risk and well-being:

Annotating harmful content can be psychologically distressing. We advise any annotator who feels anxious or uncomfortable during the process to take a break or stop the task and seek help. Early intervention is the best way to cope.

B
Additional Material

This guideline is designed to guide and employ the annotation task , which comprises of three sub-tasks: category detection, target community identification, and hatefulness and offensiveness intensity rating of tweets. The datasets utilized in Chapter 5 are created based on this guideline.

Annotation Guideline for Amharic Hate Speech (Category, Target and Intensity using Rating Scales)

Introduction

This guideline presents the concepts and rules on how to annotate and rate potentially harmful tweets that can cause emotional distress to individuals, incite violence, or discriminate against, and exclude social groups.

Annotators are expected to be **objective (as much as possible).** We welcome your feedback on how we can update the guidelines based on your feedback.
Always use the guidelines and you should be **objective** and **consistent** in your annotation.

- **Focus on the message conveyed in the tweets** and try not to focus on your own opinion on the topic.
- **Do not rush to finish the task** and always reach out to the researchers with questions when in doubt.

Mental health risk and well-being:

Annotating harmful content can be psychologically distressing. We advise any annotator who feels anxious or uncomfortable during the process to take a break or stop the task and seek help. Early intervention is the best way to cope.

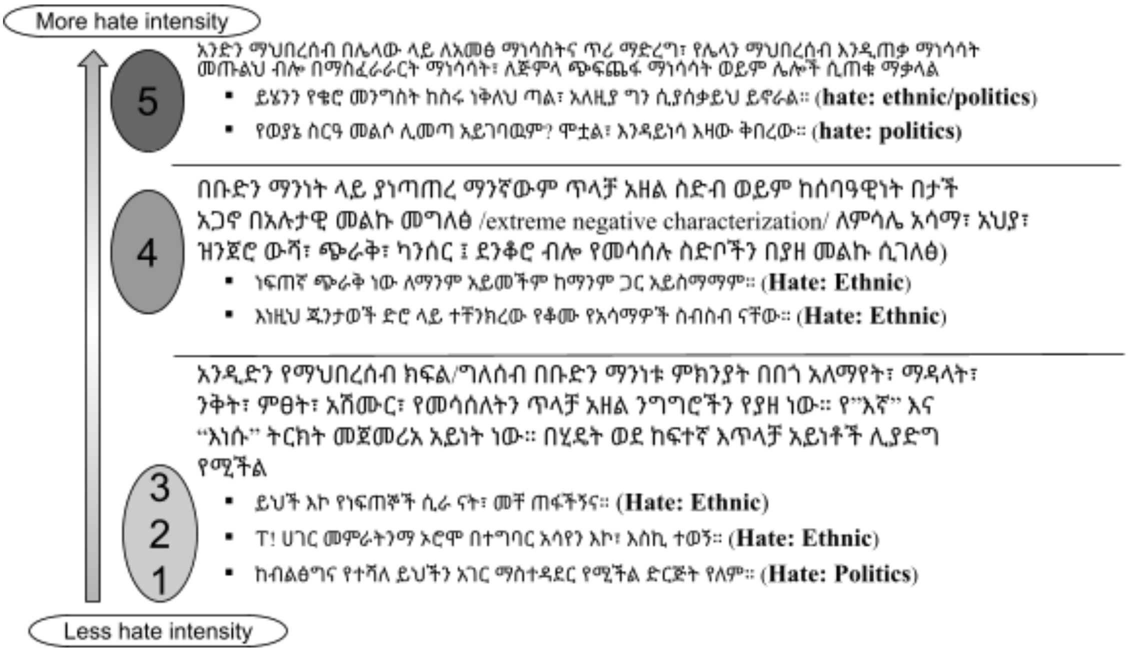
Category:

- **Offensive speech** is any form of bad language expressions including r*ude, impolite, insulting, or belittling utterance intended to offend or harm an individual.*
- **Hate speech** is language content that expresses **hatred towards a particular *group or individual*** based on their group identities such as **race**, **ethnicity**, **religion**, **gender, disability, political affiliation, or other characteristics.** It also includes **threats of violence** associating group identities.
- **Normal** is any form of expression that does not contain any bad language belonging to any of the above classifications.
- **Indeterminate** is any tweet that is not **readable** or is **completely** written in another language other than your language of annotation.

For each tweet, you are required to identify the category and for **hate** and **offensive** tweets, you are requested to indicate the severity of hate or offensive tweets by rating from **1-5** in **a Likert scale** as shown in the example.

Severity level:

1. Hate speech Severity levels:



አንድን ማህበረሰብ በሌላው ላይ ለአመፅ ማነሳስትና ጥሪ ማድረግ፣ የሌላን ማህበረሰብ እንዲጠቃ ማነሳሳት መጡልህ ብሎ በማስፈራራርት ማነሳሳት፣ ለጅምላ ጭፍጨፋ ማነሳሳት ወይም ሌሎች ሲጠቁ ማቃላል

- ይሄንን የቁሮ መንግስት ከስሩ ነቅለህ ጣል፣ አለዚያ ግን ሲያሰቃይህ ይኖራል። (**hate: ethnic/politics**)
- የወያኔ ስርዓ መልሶ ሊመጣ አይገባዉም? ሞቷል፣ እንዳይነሳ አዛው ቅበረው። (**hate: politics**)

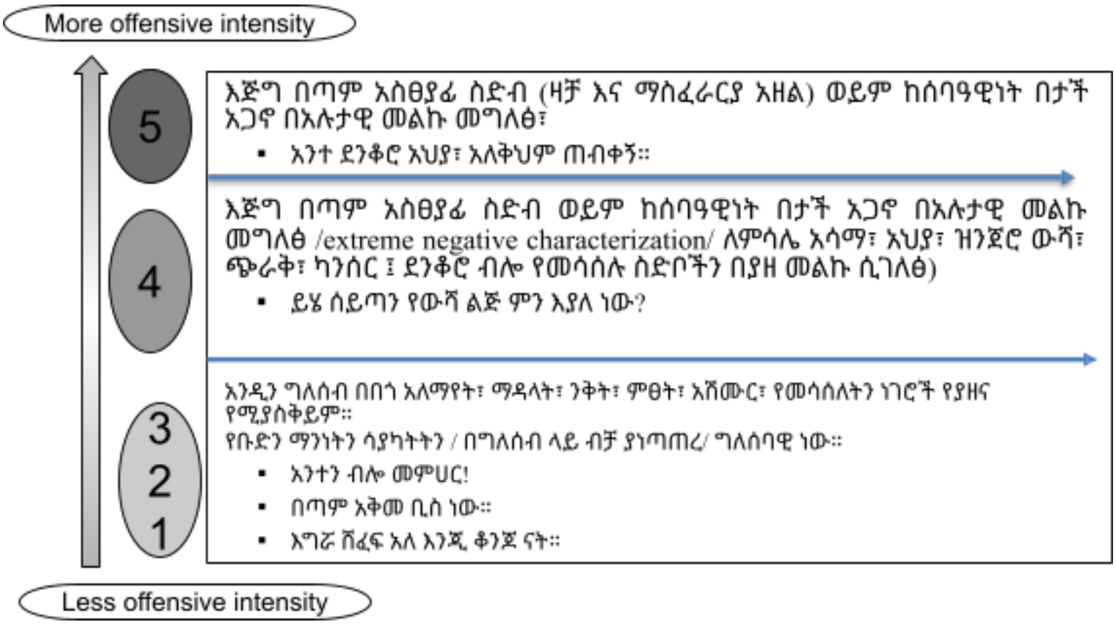
በቡድን ማንነት ላይ ያነጣጠረ ማንኛውም ጥላቻ አዘል ስድብ ወይም ከሰባዓዊነት በታች አጋኖ በአሉታዊ መልኩ መግለፅ /extreme negative characterization/ ለምሳሌ አሳማ፣ አህያ፣ ዝንጀሮ ውሻ፣ ጭራቅ፣ ካንሰር ፤ ደንቆሮ ብሎ የመሳሰሉ ስድቦችን በያዘ መልኩ ሲገለፅ)

- ነፍጠኛ ጭራቅ ነው ለማንም አይመችም ከማንም ጋር አይስማማም። (**Hate: Ethnic**)
- እነዚህ ጁንታወች ድሮ ላይ ተቸንክረው የቆሙ የአሳማዎች ስብስብ ናቸው። (**Hate: Ethnic**)

እንዲድን የማህበረሰብ ክፍል/ግለሰብ በቡድን ማንነቱ ምክንያት በበጎ አለማየት፣ ማዳላት፣ ንቅት፣ ምፀት፣ አሽሙር፣ የመሳሰለትን ጥላቻ አዘል ንግግሮችን የያዘ ነው። የ"እኛ" እና "እነሱ" ትርክት መጀመሪአ አይነት ነው። በሂደት ወደ ከፍተኛ እጥላቻ አይነቶች ሊያድግ የሚችል

- ይህች እኮ የነፍጠኞች ሲራ ናት፣ መቸ ጠፋችኝና። (**Hate: Ethnic**)
- T! ሀገር መምራትንማ ኦሮሞ በተግባር አሳየን እኮ፣ እስኪ ተወኝ። (**Hate: Ethnic**)
- ከብልፅግና የተሻለ ይህችን አገር ማስተዳደር የሚችል ድርጅት የለም። (**Hate: Politics**)

2. Offensive speech Severity levels:



እጅግ በጣም አስፀያፊ ስድብ (ዛቻ እና ማስፈራርያ አዘል) ወይም ከሰባዓዊነት በታች አጋኖ በአሉታዊ መልኩ መግለፅ፣

- አንተ ደንቆሮ አህያ፣ አለቅህም ጠብቀኝ።

እጅግ በጣም አስፀያፊ ስድብ ወይም ከሰባዓዊነት በታች አጋኖ በአሉታዊ መልኩ መግለፅ /extreme negative characterization/ ለምሳሌ አሳማ፣ አህያ፣ ዝንጀሮ ውሻ፣ ጭራቅ፣ ካንሰር ፤ ደንቆሮ ብሎ የመሳሰሉ ስድቦችን በያዘ መልኩ ሲገለፅ)

- ይሄ ሰይጣን የውሻ ልጅ ምን እያለ ነው?

እንዲን ግለሰብ በበጎ አለማየት፣ ማዳላት፣ ንቅት፣ ምፀት፣ አሽሙር፣ የመሳሰለትን ነገሮች የያዘና የሚያስቅይም።
የቡድን ማንነትን ሳያካትትን / በግለሰብ ላይ ብቻ ያነጣጠረ/ ግለሰባዊ ነው።

- አንተን ብሎ መምሀር!
- በጣም አቅመ ቢስ ነው።
- እግሯ ሽፈፍ አለ እንጂ ቆንጆ ናት።

C
Additional Material

This annotation guideline is designed to guide the data annotation task, which is utilized in Chapter 7 to create the parallel text detoxification datasets.

የስራ መመሪያ

ዓላማ፡ አስፀያፊ መልዕክትን ማፅዳትና ዋና መለእክቱን ሳይለቅ እንደገና በመፃፍ ማስተካከል።

የዚህ መመሪያ ዋና አላማ ቀጥሎ የቀረቡት ከማህበራዊ ሚዲያ ላይ የተሰበሰቡ አስፀያፊ ፅሁፎችን ትርጉማቸውን ሳይለቁ እንደጋና በመፃፍ አስፀያፊ እንዳይሆኑ ማድረግ ነው። ቀጥሎ የቀረቡትን ምሳሌዎች በማንበብ ስራውን ለመረዳት ይሞክሩ።

ማሳሰቢያ፡ የጥላቻ ንግግር ከሆነ እንዲሁም በሌላ ምክንያት አስፀያፊነቱን ለማስወገድ አስቸጋሪ ከሆነብዎት ከተመለከቱት ምርጫዎች ውስጥ ወደ አንዱ ይመድቡት።

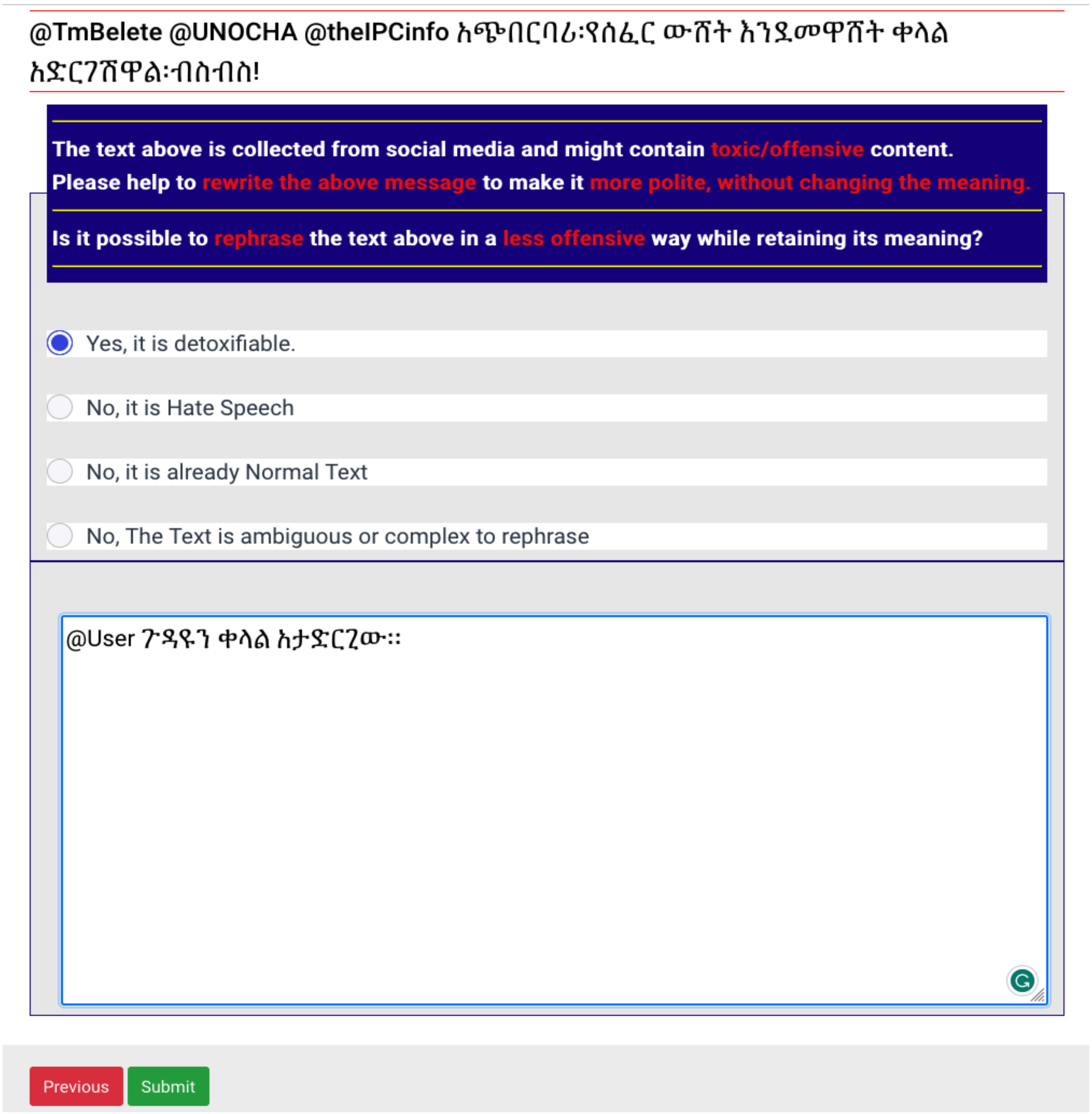
ትርጓሜ፡

- ➢ አስፀያፊ (Offensive) ንግግር ማለት አንድን ግለሰብ በቡድን ማንነቱ ሳይሆን የግል ስብዕናው መሰረት በማድረግ እንዲበሳጭ፣ እንዲናደድ፣ እንዲቀየም ወይም ሞራሉ እንዲጎዳ የሚያደርግ ተገቢ ያልሆነ ንግግር ማለት ነው።

- ➢ የጥላቻ ንግግር (Hate Speech) ማለት አንድን ቡድን ወይም ግለሰብ የቡድን ማንነቱን ማለትም ብሄር፣ ጎሳ፣ ሃይማኖት፣ ቋንቋ፣ ባህል፣ ልምድ፣ ፆታ፣ እንዲሁም አካል ጉዳተኝነትን መሰረት በማድረግ የሚገለፅ ጥላቻ ያዘለ ንግግር ነው። እንዲሁም ሰዎችን በቀጥታም ሆነ በተዘዋዋሪ ለአመፅ፣ ለወንጀል፣ ለጥፋትና ለበቀል ማነሳሳትን ያካትታል።

- ➢ መደበኛ (Normal) ንግግር ማለት አንድ ፅሁፍ ምንም አይነት የጥላቻም ሆነ አስፀያፊ ንግግር ሳይኖረው ሲቀር እና በአንባቢው ሰው አስተያየት ምንም አይነት መጥፎ ስሜት የማይፈጥር ንግግር ሲሆን ነው፤፤

ቀጥሎ የተመለከትው ምስል ለመረጃ ሰብሳቢዎች የሚቀርበው የመተግበሪያው በይነገፅ ሲሆን ዋና መልዕክቱን፤ ትዕዛዛቱን፤ አማራጮችን እንድሁም አስፀያፊ ይዘት ያለው መልዕክት አስፀያፊነቱን ማንፃት የሚቻል ከሆነ እንደገና የመፃፊያ ቦታ አካቶ ይዟል።



@TmBelete @UNOCHA @theIPCinfo አጭበርባሪ፡የሰፈር ውሸት እንደመዋሸት ቀላል አድርገሽዋል፡ብስብስ!

The text above is collected from social media and might contain toxic/offensive content. Please help to rewrite the above message to make it more polite, without changing the meaning.

Is it possible to rephrase the text above in a less offensive way while retaining its meaning?

● Yes, it is detoxifiable.

○ No, it is Hate Speech

○ No, it is already Normal Text

○ No, The Text is ambiguous or complex to rephrase

@User ጉዳዩን ቀላል አታድርገው::

<u>ጥሩ ምሳሌዎች፡</u>

Input: @User አጭበርባሪ፡የሰፈር ውሸት እንደመዋሸት ቀላል አድርገሽዋል፡ብስብስ!
Output: @User ቀላል አታድርጊው። **OR** ቀላል አይደም።

Input: ስንት አይነት ደደብ አለ!!
Output: ስንት አይነት ሰው አለ!! **OR** እንደዚህም አይነት ሰው አለ!!

Input: @User ግማታም ዶክቶር አብይ የተባረከ ምርጥ ጀግና የ ኢትዮጵያ ልጅ ነው በ አንተ ቱልቱላ ስሙ አይነሳም እሺ??? አንተ ነህ ዕድሜህ ስደት ላይ የምትጨርሰው
Output: @User ዶክቶር አብይ የተባረከ ምርጥ ጀግና የ ኢትዮጵያ ልጅ ነው በከንቱ ስሙ አይነሳም እሺ??? አንተም ዕድሜህን በስደት ላይ አትጨርሰው
Input: @User ተጨማሪ ስልጣን ፈልጎ ማሽቃበጡ ነው። ብሽቅ ካድሬ!
Output: @User ተጨማሪ ስልጣን ፈልጎ መደገፉ ነው። የመንግስት ደጋፊ!

The first kind of intellectual and artistic personality belongs to the hedgehogs, the second to the foxes …

— Sir Isaiah Berlin (2013)

References

Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. 2022. Afaan Oromo Hate Speech Detection and Classification on Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 6612–6619. Marseille, France: European Language Resources Association. (Cited on pages 26 sq.).

Zeleke Abebaw, Andreas Rauber, and Solomon Atnafu. 2022a. Design and Implementation of a Multichannel Convolutional Neural Network for Hate Speech Detection in Social Networks. *Revue d'Intelligence Artificielle* 36 (2). (Cited on pages 74, 84, 99, 113).

———. 2022b. Multi-channel convolutional neural network for hate speech detection in social media. In *Proceedings of the 9th EAI International Conference on the Advances of Science and Technology, ICAST 2021,* 603–618. Bahir Dar, Ethiopia: Springer. (Cited on pages 26 sq., 39, 47, 57, 74, 84, 99, 113).

Bekalu Tadele Abeje, Ayodeji Olalekan Salau, Habtamu Abate Ebabu, and Aleka Melese Ayalew. 2022. Comparative Analysis of Deep Learning Models for Aspect Level Amharic News Sentiment Analysis. In *2022 International Conference on Decision Aid Sciences and Applications (DASA),* 1628–1633. Chiangrai, Thailand: IEEE. (Cited on page 9).

Halefom Hailu Abraha. 2017. Examining approaches to Internet regulation in Ethiopia. *Information and Communications Technology Law* 26 (3): 293–311. (Cited on page 3).

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A Neural Language Identification Tool for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 1958–1981. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on page 10).

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire M. Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 4488–4508. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on page 12).

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal Question Answering for Memes via Rationale-Based Inferencing. In *Findings of the Association for Computational Linguistics ACL 2024,* 5042–5078. Bangkok, Thailand: Association for Computational Linguistics. (Cited on pages 100 sq.).

Shawly Ahsan, Eftekhar Hossain, Omar Sharif, Avishek Das, Mohammed Moshiul Hoque, and M. Dewan. 2024. A Multimodal Framework to Detect Target Aware Aggression in Memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers),* 2487–2500. St. Julian's, Malta: Association for Computational Linguistics. (Cited on pages 100 sqq.).

Alham Fikri Aji, Jessica Zosa Forde, Alyssa Marie Loo, Lintang Sutawika, Skyler Wang, Genta Indra Winata, Zheng-Xin Yong, Ruochen Zhang, A. Seza Doğruöz, Yin Lin Tan, and Jan Christian Blaise Cruz. 2023. Current Status of NLP in South East Asia with Insights from Multilingualism and Language Diversity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract,* 8–13. Nusa Dua, Bali: Association for Computational Linguistics. (Cited on page 46).

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics,* 4336–4349. Gyeongju, Republic of Korea: International Committee on Computational Linguistics. (Cited on pages 42, 93, 105).

Girma Neshir Alemneh, Andreas Rauber, and Solomon Atnafu. 2020. Negation handling for Amharic sentiment classification. In *Proceedings of the Fourth Widening Natural Language Processing Workshop,* 4–6. Seattle, WA, USA: Association for Computational Linguistics. (Cited on page 9).

Raghad Alshaalan and Hend Al-Khalifa. 2020. Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop,* 12–23. Barcelona, Spain (Online): Association for Computational Linguistics. (Cited on pages 39 sq.).

Hayastan Avetisyan and David Broneske. 2023. Large Language Models and Low-Resource Languages: An Examination of Armenian NLP. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings),* 199–210. Nusa Dua, Bali: Association for Computational Linguistics. (Cited on page 46).

Yohannes Eneyew Ayalew. 2020. Defining 'Hate Speech'under the Hate Speech Suppression Proclamation in Ethiopia: A Sisyphean exercise? *Ethiopian Human Rights Law Series* 12. (Cited on pages 27 sq.).

Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashaf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korenčić, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, Xintong Wang, Matti Wiegmann, Seid Muhie Yimam, and Eva Zangerle. 2024. Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction,* 231–259. Cham, Switzerland: Springer Nature. (Cited on pages 17 sqq., 21, 110 sq., 113, 117).

Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Tegegne Asfaw, and Chris Biemann. 2022. Challenges of Amharic Hate Speech Data Annotation Using Yandex Toloka Crowdsourcing Platform. In *Proceedings of the The Sixth Widening NLP Workshop (WiNLP).* Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 61, 76 sq.).

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5js in Ethiopia: Amharic Hate Speech Data Annotation Using Toloka Crowdsourcing Platform. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA),* 114–120. Bahir Dar, Ethiopia: IEEE. (Cited on pages 3, 17 sq., 21, 26 sq., 35 sq., 39, 46, 48, 50, 62, 84, 99, 113, 118 sq.).

Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023. Multilingual Racial Hate Speech Detection Using Transfer Learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing,* 41–48. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria. (Cited on pages 17, 19, 21, 46, 59).

Abinew Ali Ayele, Esubalew Alemneh Jalew, Adem Chanie Ali, Seid Muhie Yimam, and Chris Biemann. 2024. Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024,* 167–178. Torino, Italia: ELRA / ICCL. (Cited on pages 17, 19, 21, 84, 105, 113).

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic Hate Speech Data Collection and Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing,* 49–59. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria. (Cited on pages 3, 17, 19, 21, 31, 70, 84, 99, 105, 113 sq., 117 sqq.).

Babak Bahador. 2023. Monitoring hate speech and the limits of current definition. In *Challenges and perspectives of hate speech research,* 12:291–298. Digital Communication Research. Berlin, Germany. (Cited on pages 84, 90 sq.).

Weldemariam Bahre. 2022. Hate Speech Detection from Facebook Social Media Posts and Comments in Tigrigna language. PhD diss., St. Mary's University. (Cited on pages 26 sq.).

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin. 2022. Validity, Agreement, Consensuality and Annotated Data Quality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2940–2948. Marseille, France: European Language Resources Association. (Cited on page 32).

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022.* Virtual Event. (Cited on page 105).

Arup Baruah, Ferdous Barbhuiya, and Kuntal Dey. 2019. ABARUAH at SemEval-2019 Task 5 : Bi-directional LSTM for Hate Speech Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation,* 371–376. Minneapolis, MN, USA: Association for Computational Linguistics. (Cited on pages 36, 41).

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish Hate Speech Dataset and Detection System. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 4177–4185. Marseille, France: European Language Resources Association. (Cited on pages 24, 26, 84).

Divya Bhadauria, Alejandro Sierra Múnera, and Ralf Krestel. 2024. The Effects of Data Quality on Named Entity Recognition. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024),* 79–88. San Ġiljan, Malta: Association for Computational Linguistics. (Cited on page 69).

Aruna Bhat, Vaibhav Vashisht, Vaibhav Raj Sahni, and Sumit Meena. 2023. Hate Speech Detection using Multimodal Meme Analysis. In *Proceedings of the 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC),* 1137–1142. Salem, India: IEEE. (Cited on page 101).

Christopher Bishop. 2006. Pattern Recognition and Machine Learning. Information Science and Statistics. Springer. (Cited on pages 34 sq.).

Victoria Bobicev and Marina Sokolova. 2018. Thumbs Up and Down: Sentiment Analysis of Medical Online Forums. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task,* 22–26. Brussels, Belgium: Association for Computational Linguistics. (Cited on page 8).

João Bran and Adeline Hulin. 2023. Social Media 4 Peace: local lessons for global practices. Countering hate speech. the United Nations Educational, Scientific / Cultural Organization (UNESCO). (Cited on page 3).

Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk,* 1–12. Los Angeles, CA, USA: Association for Computational Linguistics. (Cited on page 47).

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 321–332. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on page 101).

Cohan Sujay Carlos and Madhulika Yalamanchi. 2012. Intention Analysis for Sales, Marketing and Customer Service. In *Proceedings of COLING 2012: Demonstration Papers,* 33–40. Mumbai, India: The COLING 2012 Organizing Committee. (Cited on page 8).

Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. 2022. Hate Speech Dynamics Against African descent, Roma and LGBTQI Communities in Portugal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2362–2370. Marseille, France: European Language Resources Association. (Cited on page 60).

Pompeu Casanovas and Andre Oboler. 2018. Behavioural Compliance and casanovas2018behavioural Enforcement in Online Hate Speech. In *TERECOM@ JURIX,* 125–134. Groningen, The Netherlands. (Cited on pages 13 sq.).

Tommaso Caselli and Hylke Van Der Veen. 2023. Benchmarking Offensive and Abusive Language in Dutch Tweets. In *The 7th Workshop on Online Abuse and Harms (WOAH).* Toronto, Canada: Association for Computational Linguistics. (Cited on page 83).

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective Isms? On the Danger of Conflating Hate and Offence in Abusive Language Detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024),* 275–282. Mexico City, Mexico: Association for Computational Linguistics. (Cited on page 58).

Bharathi Raja Chakravarthi. 2020. HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media,* 41–53. Barcelona, Spain (Online): Association for Computational Linguistics. (Cited on pages 35 sq.).

Nikhil Chakravartula. 2019. HATEMINER at SemEval-2019 Task 5: Hate speech detection against Immigrants and Women in Twitter using a Multinomial Naive Bayes Classifier. In *Proceedings of the 13th International Workshop on Semantic Evaluation,* 404–408. Minneapolis, MN, USA: Association for Computational Linguistics. (Cited on pages 35 sq.).

Mohit Chandra, Ashwin Pathak, Eesha Dutta, Paryul Jain, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2020. AbuseAnalyzer: Abuse Detection, Severity and Target Prediction for Gab Posts. In *Proceedings of the 28th International Conference on Computational Linguistics,* 6277–6283. Barcelona, Spain (Online): International Committee on Computational Linguistics. (Cited on pages 35 sq., 40, 84, 91).

Muluken Asegidew Chekol, Mulatu Alemayehu Moges, and Biset Ayalew Nigatu. 2023. Social media hate speech in the walk of Ethiopian political reform: analysis of hate speech prevalence, severity, and natures. *Information, Communication & Society* 26 (1): 218–237. (Cited on pages 3, 28).

Hiroki Chida, Yohei Murakami, and Mondheera Pituxcoosuvarn. 2022. Quality Control for Crowdsourced Bilingual Dictionary in Low-Resource Languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 6590–6596. Marseille, France: European Language Resources Association. (Cited on page 46).

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An Annotated Corpus for Sexism Detection in French Tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 1397–1403. Marseille, France: European Language Resources Association. (Cited on pages 58, 60).

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs Understand Social Knowledge? Evaluating the Sociability of Large Language Models with SocKET Benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 11370–11403. Singapore, Singapore: Association for Computational Linguistics. (Cited on page 1).

Christopher Clarke, Matthew Hall, Gaurav Mittal, Ye Yu, Sandra Sajeev, Jason Mars, and Mei Chen. 2023. Rule By Example: Harnessing Logical Rules for Explainable Hate Speech Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 364–376. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 83 sq.).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20 (1): 37–46. (Cited on page 33).

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 8440–8451. Online: Association for Computational Linguistics. (Cited on pages 42 sq., 93, 118).

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20:273–297. (Cited on page 35).

Marta Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672,* (cited on page 119).

Snehil Dahiya, Shalini Sharma, Dhruv Sahnan, Vasu Goel, Emilie Chouzenoux, Víctor Elvira, Angshul Majumdar, Anil Bandhakavi, and Tanmoy Chakraborty. 2021. Would your tweet invoke hate on the fly? Forecasting hate intensity of reply threads on Twitter. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining,* 2732–2742. Singapore, Singapore: Association for Computing Machinery. (Cited on page 26).

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text Detoxification using Large Pre-trained Neural Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* 7979–7996. Punta Cana, Dominican Republic: Association for Computational Linguistics. (Cited on pages 112, 119).

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2024. Evaluating ChatGPT against Functionality Tests for Hate Speech Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 6370–6380. Torino, Italia: ELRA / ICCL. (Cited on pages 26, 111).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online,* 25–35. Florence, Italy: Association for Computational Linguistics. (Cited on pages 3, 25 sq., 58 sq.).

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media,* 11:512–515. Montréal, Canada: Association for Computational Linguistics. (Cited on pages 3, 13 sq., 25, 61, 74, 83).

Gretel Liz De la Peña Sarracén, Paolo Rosso, and Anastasia Giachanou. 2020. PRHLT-UPV at SemEval-2020 Task 8: Study of Multimodal Techniques for Memes Analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation,* 908–915. Barcelona, Spain (online): International Committee for Computational Linguistics. (Cited on page 100).

Abreham Gebremedin Debele Michael Melese and Woldeyohannis. 2022. Multimodal Amharic hate speech detection using deep learning. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA),* 102–107. IEEE. (Cited on pages 102, 105).

Naol Bakala Defersha and Kula Kekeba Tune. 2021. Detection of hate speech text in Afan Oromo social media using machine learning approach. *Indian Journal of Science and Technology* 14 (31): 2567–78. (Cited on pages 26 sq.).

Mequanent Degu, Abebe Tesfahun, and Haymanot Takele. 2023. Amharic language hate speech detection system from Facebook memes using deep learning system. *Available at SSRN 4389914,* (cited on page 102).

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 4707–4717. Hong Kong, China: Association for Computational Linguistics. (Cited on page 7).

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17),* 86–95. Venice, Italy: ITASEC17. (Cited on pages 47, 54, 75).

Donatella della Porta and Mario Diani. 2015. The Oxford Handbook of Social Movements. Oxford University Press. (Cited on page 2).

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. MultiParaDetox: Extending Text Detoxification with Parallel Data to New Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers),* 124–140. Mexico City, Mexico: Association for Computational Linguistics. (Cited on pages 111 sq.).

Daryna Dementieva, Nikolay Babakov, Amit Ronen, **Abinew Ali Ayele**, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashaf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and Explainable Text Detoxification with Parallel Corpora. In *Proceedings of the 31 International Conference on Computational Linguistics (COLING 2025),* –. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 18 sq., 21, 110 sq., 113, 117).

Daryna Dementieva, Daniil Moskovskiy, David Dale, and Alexander Panchenko. 2023. Exploring Methods for Cross-lingual Text Style Transfer: The Case of Text Detoxification. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers),* 1083–1101. Nusa Dua, Bali: Association for Computational Linguistics. (Cited on page 112).

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Methods for Detoxification of Texts for the Russian Language. *Multimodal Technol. Interact.* 5 (9): 54. (Cited on page 113).

Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021. Crowdsourcing of Parallel Corpora: the Case of Style Transfer for Detoxification. In *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021),* 35–49. Copenhagen, Denmark. (Cited on page 112).

Florenc Demrozi, Cristian Turetta, Fadi Al Machot, Graziano Pravadelli, and Philipp Kindt. 2023. A comprehensive review of automated data annotation techniques in human activity recognition. *arXiv preprint arXiv:2307.05988,* (cited on page 31).

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH),* 143–153. Seattle, WA, USA, (Hybrid): Association for Computational Linguistics. (Cited on pages 25 sq., 84, 111).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171–4186. Minneapolis, MN, USA: Association for Computational Linguistics. (Cited on pages 65, 119).

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP),* 52–64. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. (Cited on pages 42 sq., 93).

Alexey Drutsa, Viktoriya Farafonova, Valentina Fedorova, Olga Megorskaya, Evfrosiniya Zerminova, and Olga Zhilinskaya. 2019. Practice of efficient data collection via crowdsourcing at large-scale. *arXiv preprint arXiv:1912.04444,* (cited on pages 47 sq.).

Alexey Drutsa, Dmitry Ustalov, Valentina Fedorova, Olga Megorskaya, and Daria Baidakova. 2021. Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials,* 25–30. Online: Association for Computational Linguistics. (Cited on pages 48, 50).

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media,* 42–51. Palo Alto, CA, USA. (Cited on pages 25 sq.).

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* 345–363. Punta Cana, Dominican Republic, Hybride: Association for Computational Linguistics. (Cited on page 13).

Theodoros Evgeniou and Massimiliano Pontil. 2001. Support Vector Machines: Theory and Applications. In *Machine Learning and Its Applications: Advanced Lectures,* 249–257. Berlin, Heidelberg: Springer Berlin Heidelberg. (Cited on page 35).

Parvin Sadat Feizabadi and Sebastian Padó. 2014. Crowdsourcing Annotation of Non-Local Semantic Roles. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers,* 226–230. Gothenburg, Sweden: Association for Computational Linguistics. (Cited on page 45).

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 878–891. Dublin, Ireland: Association for Computational Linguistics. (Cited on page 119).

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 1905–1925. Online: Association for Computational Linguistics. (Cited on page 59).

Edgar Fieller and Egon Pearson. 1961. Tests for rank correlation coefficients: II. *Biometrika,* 29–40. (Cited on page 95).

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online,* 46–51. Vancouver, Canada: Association for Computational Linguistics. (Cited on page 70).

Joseph Fleiss'. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76 (5): 378. (Cited on page 33).

Joseph Fleiss', Bruce Levin, and Myunghee Cho Paik. 2013. Statistical methods for rates and proportions. Third Edition. john wiley & sons. (Cited on page 33).

Griffin Floto, Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Zhenwei Tang, Ali Pesaranghader, Manasa Bharadwaj, and Scott Sanner. 2023. DiffuDetox: A Mixed Diffusion Model for Text Detoxification. In *Findings of the Association for Computational Linguistics: ACL 2023,* 7566–7574. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 25 sq., 111 sq.).

Thea Forsén and Kjetil Tronvoll. 2021. Protest and political change in Ethiopia: The initial success of the Oromo Qeerroo youth movement. *Nordic Journal of African Studies* 30 (4): 19–19. (Cited on page 3).

Paula Fortuna, Monica Dominguez, Leo Wanner, and Zeerak Talat. 2022. Directions for NLP Practices Applied to Online Hate Speech Detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 11794–11805. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 50, 76).

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51 (4): 1–30. (Cited on page 13).

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 6786–6794. Marseille, France: European Language Resources Association. (Cited on pages 14, 83 sq.).

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science,* 105–114. New York City, NY, USA: Association for Computing Machinery. (Cited on pages 25 sq.).

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media,* 491–500. Palo Alto, CA, USA. (Cited on page 83).

Iginio Gagliardone, Alisha Patel, and Matti Pohjonen. 2014. Mapping and Analysing Hate Speech Online: Opportunities and Challenges for Ethiopia. University of Oxford. (Cited on page 24).

Hector Garcia-Molina, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios. 2016. Challenges in data crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* 28 (4): 901–911. (Cited on pages 45, 48, 53, 59).

Raisa Romanov Geleta, Klaus Eckelt, Emilia Parada-Cabaleiro, and Markus Schedl. 2023. Exploring Intensities of Hate Speech on Social Media: A Case Study on Explaining Multilingual Models with XAI. In *Proceedings of the 4th Conference on Language, Data and Knowledge,* 532–537. Vienna, Austria: NOVA CLUNL, Portugal. (Cited on pages 25 sq.).

Emuye Bawoke Getaneh. 2020. Amharic text hate speech detection in social media using deep learning approach. Master's thesis, Bahir Dar University, Bahir Dar, Ethiopia. (Cited on pages 26, 74).

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Binyam Ephrem Seyoum. 2018. Portable Spelling Corrector for a Less-Resourced Language: Amharic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). (Cited on page 6).

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Michael Gasser, and Andreas Nürnberger. 2018. Contemporary Amharic Corpus: Automatically Morpho-Syntactically Tagged Amharic Corpus. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing,* 65–70. Santa Fe, NM, USA: Association for Computational Linguistics. (Cited on pages 6, 23).

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Tirufat Tesifaye Lema, and Andreas Nürnberger. 2017. Manually Annotated Spelling Error Corpus for Amharic, (cited on page 6).

Natasa Gisev, Simon Bell, and Timothy Chen. 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy* 9 (3): 330–338. (Cited on page 33).

Anne Göhring and Manfred Klenner. 2022. Polar Quantification of Actor Noun Phrases for German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 1376–1380. Marseille, France: European Language Resources Association. (Cited on page 26).

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV),* 1459–1467. Snowmass Village, CO, USA: IEEE. (Cited on pages 101, 105).

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* 4:2047–2052. Montréal, Canada: IEEE. (Cited on page 40).

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning BERT for Low-Resource Natural Language Understanding via Active Learning. In *Proceedings of the 28th International Conference on Computational Linguistics,* 1158–1171. Barcelona, Spain (Online): International Committee on Computational Linguistics. (Cited on page 43).

Lewis Griffin, Bennett Kleinberg, Maximilian Mozes, Kimberly Mai, Maria Do Mar Vau, Matthew Caldwell, and Augustine Mavor-Parker. 2023. Large Language Models respond to Influence like Humans. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023),* 15–24. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 1 sq.).

Annika Grützner-Zahn, Federico Gaspari, Maria Giagkou, Stefanie Hegele, Andy Way, and Georg Rehm. 2024. Surveying the Technology Support of Languages. In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024,* 1–17. Torino, Italia: ELRA / ICCL. (Cited on page 46).

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying Text with MaRCo: Controllable Revision with Experts and Anti-Experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 228–242. Toronto, Canada: Association for Computational Linguistics. (Cited on page 112).

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic Word Embeddings for Racism Detection on Twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 926–936. Taipei, Taiwan: Asian Federation of Natural Language Processing. (Cited on page 59).

Abdalraouf Hassan and Ausif Mahmood. 2017. Deep learning approach for sentiment analysis of short texts. In *2017 3rd international conference on control, automation and robotics (ICCAR),* 705–710. Nagoya, Japan: IEEE. (Cited on page 39).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 770–778. Las Vegas, NV, USA. (Cited on page 104).

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2545–2568. Online: Association for Computational Linguistics. (Cited on page 6).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (8): 1735–1780. (Cited on pages 39 sq.).

Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 1–24. (Cited on page 59).

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. Align before Attend: Aligning Visual and Textual Features for Multimodal Hateful Content Detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop,* 162–174. St. Julian's, Malta: Association for Computational Linguistics. (Cited on page 100).

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 591–598. Berlin, Germany: Association for Computational Linguistics. (Cited on pages 7 sq.).

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 588–602. Online: Association for Computational Linguistics. (Cited on pages 6 sqq.).

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing,* 27–35. Boulder, Colorado: Association for Computational Linguistics. (Cited on pages 46, 53).

Bo Huang and Yang Bai. 2021. HUB@DravidianLangTech-EACL2021: Meme Classification for Tamil Text-Image Fusion. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages,* 210–215. Kyiv, Ukraine: Association for Computational Linguistics. (Cited on page 100).

Tzu-Yun Huang, Hsiao-Han Wu, Chia-Chen Lee, Shao-Man Lee, Guan-Wei Li, and Shu-Kai Hsieh. 2016. Crowdsourcing Experiment Designs for Chinese Word Sense Annotation. In *Proceedings of the 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016),* 82–99. Tainan, Taiwan: The Association for Computational Linguistics / Chinese Language Processing (ACLCLP). (Cited on page 53).

Mohammad Ali Hussiny and Lilja Øvrelid. 2023. Emotion Analysis of Tweets Banning Education in Afghanistan. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis,* 271–277. Toronto, Canada: Association for Computational Linguistics. (Cited on page 8).

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR Improves Machine Translation for Low-Resource Languages. In *Proceedings of the Association for Computational Linguistics: ACL 2022,* 1164–1174. Dublin, Ireland: Association for Computational Linguistics. (Cited on page 103).

Deepti Jain, Sandhya Arora, CK Jha, and Garima Malik. 2024. Transformer-based models for hate speech classification. In *AIP Proceedings of the 3072th International Conference on Intelligent and Smart Computation (ICIASC-2023).* Mohali, India: AIP Publishing. (Cited on page 25).

Alex Pappachen James. 2020. Deep Learning Classifiers with Memristive Networks. Springer. (Cited on page 40).

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. An introduction to statistical learning. Vol. 112. Springer. (Cited on pages 34 sq.).

Melese Ayichlie Jigar, **Abinew Ali Ayele**, Seid Muhie Yimam, and Chris Biemann. 2024. Detecting Hate Speech in Amharic Using Multimodal Analysis of Social Media Memes. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024,* 85–95. Torino, Italia: ELRA / ICCL. (Cited on pages 17, 19, 21, 99).

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers,* 427–431. Valencia, Spain: Association for Computational Linguistics. (Cited on page 102).

Satyajit Kamble and Aditya Joshi. 2018. Hate Speech Detection from Code-mixed Hindi-English Tweets Using Deep Learning Models. In *Proceedings of the 15th International Conference on Natural Language Processing,* 150–155. Hyderabad, India: NLP Association of India. (Cited on pages 39 sqq.).

Lata Guta Kanessa and Solomon Gizaw Tulu. 2021. Automatic Hate and Offensive speech detection framework from social media: the case of Afaan Oromoo language. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA),* 42–47. Bahir Dar, Ethiopia: IEEE. (Cited on pages 26 sq.).

Prashant Kapil and Asif Ekbal. 2020. Leveraging Multi-domain, Heterogeneous Data using Deep Multitask Learning for Hate Speech Detection. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON),* 491–500. Patna, India: NLP Association of India (NLPAI). (Cited on pages 25 sq.).

Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2023. Multimodal Hate Speech Detection from Bengali Memes and Texts. In *Proceedings of the first International Conference on Speech and Language Technologies for Low-Resource Languages,* 293–308. Beijing, China: Springer International Publishing. (Cited on pages 104 sq.).

Simon Kemp. 2024. Digital 2024: Global Overview Report. Technical report. Last accessed: March 1, 2024. DataReportal. (Cited on page 2).

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment Analysis: It's Complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* 1886–1895. New Orleans, Louisiana: Association for Computational Linguistics. (Cited on page 7).

Christoph Kern, Stephanie Eckman, Jacob Beck, Rob Chew, Bolei Ma, and Frauke Kreuter. 2023. Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 14874–14886. Singapore, Singapore: Association for Computational Linguistics. (Cited on page 69).

Geoffrey Khan, Michael P Streck, and Janet CE Watson. 2011. The Semitic languages: An international handbook. Walter de Gruyter. (Cited on page 23).

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020),* 2611–2624. Vancouver, Canada. (Cited on pages 100 sq.).

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *J. Artif. Intell. Res.* 71:431–478. (Cited on page 3).

Büşra Kocaçınar, Nasibullah Qarizada, Cihan Dikkaya, Emirhan Azgun, Elif Yıldırım, and Fatma Patlar Akbulut. 2023. Analysis of the Lingering Effects of Covid-19 on Distance Education. In *IFIP International Conference on Artificial Intelligence Applications and Innovations,* 189–200. Cham, Switzerland: Springer. (Cited on page 8).

Nancy Krieger. 1999. Embodying inequality: a review of concepts, measures, and methods for studying health consequences of discrimination. *International journal of health services* 29 (2): 295–352. (Cited on page 59).

Anoop Kunchukuttan, Rajen Chatterjee, Shourya Roy, Abhijit Mishra, and Pushpak Bhattacharyya. 2013. TransDoop: A Map-Reduce based Crowdsourced Translation for Complex Domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 175–180. Sofia, Bulgaria: Association for Computational Linguistics. (Cited on page 45).

Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1 (4): 541–551. (Cited on page 39).

Yuh-Jye Lee, Yi-Ren Yeh, and Hsing-Kuo Pao. 2012. Introduction to Support Vector Machines and Their Applications in Bankruptcy Prognosis. In *Handbook of Computational Finance,* 731–761. Berlin, Heidelberg: Springer Berlin Heidelberg. (Cited on page 35).

Paul LeGendre, Asuman İnceoğlu, Michael Lieberman, Alina Plata, Andreas Stegbauer, and Alexander Verkhovsky. 2022. Hate Crime Laws: A Practical Guide. Warsaw, Poland: Organization for Security / Co-operation in Europe (OSCE)/Office for Democratic Institutions / Human Rights (ODIHR). (Cited on page 13).

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* 10528–10539. Punta Cana, Dominican Republic: Association for Computational Linguistics. (Cited on page 70).

Jiyi Li and Fumiyo Fukumoto. 2019. A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP,* 24–28. Hong Kong: Association for Computational Linguistics. (Cited on page 45).

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 9114–9128. Singapore, Singapore: Association for Computational Linguistics. (Cited on page 100).

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 3125–3135. Florence, Italy: Association for Computational Linguistics. (Cited on page 12).

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23 (1): 18. (Cited on page 34).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 8:726–742. (Cited on page 119).

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with Parallel Data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 6804–6818. Dublin, Ireland: Association for Computational Linguistics. (Cited on pages 25 sq., 111, 113, 119).

Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *31st Conference on Neural Information Processing Systems (NIPS 2017),* 1–10. Long Beach, CA, USA: Curran Associates, Inc. (Cited on page 116).

Chu Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Legally Enforceable Hate Speech Detection for Public Forums. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 10948–10963. Singapore, Singapore: Association for Computational Linguistics. (Cited on page 13).

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms,* 150–161. Online: Association for Computational Linguistics. (Cited on pages 24, 101).

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264,* (cited on page 46).

Christopher Manning and Hinrich Schütze. 1999. Foundations of statistical natural language processing. MIT press. (Cited on pages 37 sq.).

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. Cambridge university press. (Cited on page 37).

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoıt Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 7203–7219. Online: Association for Computational Linguistics. (Cited on page 65).

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media* 22 (2): 205–224. (Cited on page 59).

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence,* 14867–14875. Palo Alto, CA, USA: Association for the Advancement of Artificial Intelligence. (Cited on pages 3, 13 sq., 25, 47, 54, 59 sq., 65, 74 sq., 83).

Qing Meng, Tharun Suresh, Roy Ka-Wei Lee, and Tanmoy Chakraborty. 2023. Predicting hate intensity of Twitter conversation threads. *Knowledge-Based Systems* 275:1–14. (Cited on page 26).

Muhabie Mekonnen Mengistu. 2015. Ethnic federalism: A means for managing or a triggering factor for ethnic conflicts in Ethiopia. *Social Sciences* 4 (4): 94–105. (Cited on pages 27 sq.).

Ibomoiye Domor Mienye, Theo G. Swart, and George Obaido. 2024. Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information* 15 (9). (Cited on pages 39 sq.).

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering Aggression from the Multilingual Social Media Feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018),* 199–207. Santa Fe, NM, USA: Association for Computational Linguistics. (Cited on page 58).

Arturo Montejo-Ráez and Salud María Jiménez-Zafra. 2022. Current Approaches and Applications in Natural Language Processing. *Applied Sciences* 12 (10): 4859. (Cited on page 6).

Zewdie Mossie and Jenq-Haur Wang. 2018. Social network hate speech detection for Amharic language. *Computer Science & Information Technology,* 41–55. (Cited on pages 26, 35 sq., 47, 57, 74, 84, 99, 113).

———. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management* 57 (3). (Cited on pages 26 sq., 47, 57, 74).

Ghaderi Hajat Mostafa and Mirzaei Tabar Meysam. 2023. The impact of spatial injustice on ethnic conflict in Ethiopia. *Geopolitics Quarterly* 19 (70): 41–65. (Cited on pages 28, 89).

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, **Abinew Ali Ayele**, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 13968–13981. Singapore, Singapore: Association for Computational Linguistics. (Cited on pages 9 sq., 18, 20).

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, **Abinew Ali Ayele**, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023),* 2319–2337. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 12, 18, 20).

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr. Ojha, John P. McCrae, and Ondrej Dusek. 2023. Text Detoxification as Style Transfer in English and Hindi. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON),* 133–144. Goa, India: NLP Association of India (NLPAI). (Cited on page 112).

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online,* 111–118. Florence, Italy. (Cited on page 74).

Wondwossen Mulugeta, Michael Gasser, and Baye Yimam. 2012. Incremental Learning of Affix Segmentation. In *Proceedings of COLING 2012,* 1901–1914. Mumbai, India: The COLING 2012 Organizing Committee. (Cited on page 6).

Made Nindyatama Nityasya, Haryo Wibowo, Alham Fikri Aji, Genta Winata, Radityo Eko Prasojo, Phil Blunsom, and Adhiguna Kuncoro. 2023. On "Scientific Debt" in NLP: A Case for More Rigour in Language Model Pre-Training Research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 8554–8572. Toronto, Canada: Association for Computational Linguistics. (Cited on page 2).

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web,* 145–153. Montréal, Canada: ACM Digital Library. (Cited on page 25).

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning,* 116–126. Punta Cana, Dominican Republic: Association for Computational Linguistics. (Cited on pages 42 sq., 93).

Emily Öhman. 2020. Challenges in annotation: annotator experiences from a crowdsourced emotion annotation task. In *Proceedings of the 5th Conference on Digital Humanities in the Nordic Countries,* 293–301. Riga, Latvia: CEUR Workshop Proceedings. (Cited on pages 45 sq., 48, 53, 59, 62).

Megersa Oljira. 2020. Sentiment Analysis of Afaan Oromo using Machine learning Approach. *International Journal of Research Studies in Science, Engineering and Technology* 7 (9): 7–15. (Cited on page 9).

Abdurahman Omar. 2020. The Ethiopian Muslims Protest in the Era of Social Media Activism. Master's thesis, Uppsala University, Uppsala, Sweden. (Cited on page 3).

OpenAI. 2024. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774,* (cited on pages 7, 26, 111).

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 4675–4684. Hong Kong, China: Association for Computational Linguistics. (Cited on pages 25 sq., 47, 54, 75).

Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & intelligent systems* 9 (3): 2827–2842. (Cited on page 13).

Hyeoun-Ae Park. 2013. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean academy of nursing* 43 (2): 154–164. (Cited on page 36).

Hyoungjun Park, Ho Shim, and Kyuhan Lee. 2023. Uncovering the Root of Hate Speech: A Dataset for Identifying Hate Instigating Speech. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 6236–6245. Singapore, Singapore: Association for Computational Linguistics. (Cited on page 25).

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper Attention to Abusive User Content Moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 1125–1135. Copenhagen, Denmark: Association for Computational Linguistics. (Cited on pages 3, 113).

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The Portable Text Annotation Tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 327–337. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 86, 114).

Wondwossen Philemon and Wondwossen Mulugeta. 2014. A Machine Learning Approach to Multi-Scale Sentiment Analysis of Amharic Online Posts. *HiLCoE Journal of Computer Science and Technology* 2 (2): 8. (Cited on page 9).

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 4996–5001. Florence, Italy: Association for Computational Linguistics. (Cited on page 105).

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 10671–10682. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on page 32).

Flor Miriam Plaza-del-arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? Using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH),* 60–68. Toronto, Canada: Association for Computational Linguistics. (Cited on page 83).

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers,* 186–191. Brussels, Belgium: Association for Computational Linguistics. (Cited on page 119).

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021,* 4439–4455. Punta Cana, Dominican Republic: Association for Computational Linguistics. (Cited on pages 100 sq.).

Jack Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446,* (cited on page 69).

Ochilbek Rakhmanov and Tim Schlippe. 2022. Sentiment Analysis for Hausa: Classifying Students' Comments. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages,* 98–105. Marseille, France: European Language Resources Association. (Cited on page 8).

Justus Randolph. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission,* (cited on pages 10, 33).

Kyle Rapp. 2021. Social media and genocide: The case for home state responsibility. *Journal of Human Rights* 20 (4): 486–502. (Cited on page 13).

Megersa Oljira Rase. 2020. Sentiment analysis of Afaan Oromoo facebook media using deep learning approach. *New Media and Mass Communication* 90 (2020): 2224–3267. (Cited on page 9).

Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter van Atteveldt. 2021. Are we human, or are we users? The role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *Proceedings of the 1st Workshop on NLP for Positive Impact,* 47–59. Online: Association for Computational Linguistics. (Cited on page 2).

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* 60 (5): 503–520. (Cited on page 25).

Gabriel Roccabruna, Steve Azzolin, and Giuseppe Riccardi. 2022. Multi-source Multi-domain Sentiment Analysis with BERT-based Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 581–589. Marseille, France: European Language Resources Association. (Cited on pages 7 sq.).

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to Action on Social Media: Detection, Social Impact, and Censorship Potential. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda,* 36–44. Hong Kong, China: Association for Computational Linguistics. (Cited on pages 2 sq.).

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society* 46 (5): 621–647. (Cited on page 59).

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118,* (cited on page 61).

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 175–190. Seattle, WA, USA: Association for Computational Linguistics. (Cited on pages 31 sq.).

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-Efficient Strategies for Expanding Hate Speech Detection into Under-Resourced Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 5674–5691. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 6, 25).

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH),* 154–169. Seattle, WA, USA: Association for Computational Linguistics. (Cited on page 25).

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The Measuring Hate Speech Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022,* 83–94. Marseille, France: European Language Resources Association. (Cited on page 84).

Abiodun Salawu and Asemahagn Aseres. 2015. Language policy, ideologies, power and the Ethiopian media. *South African Journal for Communication Theory and Research* 41 (1): 71–89. (Cited on page 23).

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM* 18 (11): 613–620. (Cited on page 37).

Eric Sanders and Antal van den Bosch. 2020. Optimising Twitter-based Political Election Prediction with Relevance andSentiment Filters. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 6158–6165. Marseille, France: European Language Resources Association. (Cited on page 8).

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). (Cited on page 26).

Ahmed El-Sayed and Omar Nasr. 2024. AAST-NLP at Multimodal Hate Speech Event Detection: A Multimodal Approach for Classification of Text-Embedded Images Based on CLIP and BERT-Based Models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024),* 139–144. St. Julians, Malta: Association for Computational Linguistics. (Cited on page 25).

Salim Sazzed. 2023. Discourse Mode Categorization of Bengali Social Media Health Text. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis,* 52–57. Toronto, Canada: Association for Computational Linguistics. (Cited on page 2).

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media,* 1–10. Valencia, Spain: Association for Computational Linguistics. (Cited on pages 100 sqq.).

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia* 126 (5): 1763–1768. (Cited on page 95).

Egon Schwelb. 1966. The international convention on the elimination of all forms of racial discrimination. *International & Comparative Law Quarterly* 15 (4): 996–1068. (Cited on pages 58 sq.).

Qinlan Shen and Carolyn Rose. 2019. The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. In *Proceedings of the Third Workshop on Abusive Language Online,* 58–69. Florence, Italy: Association for Computational Linguistics. (Cited on pages 13 sq.).

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH),* 102–108. Seattle, WA, USA (Hybrid): Association for Computational Linguistics. (Cited on page 60).

Alane Suhr, Clara Vania, Nikita Nangia, Maarten Sap, Mark Yatskar, Samuel R. Bowman, and Yoav Artzi. 2021. Crowdsourcing Beyond Annotation: Case Studies in Benchmark Data Collection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts,* 1–6. Punta Cana, Dominican Republic & Online: Association for Computational Linguistics. (Cited on pages 45 sq., 59).

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying,* 32–41. Marseille, France: European Language Resources Association (ELRA). (Cited on page 101).

Narges Tabari, Armin Seyeditabari, and Wlodek Zadrozny. 2017. SentiHeros at SemEval-2017 Task 5: An application of Sentiment Analysis on Financial Tweets. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017),* 857–860. Vancouver, Canada: Association for Computational Linguistics. (Cited on pages 7 sq.).

Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation: A Survey. *arXiv preprint arXiv:2402.13446,* (cited on page 32).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual Translation from Denoising Pre-Training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 3450–3466. Online: Association for Computational Linguistics. (Cited on page 119).

Xiangyu Tao and Celia Fisher. 2022. Exposure to social media racial discrimination and mental health among adolescents of color. *Journal of youth and adolescence* 51 (1): 30–44. (Cited on page 58).

Bekalu Atnafu Taye. 2017. Ethnic federalism and conflict in Ethiopia. *African journal on conflict resolution* 17 (2): 41–66. (Cited on page 28).

Abrhalei Frezghi Tela. 2020. Sentiment Analysis for Low-Resource Language: The Case of Tigrinya. Master's thesis, Itä-Suomen yliopisto. (Cited on page 9).

Surafel Getachew Tesfaye and Kula Kakeba. 2020. Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network. *Preprint paper,* (cited on pages 47, 74, 84, 99, 113).

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Jafri, Shuvam Shiwakoti, Hariram Veeramani, Raghav Jain, Guneet Singh Kohli, Ali Hürriyetoğlu, and Usman Naseem. 2024. Stance and Hate Event Detection in Tweets Related to Climate Activism - Shared Task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024),* 234–247. St. Julians, Malta: Association for Computational Linguistics. (Cited on page 101).

Surendrabikram Thapa, Aditya Shah, Farhan Jafri, Usman Naseem, and Imran Razzak. 2022. A Multi-Modal Dataset for Hate Speech Detection on Social Media: Case-study of Russia-Ukraine Conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE),* 1–6. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. (Cited on page 101).

Christoph Tillmann, Aashka Trivedi, Sara Rosenthal, Santosh Borse, Rong Zhang, Avirup Sil, and Bishwaranjan Bhattacharjee. 2023. Muted: Multilingual Targeted Offensive Speech Identification and Visualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 229–236. Singapore, Singapore: Association for Computational Linguistics. (Cited on page 26).

Tanya Tiwari, Tanuj Tiwari, and Sanjay Tiwari. 2018. How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different? *International Journal of Advanced Research in Computer Science and Software Engineering* 8 (2): 1–9. (Cited on pages 25, 34).

Toloka. 2024. Toloka: Crowdsourcing Platform for Data Labeling and AI Training. Accessed: 2024-11-27. (Cited on page 65).

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, **Abinew Ali Ayele**, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023),* 126–139. Dubrovnik, Croatia: Association for Computational Linguistics. (Cited on pages 6, 9, 18, 20, 26).

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics* 126 (1): 157–179. (Cited on page 3).

Minh Tran, Yipeng Zhang, and Mohammad Soleymani. 2020. Towards A Friendly Online Community: An Unsupervised Style Transfer Framework for Profanity Redaction. In *Proceedings of the 28th International Conference on Computational Linguistics,* 2107–2114. Barcelona, Spain (Online): International Committee on Computational Linguistics. (Cited on page 113).

Thanh Tran, Yifan Hu, Changwei Hu, Kevin Yen, Fei Tan, Kyumin Lee, and Se Rim Park. 2020. HABERTOR: An Efficient and Effective Deep Hatespeech Detector. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 7486–7502. Online: Association for Computational Linguistics. (Cited on pages 39 sqq.).

Natalia Vanetik and Elisheva Mimoun. 2022. Detection of Racist Language in French Tweets. *Information* 13 (7). (Cited on page 59).

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 7174–7183. Marseille, France: European Language Resources Association. (Cited on pages 35 sq.).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems,* 6000–6010. NIPS'17. Long Beach, CA, USA: Curran Associates Inc. (Cited on pages 41 sq.).

Jennifer Wortman Vaughan. 2017. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *J. Mach. Learn. Res.* 18 (1): 7026–7071. (Cited on page 47).

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975,* (cited on page 101).

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online,* 80–93. Florence, Italy: Association for Computational Linguistics. (Cited on page 13).

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation* 47 (1): 9–31. (Cited on page 46).

Zhenji Wang, Dan Tao, and Pingping Liu. 2015. Development and Challenges of Crowdsourcing Quality of Experience Evaluation for Multimedia. In *In Proceedings of the First International Conference on Big Data Computing and Communications - BigCom 2015,* 9196:444–452. Lecture Notes in Computer Science. Taiyuan, China: Springer. (Cited on pages 45 sq., 59).

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media,* 19–26. Montréal, Canada: Association for Computational Linguistics. (Cited on page 25).

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop,* 88–93. San Diego, California: Association for Computational Linguistics. (Cited on pages 3, 25, 70, 83).

Negessa Wayessa and Sadik Abas. 2020. Multi-Class Sentiment Analysis from Afaan Oromo Text Based On Supervised Machine Learning Approaches. *International Journal of Research Studies in Science, Engineering and Technology* 7 (7): 10–18. (Cited on page 9).

Kevin Winter and Roman Kern. 2019. Know-Center at SemEval-2019 Task 5: Multilingual Hate Speech Detection on Twitter using CNNs. In *Proceedings of the 13th International Workshop on Semantic Evaluation,* 431–435. Minneapolis, MN, USA: Association for Computational Linguistics. (Cited on page 25).

Valerie Wirtschafter. 2021. How George Floyd changed the online conversation around BLM, (cited on page 60).

Michael Wojatzki, Torsten Zesch, Saif Mohammad, and Svetlana Kiritchenko. 2018. Agree or Disagree: Predicting Judgments on Nuanced Assertions. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics,* 214–224. New Orleans, LA, USA: Association for Computational Linguistics. (Cited on page 90).

Fan Wu, Guolian Chen, Junkuo Cao, Yuhan Yan, and Zhongneng Li. 2024. Multimodal Hateful Meme Classification Based on Transfer Learning and a Cross-Mask Mechanism. *Electronics* 13 (14). (Cited on pages 101 sq.).

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional BERT Contextual Augmentation. In *Computational Science – ICCS 2019,* 84–95. Cham, Switzerland: Springer International Publishing. (Cited on page 112).

Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018. Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). (Cited on page 8).

Tariku Birhanu Yadesa, Syed Umar, and Tagay Takele Fikadu. 2020. Sentiment Mining Model for Opinionated Afaan Oromo Texts, (cited on page 9).

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual Content Moderation: A Case Study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics,* 3828–3844. Dubrovnik, Croatia: Association for Computational Linguistics. (Cited on page 3).

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, **Abinew Ali Ayele**, and Chris Biemann. 2020. Exploring Amharic Sentiment Analysis from Social Media Texts: Building Annotation Tools and Classification Models. In *Proceedings of the 28th International Conference on Computational Linguistics,* 1048–1060. Barcelona, Spain (Online): International Committee on Computational Linguistics. (Cited on pages 9 sq.).

Seid Muhie Yimam, Abinew Ali Ayele, and Chris Biemann. 2019. Analysis of the Ethiopic Twitter dataset for abusive speech in Amharic. *arXiv preprint arXiv:1912.04419,* (cited on pages 49, 73, 101).

Seid Muhie Yimam, **Abinew Ali Ayele**, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for Amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet* 13 (11): 275. (Cited on pages 6, 23, 26, 42 sq., 93, 105).

Seid Muhie Yimam, Daryna Dementieva, Tim Fischer, Daniil Moskovskiy, Naquee Rizwan, Punyajoy Saha, Sarthak Roy, Martin Semmann, Alexander Panchenko, Chris Biemann, et al. 2024. Demarked: A Strategy for Enhanced Abusive Speech Moderation through Counterspeech, Detoxification, and Message Management. *arXiv preprint arXiv:2406.19543,* (cited on pages 111, 113).

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A Flecxible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 1–6. Sofia, Bulgaria: Association for Computational Linguistics. (Cited on pages 74, 77).

Elif Yıldırım, Harun Yazgan, Onur Özbek, Ahmet Can Günay, Büşra Kocaçınar, Öznur Şengel, and Fatma Patlar Akbulut. 2023. Sentiment Analysis of Tweets on Online Education during COVID-19. In *IFIP International Conference on Artificial Intelligence Applications and Innovations,* 240–251. Cham, Switzerland: Springer. (Cited on page 8).

Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2022. A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task. In *Proceedings of the Natural Legal Language Processing Workshop 2022,* 53–64. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. (Cited on pages 3, 24).