Journal of Computer Science

Original Research Paper

EthioLSocMDMTLM: Exploring Application of Topic Modeling for Building Ethiopian Language Social Media Data-Based Multilingual Transformer Language Models for Multilingual Hateful Content Detection

[1]Naol Bakala Defersha, [1]Kula Kekeba Tune and [2]Solomon Teferra Abate

[1]*Software Engineering, College of Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia*
[2]*School of Information Science, College of Natural and Computational Science, Addis Ababa University, Addis Ababa, Ethiopia*

Article history
Received: 29-10-2024
Revised: 14-11-2024
Accepted: 25-11-2024

Corresponding Author:
Naol Bakala
Software Engineering, College of Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia
Email: naol.bakala@aastustudent.edu.et

Abstract: This study proposes topic modeling techniques to develop Ethiopian Language Social Media Data Based Multilingual Transformer Language Models for multilingual hateful content detection. We modified various multilingual pretrained models, investigated the challenges of using pre-trained transformer language models, and built multilingual hateful content detection models. Topic words with rows of 1561, 70, and 1044 extracted from Afaan Oromo, Tigrigna, and Amharic Afaan Oromo, Amharic, and Tigrigna respectively used to train transformers. The proposed models were also tested by developing a multilingual hateful content detection model for low-resource Ethiopian languages using deep learning techniques. A total of 45522, 59529, and 48882, Tex documents of Amharic, Afaan Oromo, and Tigrigna were collected and three annotators annotated the data into binary classes where the agreement among annotators result scored 87% for Amharic, 82% for Tigrigna and 84% for Afaan Oromo. LSTM, CNN, and BiLSM deep learning algorithms applied algorithms, that includes integration of EthioLan_mBERT, EthioLan_BERT, and EthioLan_XLM-Roberta contextual embeddings. Among applied the techniques; LSTM+ EthioLan_mBERT outperforms the score performance of F1score 81%. We publicly release the modified pre-trained models, dataset, and related codes.

Keywords: Afaan Oromo, Low Resource Languages, Amharic, Hateful Content, EthioLSocMDMTLM, Transformer Language Model, Tigrigna

Introduction

The proliferation of social media networks and online communication platforms has significantly increased the dissemination of both positive and negative content (Korre *et al*., 2024; Biradar *et al*., 2022; Lee *et al*., 2023; Pradanna and Abdulkarim, 2023; Unlu *et al*., 2024). The UNSECO report highlights the widespread spread of hateful content on various social media platforms such as Facebook, TikTok, Instagram, Telegram, X/Twitter, Whatsup, Snapchat, Signal, and others, with Facebook being the most prominent network contributing to this issue (UNESCO/Ipsos, 2023). Across the world, hate speech is disseminated in different countries such as Algeria, Austria, Bangladesh, Belgium, Croatia, "El Salvador, Dominican Republic, Ghana, India, Indonesia, Mexico, Romania, Senegal, South Africa, Ukraine, and the United States" (UNESCO/Ipsos, 2023).

To minimize the impact of hateful content, countries like Ethiopia set various rules and regulations (Defersha *et al*., 2024; Ayele *et al*., 2023). Moreover, the researcher's interest also increased in developing a technological model to overcome those challenges. Accordingly, researchers such as Biradar *et al*. (2022); (Ababu and Woldeyohannis, 2022) have proposed a bilingual hateful content detection model for Ethiopian indigenous languages such as Amharic and Afaan Oromo. Researchers also attempted to develop multilingual hateful content detection models and developed multilingual datasets for Ethiopian languages such as Tigrigna, Amharic, and Afaan Oromo (Defersha *et al*., 2024). Although the researchers claim the current state of arts performance is based on their own dataset, the model cannot be implemented for all languages and datasets (Gröndahl *et al*., 2018).

Artificial intelligence has advanced significantly with the use of large language models (Kasneci *et al*., 2023).

Particularly little focus has been placed on African languages in this regard (Ogueji *et al*., 2021). Despite the transformer language models having great advantages over resource-scarce languages, most researchers did not explore the application of the Transformer language model in low-resource language hateful content detection.

Ethiopia like other country try to develop legislation to minimize the impact of hateful content disseminated on social media platforms. This is because of the anonymous internet users who propagate hostile words while hiding behind their screens and the complicated nature of the online community, which is hard to regulate by local authorities (Ayele *et al*., 2023). Moreover, the existing transformer language models and hateful content detection model developed for resource-rich languages such as English are not applicable for resource-scarce languages such as Ethiopian languages due to cultural, linguistic social, and political differences. Therefore, we need to train Ethiopian Language Multilingual Social Media Data Transformer Language Models (EthioLSocMDMTLM) and explore their application in developing a multilingual hateful content detection model for low-resource Ethiopian languages.

Hence, in this study, we train Ethiopian language Multilingual Social Media Data Transformer Language Models (EthioLSocMDMTLM) by exploring the application of topic modeling techniques. This study addressed the following research questions:

- RQ1: Can we apply the topic extraction approach (BERTopic, PLSA, LSA, NMFI, and LDA) to train EthioLSocMDMTLM (EthioLan_mBERT, EthioLan_BERT and EthioLan_XLM-Roberta) by using transformer language models such as Roberta, mBERT, xlmr-roberta and BERT to train models in cost-effective?
- RQ2: Which EthioLSocMDMTLM from EthioLan_BERT, EthioLan_mBERT, and EthioLan-xlm-roberta-base work effectively with a multilingual dataset of Afaan Oromo, Tigrigna and Amharic in developing multilingual hateful content detection model?
- RQ3: Which deep learning algorithms (CNN, lstm, and bilstm) outperform in developing a hateful Content detection model based on EthioLSocMDMTLM?

Our work contributes the following:

(1) Producing transformer language models such as EthioLan_BERT, EthioLan_mBERT, and EthioLan-xlm-roberta-base for low-resource Languages, including Afaan Oromo, Tigrigna, and Amharic in a cost-effective way
(2) Investigating the application of topic modeling in generating EthioLSocMDMTLM for detecting hateful content for low-resource Ethiopian languages using deep learning algorithms such as LSTM, CNN, and BiLSTM
(3) Publishing multilingual Ethiopian Languages social media-based multilingual pretrained Transformer language models

Literature Review

Hateful content is the content that media users post on social media to attack others based on their political, social, cultural, religious, ethnic, or other perspectives that lead to violence. Many researchers attempted to address the prevalence of hate speech based on data gathered from social media platforms.

The proliferation of hateful content is a significant issue that prevails significant threats to individuals and communities (Wang *et al*., 2022). Hate speech, which incites violence against individuals based on ethnicity, race, religion, sexual orientation, or gender poses a great challenge to creating inclusive and respectful online environments (Ababa *et al*., 2021; Naidu and Kumar, 2021; Ayele *et al*., 2023; Zufall *et al*., 2022). Yimam *et al*. (2019), confirmed that online hate speech can have real-world repercussions by dividing society, escalating hatred, and, in certain cases, inspiring violence in Amharic (Yimam *et al*., 2019). Hateful content emerging on social media also leads to violence and genocide (Defersha *et al*., 2021; 2024; Ababa *et al*., 2021; Mcgowan *et al*., 2024). Consequently, platform managers, content moderators, and legislators attempt to identify and address hate speech through online conversations. Researchers are also putting more effort into creating plans to minimize the impact of hateful content (Ababa *et al*., 2021; Ababu and Woldeyohannis, 2022; Ayele *et al*., 2024). In the last few years, researchers from a wide range of academic fields, including computer science, media and communications studies, psychology, social science and psychology, have been more interested in studying hate speech (Ababu and Woldeyohannis, 2022; Defersha *et al*., 2022; 2024; Mathew *et al*., 2021; Kemal *et al*., 2023; Tontodimamma *et al*., 2021; Tesfaye and Tune, 2020; Kasule *et al*., 2023) and others. According to Bahador (2023), there is a continuum of continuity in the categorization of hate speech (Bahador, 2023). The authors applied transformers (Truică and Apostol, 2022); (Shifath *et al*., 2021), word embedding (Truică and Apostol, 2023; Faris *et al*., 2020; Lu, 2023), and topic modeling (Defersha *et al*., 2024; Calderón *et al*., 2020) and lexicons (Hatzivassiloglou and McKeown, 1997; Ibrahim *et al*., 2024) to develop a highly content-detection model for developed languages.

The study evaluates Roberta and XLNet transformer-based language models on social media datasets, revealing significant performance improvements over baselines and demonstrating minimal computational cost and complex model engineering (Mukherjee and Das, 2021), biomedical (Calderón *et al*., 2020) and social media (Beltagy *et al*., 2019). The approach outperforms RoBERTa and XLNet on a scale using less than 1/4 of their compute resources (Clark *et al*., 2020). The study reveals that BERT-based models outperform LASER embedding with logistic regression in high-resource scenarios, while Portuguese and Italian perform well in zero-shot classification (Aluru *et al*., 2020). The multilingual BERT (mBERT) model outperforms the others with a fine-tuning process for hateful content identification in resource scarcity. In Natural Language Processing (NLP) tasks, the effectiveness of Pre-Trained Language Models (PLMs) on domain-specific data has been demonstrated (Guo and Sarker, 2023). For example, Guo and Sarke (2023) created a language model (SocBERT) that functions well for natural language processing (NLP) applications utilizing data gathered from Reddit and Twitter for English (Guo and Sarker, 2023). The experiment showed that domain-specific data became effective in natural language processing tasks using information from several social media platforms (Guo and Sarker, 2023).

Researchers have explored various algorithms and methodologies for extracting topics from social media content, focusing on social media networks such as Facebook, YouTube, Reddit, and Twitter because of the large amount of data emerging on social media platforms (Defersha *et al*., 2024; Calderón *et al*., 2020; Liu and Forss, 2015; Vayansky and Kumar, 2020; Zhang *et al*., 2022; Alshalan *et al*., 2020).

The existing studies for Afaan Oromo (Ababa *et al*., 2021; Defersha *et al*., 2021; Ababu and Woldeyohannis, 2022; Ganfure, 2022) for Amharic (Ayele *et al*., 2023; Tesfaye and Tune, 2020; Melat, 2022; Mossie and Wang, 2018) and Tigrigna (Weldemariam, 2022) use monolingual hateful content detection approaches. The other study also focuses on developing bilingual hateful content detection for these languages (Ababu and Woldeyohannis, 2022; Kemal *et al*., 2023).

Research Gaps

Resources-scarce languages like Amharic, Tigrigna, and Afaan Oromo require big text documents for training on non-GPU-based devices, making the transformer language model which is used to detect hateful content in resource-rich languages like English unsuitable for these languages. Despite filtering representative terms and creating dimensionally reduced documents, researchers have not yet investigated the applicability of the topic modeling approach in the building of transformer language models.

Materials and Methods

In this methodology section, we introduced all the materials and methods that have been used to explore the application of topic modeling in building EthioLSocMDMTLM and applied it to develop a multilingual hateful content detection model for resource-scarce Ethiopian languages such as Afaan Oromo, Tigrigna, and Amharic (Fig. 1).

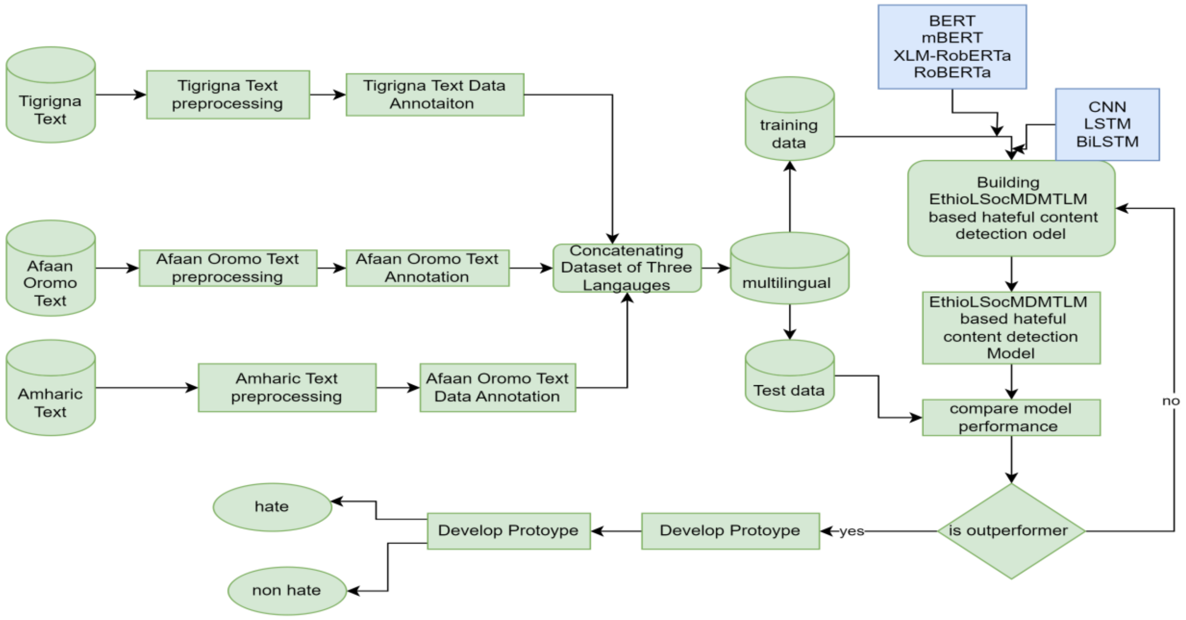

Tigrigna Text
Tigrigna Text preprocessing
Tigrigna Text Data Annotaiton
Afaan Oromo Text
Afaan Oromo Text preprocessing
Afaan Oromo Text Annotation
Amharic Text
Amharic Text preprocessing
Afaan Oromo Text Data Annotation
Concatenating Dataset of Three Langauges
multilingual
training data
Test data
BERT mBERT XLM-RobERTa RoBERTa
CNN LSTM BiLSTM
Building EthioLSocMDMTLM based hateful content detection odel
EthioLSocMDMTLM based hateful content detection Model
compare model performance
is outperformer
no
yes
Develop Protoype
Develop Protoype
hate
non hate

Fig. 1: Proposed multilingual hateful content detection model for Ethiopian languages using EthioLSocMDMTLM

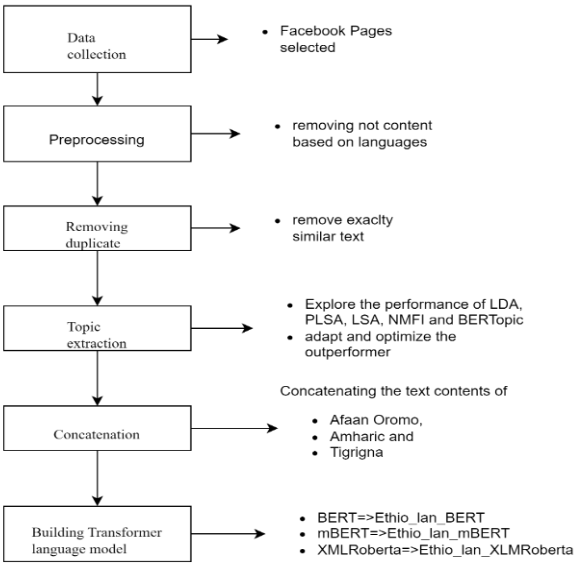

Data collection
Facebook Pages selected
Preprocessing
removing not content based on languages
Removing duplicate
remove exaclty similar text
Topic extraction
Explore the performance of LDA, PLSA, LSA, NMFI and BERTopic
adapt and optimize the outperformer
Concatenation
Concatenating the text contents of
Afaan Oromo,
Amharic and
Tigrigna
Building Transformer language model
BERT=>Ethio_lan_BERT
mBERT=>Ethio_lan_mBERT
XMLRoberta=>Ethio_lan_XLMRoberta

Fig. 2: EthioLSocMDMTLM construction through topic modeling pipeline

The study investigates the use of topic modeling in the development of the multilingual hateful content detection model for Ethiopians, known as EthioLSocMDMTLM, and its implementation in multilingual hateful content detection (Fig. 2).

Collection of Data

To prepare a dataset for the experiment we used our dataset from (Defersha *et al*., 2024). As indicated in (Defersha *et al*., 2024) Facebook is predominantly used in Ethiopia, with about 6.9 million users as of November 2022. Information from Facebook pages was collected manually between September 1, 2019, and August 1, 2022 (Defersha *et al*., 2024). Data was collected using selected and trained data (Defersha *et al*., 2024).

Sampling of data: We employed different techniques for data collection, cleaning, and sampling approach. Duplicate data and different languages were deleted from each dataset. Experts in three languages collected data for three languages a specified duration from selected sources due to the emergence of COVID-19, the Federal Government of Ethiopia's with the Oromo Liberation Army and Tigray People's Liberation Front, the imprisonment of political leaders, and the assassinations of Artists Hachalu Hundessa.

To annotate these datasets, we also used keywords collected from native speakers of three languages using a Google form.

Text Preprocessing

Text processing methods tailored to Twitter data are used to improve model performance (Macias *et al*., 2023). Text preprocessing tasks include removing HTML, removing EMOJIS, lowercase case conversion, apostrophes, punctuation, numbers, repeated characters, and alphanumeric words from text documents to obtain cleaned documents (Macias *et al*., 2023). Text preprocessing tasks vary depending on the types of language in text documents.

Tokenization: Tokenization is the task of splitting plain text into words based on white space between the words in the sentence, clause, and phrase in the text document. Accordingly, tokenization tasks were applied to Afaan Oromo (Defersha *et al*., 2024), Tigrigna (Weldemariam, 2022), and Amharic (Tesfaye and Tune, 2020) text documents to split the plain text of languages into a list of words in languages.

Lowercasing: In Afaan Oromo, one letter can be written in capital and small letters. Therefore, at the lowering stage, all uppercase letters are converted into lowercase.

Normalization: In Amharic, there is a list of letters that have the same sound. For example, ሓ, ሐ, ኀ, ሃ and ኃ are normalized to ሀ (Teshome, 2013).

Removing numbers: Three types of numbers need to be removed from the text document to produce a text document free from numbers. These numbers inlcude Roman numerals such as I, II, III, IV, etc., Arabic numbers 0, 1, 2, 3, 4, etc., and geez numbers such as ፩, ፪, ፫, etc. Removing those numbers from text documents of Ethiopian languages is applied in the handling number phase.

Removing words of other languages: In the Afaan Oromo, Amharic, and Tigrigna text documents, users are mixing words from other languages when writing posts on social media platforms. In this study, to clean the text documents, we removed all non-Afaan Oromo language words from the Afaan Oromo text document, non-Amharic words from the Amharic text, and non-Tigrigna language words from the Tigrigna text document. The text preprocessing pipe line applied in this study indicated in Fig. (3).

Data annotation: Tigrigna, Afaan Oromo, and Amharic experts were employed in the study to label text into hate and non-hate categories. Funds and time were needed for the annotation process, but the annotators were selected based on their proficiency in reading, writing, and comprehending Amharic. To aid the annotators in accurately identifying the content, the annotations contained definitions, rules, and examples of both normal and hateful speech.

Annotation guideline: This study compiles hateful languages in Afaan Oromo, Amharic, and Tigrigna from surveys of Amhara, Oromo and Tigre communities, following hate speech guidelines from the Ethiopian government available at https://www.accessnow.org/cms/assets/uploads/2020/05/Hate-Speech-and-Disinformation-Prevention-and-Suppression-Proclamation.pdf.

The following are a few examples:

- The text delivers dehumanization comments based on the types of clothing the people are wearing and considers that people as uncivilized (ቆምጬ)
- If the message indicates the group of people as one who is moving without command and thinking(መንጋ)
- The term intends to dehumanize or degrade humans because of her/his skin color (garbicha)
- If the intention of the message is against the religion of others (መናፍቅ, ቂጥ አጣቢ)
- If the text comment delivers information that indicates that someone is not Indigenous and unwelcome (qubattoota, ሰፋሪ,)
- The message delivered by comments delivers a message to immigrants (መጤ)
- If the text comments indicate the gunmen of the dictator and oppress the people ነፍጠኛ

With labeling approved by a majority vote and revised guidelines, the study used Fleiss-Kappa Statics to evaluate inter-rater agreement among annotators, obtaining 87% for Amharic, 85% for Tigrigna, and 84% for Afaan Oromo. Few annotation guidelines also indicated in Table (1).

Sample of text data annotation indicated in Table (2) for Amharic, in Table (3) for Afaan Oromo and Table (4) for Tigrigna.

Topic Extraction

In our previous study, we applied topic modeling algorithms such as BERTopic, LDA, LSA, PLSA, and NMF to build a topic extraction model based on Fig. (4).

Adapting and optimizing BERTopic

Ghasiya and Okamura's topic modeling reduces computational complexity in finding related word similarities by clustering lower-dimension approximations (Ghasiya and Okamura, 2021). A topic modeling technique called BERTopic eliminates the need for human judgment and involvement during model training (Defersha *et al*., 2024). It is indicated in Fig. (5).

Table 1: Data annotation guideline

Text comments that criticize the beliefs of other groups or people should be flagged as antagonistic.
If the text comments include insults directed at specific people or organizations
Written texts that make derogatory remarks about specific people or groups ought to be classified as hate speech.
Text comments that contain language that implies the intention to cause pain, destruction, or injury, or that compel people or groups to do something, or not do something, should be considered threats.

Table 2: Sample Amharic text data Annotation by table

Facebook text comments	Label
ጋላ አባሯቸው በጅራፍ ውሻ ጭሮው ተዋወቋት ስለሚገርፏቸው ቆንዶፍ ተበዳግም ሱማሌውን	hate
ቤት አካባቢየ በቀጠሮዋ አይለይምቱሀ ከታሰረችበት ከፍርድ መታሰርም አገናጅቶ እንዳላችሁት ገበርኩልሽ	non_hate

Table 3: Sample Afaan Oromo text data Annotation by table

Facebook Text Comments	Label
haattuu opdo tolfate iss ulfiinaa jalaatuu tiger harreedha hooma rakkin	hate
waaqayyo waaqayyoo hammata waaqa baasaniifii fayyinnis cuuphamuun hammaataadha barsiisu naafii	non_hate

Table 4: Example for Tigrigna text data annotation by table

Facebook text documents	Label
ከምኡ_ንዓባስ_ንሃርሞ_ኣይምዓበናን	hate
ጉጅለ_ግብረሽበራዊ_ሽርክን_ሃሃ	non_hate

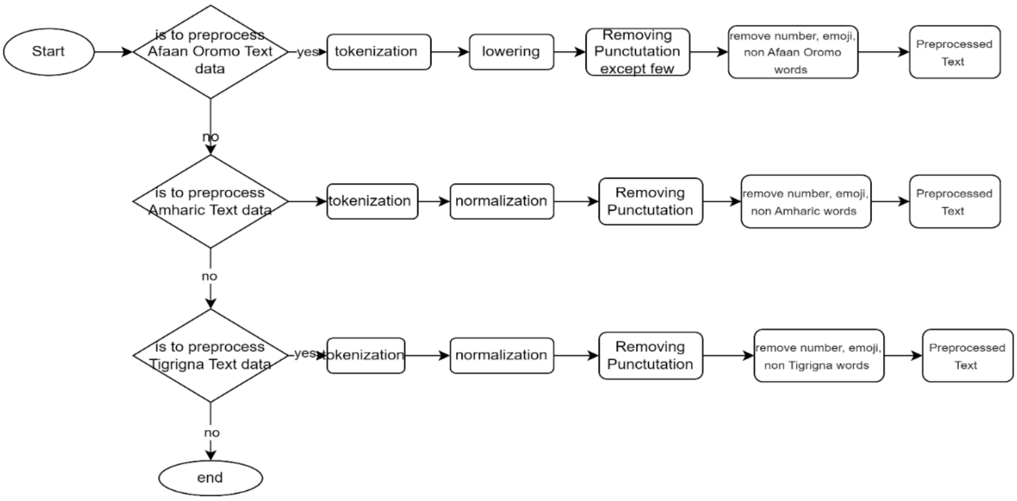

Start
is to preprocess Afaan Oromo Text data
yes
tokenization
lowering
Removing Punctutation except few
remove number, emoji, non Afaan Oromo words
Preprocessed Text
no
is to preprocess Amharic Text data
tokenization
normalization
Removing Punctutation
remove number, emoji, non Amharic words
Preprocessed Text
no
is to preprocess Tigrigna Text data
yes
tokenization
normalization
Removing Punctutation
remove number, emoji, non Tigrigna words
Preprocessed Text
no
end

Fig. 3: Proposed shallow Ethiopian languages' preprocessing pipeline

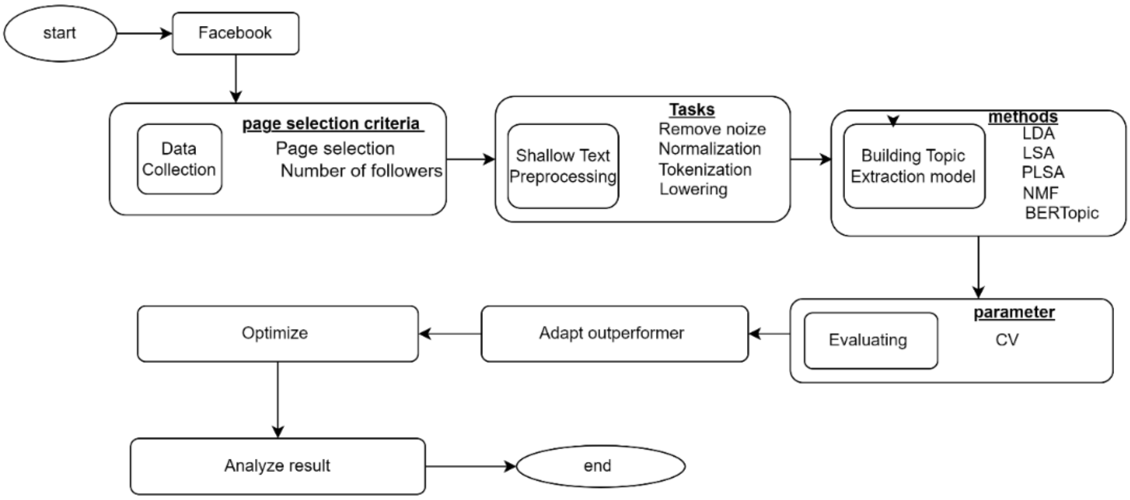


Fig. 4: Proposed framework for exploring the application of Topic modeling in developing a topic extraction model

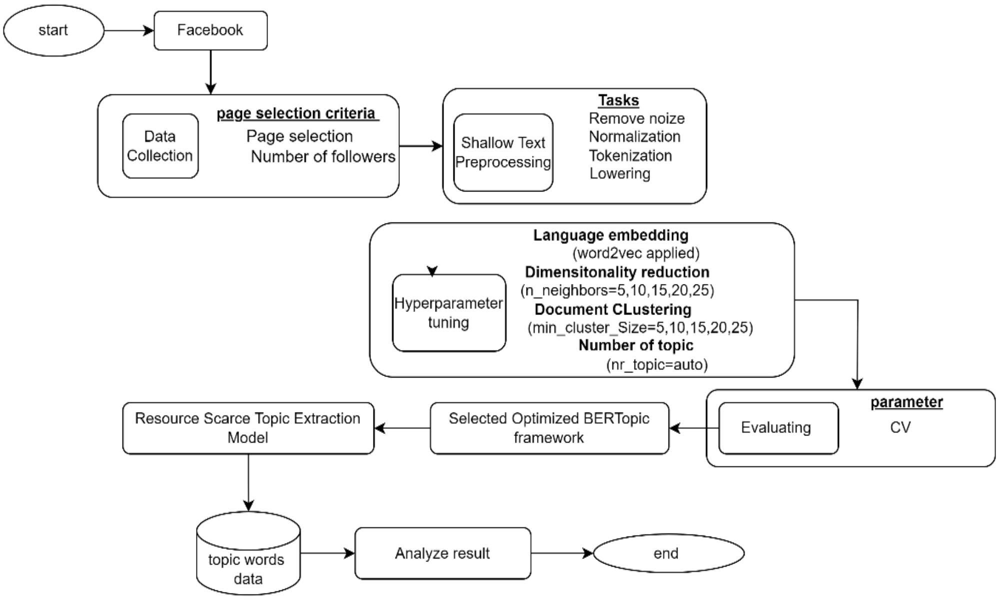


Fig. 5: BERTopic adaption and optimization framework for resource Ethiopian languages

It captures semantic similarity in documents using widely used embeddings such as Hugging Face Transformers and Sentence Transformers (Defersha *et al*., 2024). In our previous work, we employed fine-tuning, a frequent task in natural language processing, to learn patterns and semantics in social media material in Afaan Oromo, Amharic, and Tigrigna to organize enormous text corpora (Defersha *et al*., 2024).

The modified BERTopic algorithm was tested on social media data from Afaan Oromo, Amharic, and Tigrigna, with Afaan Oromo scoring 82.74, Amharic at 87.85% and Tigrigna at 81.79% (Fig. 6).

Term weight schema is numerical statistics that represent how the given term is essential to the given document in the corpus (Truica *et al*., 2016). Because BERTopic used c-TF-IDF as a weight schema mechanism to perform clustering by analyzing the importance of the terms for the given corpus, in this study, c-TF-IDF was applied as a weight schema mechanism.

Defersha *et al*. (2024), identified that the performance of the clustering technique which includes k-means, agglomerative clustering, cuML HDBSCAN, and HDBSCAN improves the accuracy of topic representations (Defersha *et al*., 2024). BERTopic modeling employed HDBSCAN to cluster text documents that do not require the manual setting of several clusters (Moreno-Ortiz, 2024).

Building Ethiopian Languages Transformer Languages Models

Several studies have applied transformer languages to develop hateful content detection models and fake news detection models. Shifath *et al*. (2021), applied the transformer languages models to develop a fake news detection model (Shifath *et al*., 2021), whereas (Mukherjee and Das, 2021; Zia *et al*., 2022; Roy *et al*., 2020) hate speech detection models. The following framework used in Fig. (7) is used to develop Facebook data-based EthioLSocMDMTLM.

BERT: BERT is a transformer-based NLP technique that can produce contextualized embeddings (Ghosh and Senapati, 2022). We applied BERT models on three resource-poor Ethiopian languages including Tigrigna, Afaan Oromo, and Amharic to develop the EthioLan_BERT pretrained languages transformers model.

XLM-Roberta: ROBERTA is a multilingual data extension of BERT, designed for 100 languages, and has been trained on 2.5 TB of data (Ghosh and Senapati, 2022). To train downstream XLM-Roberta models for Afaan Oromo, Tigrigna, and Amharic, we used a topic words-based dataset extracted from Tigrigna, Afaan Oromo, and Amharic.

mBERT: MBERT is a deep learning model that has been pre-trained on Wikipedia to encode collective knowledge of 104 languages, including popular worldwide languages like Hindi, Bengali, and Marathi (Ghosh and Senapati, 2022). We loaded the Afaan Oromo, Amharic, and Tigrigna topic words-based dataset into the mBERT model to develop EthioLan_ mBERT.

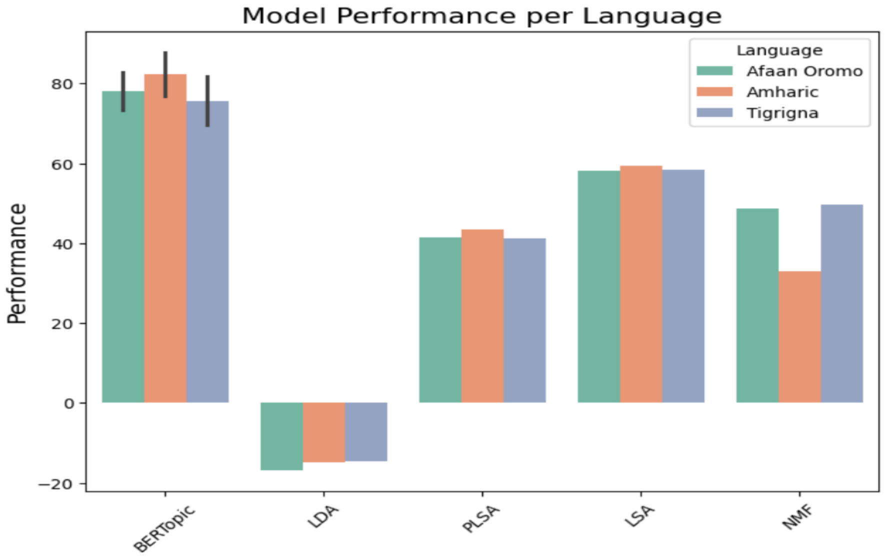


Fig. 6: Topic extraction results performance per language

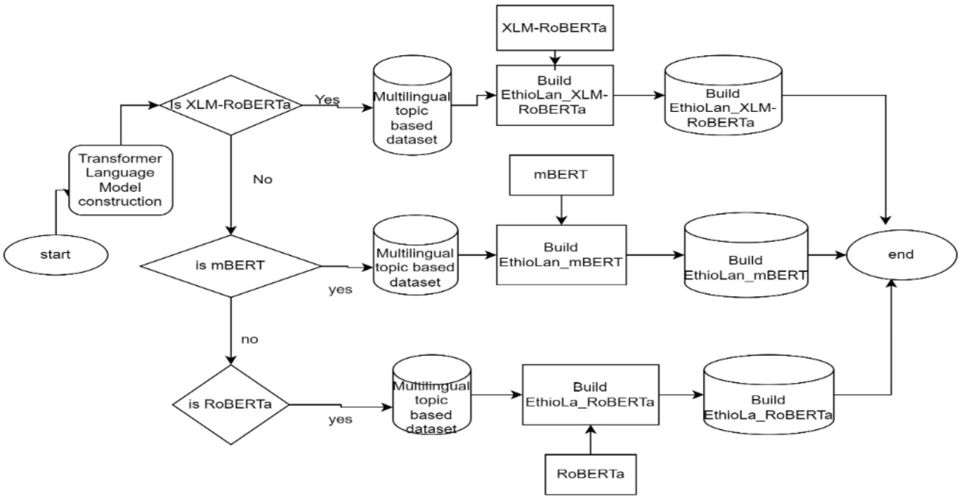

start
Transformer Language Model construction
Is XLM-RoBERTa
Yes
Multilingual topic based dataset
XLM-RoBERTa
Build EthioLan_XLM-RoBERTa
Build EthioLan_XLM-RoBERTa
No
mBERT
is mBERT
yes
Multilingual topic based dataset
Build EthioLan_mBERT
Build EthioLan_mBERT
end
no
is RoBERTa
yes
Multilingual topic based dataset
Build EthioLa_RoBERTa
Build EthioLa_RoBERTa
RoBERTa

Fig. 7: EthioLSocMDMTLM building architecture

Training Multilingual Hateful Content Detection Model

A topic-based dataset and multilingual text comments from our earlier work are used to construct the suggested model (Defersha *et al*., 2024). Using a topic words-based dataset, we created multilingual EthioLSocMDMTLM. We then used deep learning techniques to integrate it into a framework for detecting hateful content, analyzed its accuracy, and selected outperformer algorithms.

Deep Learning Techniques

The following deep learning algorithms were selected to develop the proposed model.

LSTM: Three gates are used by LSTM networks, which are trained using backpropagation, to control information flow and used in speech recognition, time series analysis, and natural language processing.

CNN: Deep learning techniques CNNs applied to classify images, find objects, recognize faces, analyze medical images, natural language processing, speech recognition and perform other computer vision tasks. CNNs are trained by backpropagation and pre-trained models like VGG, ResNet, and Inception, which can be fine-tuned for specific applications or utilized as feature extractors.

Bidirectional Long Short-Term Memory (BiLSTM) Bi-LSTM is a neural network that stores sequence information in both directions, unlike traditional RNN and LSTM models. It allows input to flow in both directions, preserving future and past data and ensuring accuracy by cross-checking both directions. To evaluate the performance of the proposed model F1-score was employed.

Experiments and Results

Dataset

In this study, to create multilingual sociMBa Transformer language models, we employed topic-word-based text comments from 1561, 1044, and 70 Afaan Oromo, Amharic, and Tigrigna languages, respectively. We analyzed 59529 (Afaan Oromo), 45522 (Amharic), and 48882 (Tigrigna) datasets we adopted from (Defersha *et al*., 2024). We used non-hate and hate speech content from the selected dataset to create a dataset. The experiment utilized 80% data for the training model and 20% for testing its performance.

Building Multilingual Hateful EthioLSocMDMTLM

We used Amharic, Tigrigna, and Afaan Oromo topic words from (Defersha *et al*., 2024) to develop EthioLSocMDMTLM such as EthioLan_mBERT, EthioLan_XML-RoBERta, EthioLan_BERT. After three pre-trained multilingual hateful content detection models developed for resource three Ethiopian Languages, three pre-trained Ethiopian languages Transformers language models were integrated into a deep learning framework. Then, the text documents consist of 45522, 59529, and 48882, text documents of Amharic, Afaan Oromo, and Tigrigna collected from our study (Defersha *et al*., 2024) loaded into transformers languages and deep learning models based multilingual hateful content detection framework. From the dataset prepared; 80% was shared as training data and the remaining 20% was shared as testing data.

Results and Discussion

EthioLSoc MDMTLM Result and Discussion

We employed the topics-based dataset prepared by Defersha *et al*. (2024) for three Ethiopian Languages to develop EthioLan_XLM-Roberta, EthioLan_BERT, and EthioLan_mBERT multilingual transformers language models for hateful content detection. The performance of the proposed approach was also tested by constructing a multilingual transformer language model tested with, BiLSTM, LSTM, and CNN were used to develop the proposed model with different embedding dimensions, epochs, batch size, activation, and optimizers. The sections result from the experiments. As indicated in Table (5), the LSTM achieved score values of 81% than others.

Table 5: Performance of the proposed approach for Ethiopian languages

Algorithm + Ethiopian languages multilingual transformers language model	Precision	Recall	F1-score
LSTM+ EthioLan_BERT	0.66	0.67	0.66
CNN + EthioLan_BERT	0.70	0.63	0.69
BiLSTM+ EthioLan_BERT	0.64	0.54	0.61
CNN + EthioLan_XLM-Roberta	075	0.67	0.68
LSTM+ EthioLan_XLM-Roberta	0.59	0.65	0.55
BiLSTM + EthioLan_XLM-Roberta	0.42	0.65	0.51
CNN + EthioLan_mBERT	0.58	0.60	0.53
LSTM+ EthioLan_mBERT	0.82	0.85	0.81
BiLSTM + EthioLan_mBERT	0.50	0.62	0.45

Comparing the Performance of Proposed Approaches

We experimented using the sociMBaM-Transformers model by integrating them into the resource-scarce multilingual Hateful content detection framework. In Ethiopian language transformer models the LSTM+ EthioLan_mBERT outperforms others by scoring f1score 81% which is higher than others.

Multilingual Hateful Content Detection Prototype Evaluation

In this study, LSTM+ EthioLan_mBERT was selected as it outperforms other proposed approaches. Accordingly, the prototype was developed with two options for taking input text data: (1) writing or (2) loading input text into the prototype using the load CSV button. Amharic Facebook text comments "በባንዳነት ተግባር ተሰማርተው ሲወጓት ከከረሙት ታጣቂዎች በላይ በክትክታ ዱላ ሲዋጋላት የሰነበተው ፋኖ የስጋት ምንጭ ሆኖ ይታያል" loaded into the multilingual hateful content detection prototype (Fig. 8) and class predicted as hate correctly.

We also fed Tigrigna Facebook text "በርታልን ኣብ ቀጻሊ ብዙሕ ሰብ ኣብ ጎንኺ ኣሎ እዚ እዩ ዓስበይ ናይዚ ሰናይ ስራሕ" which means "work hard, many people will be on your side in the future" and predicted as non-hate correctly (Fig. 9).

Afaan Oromo's text also loaded "ati nama mit harreedha" which means "you are not a human being; you are a donkey" into the prototype of the model and the model detected correctly as hate (Fig. 10).

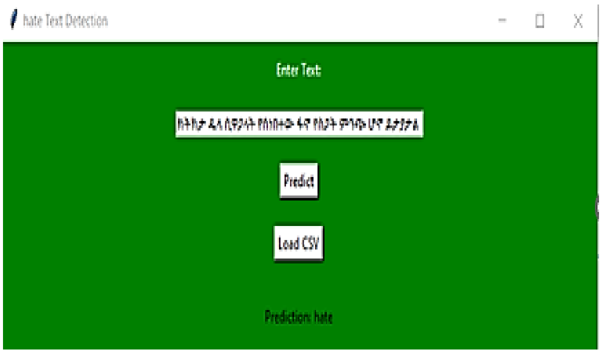
Expert Evaluation of the Multilingual Hateful Content Detector Prototype

The EthioLSocMDMTLM model-based multilingual hateful content detection prototype was evaluated by three experts in language, journalism, law, and psychology. The model was assigned hate classes to Facebook texts and assessed quantitatively. Open-ended questionnaires were prepared to gather opinions on the system's importance in solving real-world problems and its role in combating hateful content on Facebook.

Comparison of the Results with Existing Work

Although the same dataset was not used in the experiments, we also attempted to develop multilingual low-resource Ethiopian Languages and compared our model with hateful content detection models developed in Ethiopian languages using another approach. The table below compares the methods used in our new proposed approach and the existing hate speech detection. LSTM achieved a performance of 81% using EthioLan_mBERT as indicated in Table (6).

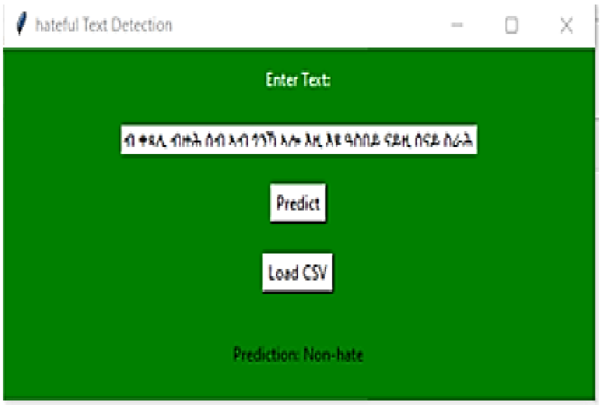
Table (7) shows questionnaire responses and questionnaire results. Respondents agreed on the prototype's efficacy and efficiency, citing its simplicity, test cases, content identification, and resource-saving capabilities. The expert evaluation concluded it saves resources and detects hateful content from Facebook text comments, supporting government officials.



ክትክታ ዱላ ሲዋጋላት የሰነበተው ፋኖ የስጋት ምንጭ ሆኖ ይታያል

Prediction: hate

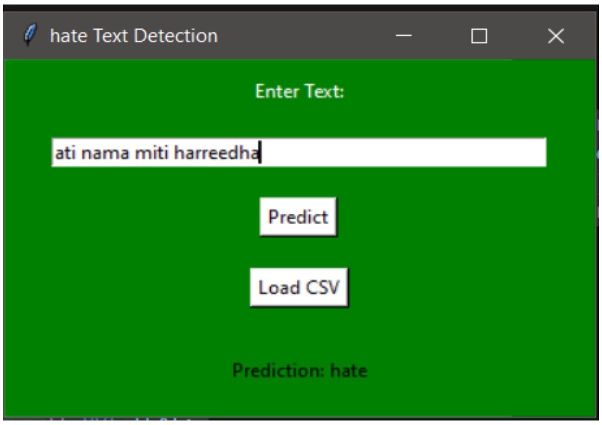
Fig 8: Prototype evaluation on Amharic text



ብ ቀጻሊ ብዙሕ ሰብ ኣብ ጎንኺ ኣሎ እዚ እዩ ዓስበይ ናይዚ ሰናይ ስራሕ

Prediction: Non-hate

Fig. 9: Prototype evaluation on Tigrigna text



ati nama miti harreedha

Prediction: hate

Fig. 10: Prototype evaluation of Afaan Oromo text

Table 6: Comparison of the proposed model with the existing study

Ref.	Languages of dataset	Transformer Languages models	Bilingual/ Multilingual	Algorithms	Features	F1-score (%)
Ababu and Woldeyohannis (2022)	Amharic and Afaan Oromo	No	Bilingual	CNN, LSTM, BiLSTM & GRU	Keras word embedding, word2vec & Fasttext	78.05
Kemal *et al*. (2023)	Amharic and Afaan Oromo	No	Bilingual	LSTM, CNN, GRU, Bi-GRU, BiLSTM+Attention, Bi-GRU+Attention & BiLSTM	word2vec	94.3
Ours	Amharic, Afaan Oromo and Tigrigna	yes	Multilingual	LSTM, CNN & BiLSTM	EthioLan_mBERT, EthioLan_BERT, EthioLan_XLM-Roberta	81

Table 7: The expert-based prototype evaluation result

No	Questions	Response
1	GUI of the prototype is easy to use and understand	Yes
2	The system able to load CSV data for testing	Yes
3	The system allows the user to enter desired input text data for testing	Yes
4	The prototype detects the class of text correctly	Yes
5	The system saves resource	Yes

Limitations of the Study

Although this study has several significances in detecting hateful content for resource-scarce Ethiopian languages, it also follows various limitations. The datasets used to train and evaluate models are limited to three Ethiopian languages and the results may not be generalized to other languages. Despite training transformer language demands a large size of text data, we applied a small size of the dataset in this experiment using only BERT, mBERT, and XLM-RoBERTa to build transformer language models, there be other opportunities to develop lexicons using other approaches and there may be cases where transformer language models, may be effective and efficient. Training transformer language model is computationally intensive and demands high-performance computing machines. In low-resource situations, the problem of fine-tuning domain-specific data is made worse by the absence of infrastructure required for such tasks.

Conclusion

The overall, objectives of this study are to develop a multilingual hateful content detection model for resource-scarce Ethiopian languages by developing and comparing the performance of linguistic resources such as the social media data-based Ethiopian language transformer language model. The develop that linguistic resource we employed topic word-based datasets extracted from Amharic, Tigrigna, and Afaan Oromo and then utilized them to develop a multilingual hateful content detection model using deep-learning techniques such as CNN, LSTM, and BiLSTM. To conduct this study, first, we adopted a topic word-based data set from our previous work. Secondly, we constructed a topics-based multilingual hateful EthioLSocMDMTLM (Ethiolan_mBERT, Ethiolan_BERT, and Ethiolan_XLM-RoBERTa) using three language topic datasets. Thirdly, linguistic resources are integrated into the Ethiopian language's multilingual hateful content detection framework. To develop the proposed model, we used deep learning techniques such as CNN, LSTM, and BiLSTM for Afaan Oromo, Amharic, and Tigrigna. The experiment's results indicated that the transformer model LSTM+EthioLan_mBERT outperforms in developing multilingual hateful content detection for low-resource Ethiopians. Overall, this study demonstrates the efficacy of integrating topic modeling for developing linguistic resources which helps where there is a limitation of computational resources. Experts also evaluated the performance of the prototype through a questionnaire prepared and provided to them. The experts concluded that the system is essential for the Ethiopian government office to minimize the dissemination of hateful content over social media such as Facebook.

Acknowledgment

We appreciate the Editors for the quick response they provided to us and also the reviewers for the constructive comments they have given for the improvement of our paper.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

Naol Bakala Defersha: Ideas; formulation or evolution of overarching research goals and aims with development or design of methodology; creation of models.

Kula KekebaTune: Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.

Solomon Teferra Abate: Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs. Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.

Ethics

Authors should address any ethical issues that may arise after the publication of this manuscript.

Reference

Ababa, A., Defersha, N. B., & Tune, K. K. (2021). Detection of Hate Speech Text in Afan Oromo Social Media Using Machine Learning Approach. *Indian Journal of Science and Technology*, *14*(31), 2567–2578. https://doi.org/10.17485/ijst/v14i31.1019

Ababu, T. M., & Woldeyohannis, M. M. (2022). Afaan Oromo Hate Speech Detection and Classification on Social Media. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6612–6619.

Alshalan, R., Al-Khalifa, Hend, Alsaeed, D., Al-Baity, H., & Alshalan, S. (2020). Detection of Hate Speech in COVID-19–Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach. *Journal of Medical Internet Research*, *22*(12), e22609. https://doi.org/10.2196/22609

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep Learning Models for Multilingual Hate Speech Detection. In *arXiv:2004.06465v3*. https://doi.org/10.48550/arXiv.2004.06465

Ayele, A. A., Jalew, E. A., Ali, Adem Chanie, Yimam, S. M., & Biemann, C. (2024). Exploring Boundaries and Intensities in Offensive and Hate Speech: Unveiling the Complex Spectrum of Social Media Discourse. In *arXiv:2404.12042v1*. https://doi.org/10.48550/arXiv.2404.12042

Ayele, A. A., Yimam, S. M., Belay, T. D., & Biemann, C. (2023). Exploring Amharic Hate Speech Data Collection and Classification Approaches. *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, 49–59. https://doi.org/10.26615/978-954-452-092-2_006

Bahador, B. (2023). Monitoring Hate Speech and The Limits of Current Definition. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (Vol. 12, pp. 291–298). Open Access Repository. https://doi.org/10.48541/dcr.v12.17

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. In *arXiv:1903.10676v3*. https://doi.org/10.48550/arXiv.1903.10676

Biradar, S., Saumya, S., & Chauhan, A. (2022). Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach. *Social Network Analysis and Mining*, *12*(1), 87. https://doi.org/10.1007/s13278-022-00920-w

Calderón, C. A., de la Vega, G., & Herrero, D. B. (2020). Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain. *Social Sciences*, *9*(11), 188. https://doi.org/10.3390/socsci9110188

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-Training Text Encoders as Discriminators Rather Than Generators. *8th International Conference on Learning Representations, ICLR 2020*, 1–18.

Defersha, N. B., Abawajy, J., & Kekeba, K. (2022). Deep Learning based Multilabel Hateful Speech Text Comments Recognition and Classification Model for Resource Scarce Ethiopian Language: The case of Afaan Oromo. *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*, 1–11. https://doi.org/10.1109/ccet56606.2022.10080837

Defersha, N. B., Kekeba, K., & Kaliyaperumal, K. (2021). Tuning Hyperparameters of Machine Learning Methods for Afan Oromo Hate Speech Text Detection for Social Media. *{2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, 596–604. https://doi.org/10.1109/iccct53315.2021.9711850

Defersha, N. B., Tune, K. K., & Abate, S. T. (2024). Adapting Outperformer from Topic Modeling Methods for Topic Extraction and Analysis: The Case of Afaan Oromo, Amharic and Tigrigna Facebook Text Comments. *International Journal of Advanced Computer Science and Applications*, *15*(3). https://doi.org/10.14569/ijacsa.2024.0150391

Faris, H., Aljarah, I., Habib, M., & Castillo, P. (2020). Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context. *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, 453–460. https://doi.org/10.5220/0008954004530460

Ganfure, G. O. (2022). Comparative analysis of deep learning based Afaan Oromo hate speech detection. *Journal of Big Data*, *9*(1), 76. https://doi.org/10.1186/s40537-022-00628-w

Ghasiya, P., & Okamura, K. (2021). Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access*, *9*, 36645–36656. https://doi.org/10.1109/access.2021.3062875

Ghosh, K., & Senapati, A. (2022). Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, 853–865.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All You Need is “Love”: Evading Hate Speech Detection. *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2–12. https://doi.org/10.1145/3270101.3270103

Guo, Y., & Sarker, A. (2023). SocBERT: A Pretrained Model for Social Media Text. *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, 45–52. https://doi.org/10.18653/v1/2023.insights-1.5

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 174–181. https://doi.org/10.3115/976909.979640

Ibrahim, N., Mulford, F., Lawrence, M., & Batista-navarro, R. (2024). Resources for Annotating Hate Speech in Social Media Platforms Used in Ethiopia: A Novel Lexicon and Labelling Scheme. *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, 115–123.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kasule, A., Mutebi, B., Balunywa, A., Makubuya, R., & Kyeyune, R. (2023). Readiness of Graduates from Ugandan Higher Institutions of Learning for Work in the Fourth Industrial Revolution. *The Uganda Higher Education Review*, *11*(2), 69–81. https://doi.org/10.58653/nche.v11i2.6

Kemal, B. S., Abebe, T. U., Pendem, G. K., Krishna, T. G., & Gemeda, K. A. (2023). Bilingual Social Media Text Hate Speech Detection for Afaan Oromo and Amharic Languages Using Deep Learning. *Journal of Namibian Studies: History Politics Culture*, *34*, 250–281. https://doi.org/10.59670/jns.v34i.1446

Korre, K., Muti, A., & Barrón-Cedeño, A. (2024). The Challenges of Creating a Parallel Multilingual Hate Speech Corpus: An Exploration. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 15842–15853.

Lee, N., Jung, C., & Oh, A. (2023). Hate Speech Classifiers are Culturally Insensitive. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 35–46. https://doi.org/10.18653/v1/2023.c3nlp-1.5

Liu, S., & Forss, T. (2015). New Classification Models for Detecting Hate and Violence Web Content. *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 487–495. https://doi.org/10.5220/0005636704870495

Lu, J. (2023). Hate Speech Detection Based on Multiple Machine Learning Algorithms. *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, 244–252. https://doi.org/10.2991/978-94-6463-300-9_25

Macias, C., Soto, M., Alcántara, T., & Calvo, H. (2023). Impact of Text Preprocessing and Feature Selection on Hate Speech Detection in Online Messages Towards the LGBTQ+ Community in Mexico. *CEUR Workshop Proceedings*, 1–10.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745

McGowan, L., Gaston, E., Day, A., & Kern, L. (2024). Hate Speech Case Study. In *Centre for Policy Research*.

Melat, F. A. (2022). *Hate Speech Detection for Amharic Language on Facebook Using Deep Learning*. Bahir Dar Institute of Technology.

Moreno-Ortiz, A. (2024). COVID-19 Corpora. In *Making Sense of Large Social Media Corpora: Keywords, Topics, Sentiment* and *Hashtags in the Coronavirus Twitter Corpus* (pp. 19–30). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-52719-7_2

Mossie, Z., & Wang, J.-H. (2018). Social Network Hate Speech Detection for Amharic Language. *Computer Science \& Information Technology*, 41–55. https://doi.org/10.5121/csit.2018.80604

Mukherjee, S., & Das, S. (2021). Application of Transformer-Based Language Models to Detect Hate Speech in Social Media. *Journal of Computational and Cognitive Engineering*, *2*(4), 278–286. https://doi.org/10.47852/bonviewjcce2022010102

Naidu, T. A., & Kumar, S. (2021). Hate Speech Detection Using Multi-Channel Convolutional Neural Network. *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 908–912. https://doi.org/10.1109/icac3n53548.2021.9725696

Ogueji, K., Zhu, Y., & Lin, J. (2021). Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 116–126. https://doi.org/10.18653/v1/2021.mrl-1.11

Pradanna, S. A., & Abdulkarim, A. (2023). The role of social media in strengthening multicultural tolerance among digital citizenship. *Proceeding of International Conference on Innovations in Social Sciences Education and Engineering*, 1–11.

Roy, S. G., Narayan, U., Raha, T., Abid, Z., & Varma, V. (2020). Leveraging multilingual transformers for hate speech detection. In *arXiv:2101.03207v1*. https://doi.org/10.48550/arXiv.2101.03207

Shifath, S. M. S.-U.-R., Khan, M. F., & Islam, Md. S. (2021). A transformer based approach for fighting COVID-19 fake news. In *arXiv:2101.12027v1*. https://doi.org/10.48550/arXiv.2101.12027

Tesfaye, S. G., & Tune, K. (2020). Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network. In *Research Square*. https://doi.org/10.21203/rs.3.rs-114533/v1

Teshome, W. (2013). Designing Amharic Definitive Question Answering. In *Addis Ababa University College of Natural Sciences School of Information Science*.

Tontodimamma, A., Nissi, E., Sarra, A., & Fontanella, L. (2021). Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, *126*(1), 157–179. https://doi.org/10.1007/s11192-020-03737-6

Truică, C.-O., & Apostol, E.-S. (2022). MisRoBÆRTa: Transformers versus Misinformation. *Mathematics*, *10*(4), 569. https://doi.org/10.3390/math10040569

Truică, C.-O., & Apostol, E.-S. (2023). It's All in the Embedding! Fake News Detection Using Document Embeddings. *Mathematics*, *11*(3), 508. https://doi.org/10.3390/math11030508

Truica, C.-O., Radulescu, F., & Boicea, A. (2016). Comparing Different Term Weighting Schemas for Topic Modeling. *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 307–310. https://doi.org/10.1109/synasc.2016.055

UNESCO/Ipsos. (2023). *Survey on the impact of online disinformation and hate speech*. Ipsos. https://www.unesco.org/sites/default/files/medias/fichiers/2023/11/unesco_ipsos_survey.pdf?hub=71542

Unlu, A., Truong, S., & Kotonen, T. (2024). Mapping the terrain of hate: Identifying and analyzing online communities and political parties engaged in hate speech against Muslims and LGBTQ+ communities. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-024-00571-4

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582. https://doi.org/10.1016/j.is.2020.101582

Wang, C.-C., Day, M.-Y., & Wu, C.-L. (2022). Political Hate Speech Detection and Lexicon Building: A Study in Taiwan. *IEEE Access*, *10*, 44337–44346. https://doi.org/10.1109/access.2022.3160712

Weldemariam, B. (2022). *a Posts and Comments in Tigrigna language*. St. Mary's University.

Yimam, S. M., Ayele, Abinew Ali, & Biemann, C. (2019). Analysis of the ethiopic twitter dataset for abusive speech in amharic. In *arXiv:1912.04419v1*. https://doi.org/10.48550/arXiv.1912.04419

Zhang, Z., Fang, M., Chen, L., & Namazi Rad, M. R. (2022). Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3886–3893. https://doi.org/10.18653/v1/2022.naacl-main.285

Zia, H. B., Castro, I., Zubiaga, A., & Tyson, G. (2022). Improving Zero-Shot Cross-Lingual Hate Speech Detection with Pseudo-Label Fine-Tuning of Transformer Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *16*, 1435–1439. https://doi.org/10.1609/icwsm.v16i1.19402

Zufall, F., Hamacher, M., Kloppenborg, K., & Zesch, T. (2022). A Legal Approach to Hate Speech – Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task. *Proceedings of the Natural Legal Language Processing Workshop 2022*, 53–64. https://doi.org/10.18653/v1/2022.nllp-1.5