Detecting Hate Speech in Amharic Using Multimodal Analysis of Social Media Memes

Melese Ayichlie Jigar[1], **Abinew Ali Ayele**[1,2],**Seid Muhie Yimam**[2], **Chris Biemann**[2]

[1] Bahir Dar University, Ethiopia, [2] Universität Hamburg, Germany

Abstract

In contemporary society, the proliferation of hate speech is increasingly prevalent across various social media platforms, with a notable trend of incorporating memes to amplify its visual impact and reach. The conventional text-based detection approaches frequently fail to address the complexities introduced by memes, thereby aggravating the challenges, particularly in low-resource languages such as Amharic. We develop Amharic meme hate speech detection models using 2,000 memes collected from Facebook, Twitter, and Telegram over four months. We employ native Amharic speakers to annotate each meme using a web-based tool, yielding a Fleiss' kappa score of 0.50. We utilize different feature extraction techniques, namely VGG16 for images and word2Vec for textual content, and build unimodal and multimodal models such as LSTM, BiLSTM, and CNN. The BiLSTM model shows the best performance, achieving 63% accuracy for text and 75% for multimodal features. In image-only experiments, the CNN model achieves 69% in accuracy. Multimodal models demonstrate superior performance in detecting Amharic hate speech in memes, showcasing their potential to address the unique challenges posed by meme-based hate speech on social media.

Keywords: Multimodal, Meme, LSTM, BiLSTM, CNN

1. Introduction

Currently, there are 5.44 billion mobile phone users worldwide, and the number of active social media users has reached 4.76 billion, which is equivalent to 60% of the global population. During this period, the addition of new users was relatively modest, with only 137 million joining, resulting in an annual growth rate of a mere 3%, as reported by Kemp (2023). According to this analysis, Ethiopia had 20.86 million internet users in January 2023, indicating a growth of 520 thousand users from 2022, which is around 2.6%.

Social media exerts influence over a nation's social, economic, and political dimensions. It facilitates swift digital information sharing among individuals. However, it also has adverse impacts when employed for disseminating aggressive, hateful, or threatening content online (Mathew et al., 2021; Ayele et al., 2022b). Hate speech encompasses any form of communication that disparages an individual or a group because of their color, race, ethnicity, sexual orientation, gender, nationality, religion, or other qualities (Zhou et al., 2020). Hate speech can spread over social media platforms in various forms such as text, image, audio, and video. Despite hate speech spreading in various forms on social media, the majority of research works on hate speech detection tasks focus on developing unimodal, especially text-based models. Also, most of the multimodal hate speech research focuses on English and some European languages (Rana and Jha, 2022; Pramanick et al., 2021; Corazza et al., 2018; Perifanos and Goutsos, 2021; Karim et al., 2023) while low-resource languages such as Amharic received less attention.

Multimodal models that combine both text and image features are required to accurately detect hate speech in online spaces. In this paper, we address the following research questions:

1. Which multi-modal models perform better for identifying hate speech in the Amharic meme dataset?
2. What features are influential in developing a predictive multimodal hate speech model for Amharic?

This paper presents several significant contributions, which encompass, but are not limited to, the following key aspects:

1. We have presented a benchmark dataset of 2k Amharic memes dataset collected from Facebook, Twitter, and Telegram.
2. We have developed an annotation tool called **HateMemAnno**, specifically designed for annotating memes.
3. We have developed a multi-modal hate speech detection model from the Amharic memes datasets.
4. We have thoroughly examined and contrasted the effectiveness of unimodal and multimodal detection methods.
5. We have investigated the challenges of Multimodal Amharic memes and explored future research opportunities in this field.

TRAC-2024 Workshop, pages 85–95

The remainder of the paper is organized as follows. The related works are presented in Section 2. Section 3 provided a detailed description of the Amharic language. The data collection and annotation procedures are described in Section 4. Section 5 presented the experimental details. In Section 6, we presented the results and discussion. In Section 7, we provided the error analysis of the experiment. Finally, Section 8 provided a summary of the findings and outlined avenues for future work.

2. Related Works

The meaning of hate speech varies across different sources. This variation is due to the prevailing societal norms, individual perspectives, contextual factors, and collective viewpoints (Madukwe et al., 2020; Yimam et al., 2019). Hate speech is a complex problem that is intertwined with the interactions among diverse social groups. It flourishes through the intentional manipulation of language's vagueness, making it challenging to detect easily (Zufall et al., 2022; Ayele et al., 2023b). Social media provides users the opportunity to conceal their genuine identities by operating in the shelter of digital screens and anonymous usernames (Bran and Hulin, 2023; Ayele et al., 2023a). The cover of anonymity grants users the ability to disseminate hate speech without facing immediate consequences, which intensifies the difficulty of addressing hate speech in the digital era (Davidson et al., 2019; Mathew et al., 2021; Ayele et al., 2022b).

For the last decade, a lot of research has been carried out to address the detection of hate speech in social media. Most of these attempts were mainly focused on detecting hate speech by employing unimodal approaches that take features only from one input, such as text, image, or audio (Suryawanshi et al., 2020). Hate speech detection research has primarily centered on textual data sources, and there has been a lesser emphasis on considering multimodal parameters. This gap is especially critical when it comes to low-resource languages. Among the research for Amharic hate speech in this regard includes the work by Ayele et al. (2022b); Abebaw et al. (2022); Tesfaye and Kakeba (2020); Ayele et al. (2023b); Defersha and Tune (2021); Mossie and Wang (2020), which focused on text-based model building.

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. They apply fastText (Joulin et al., 2017) word embedding approaches to detect hate speech from the extracted texts by employing unimodal detection approaches. Their approach solely relies on the extracted text from memes, neglecting the image component, potentially resulting in an incomplete interpretation of the meme's intended message.

In addition, 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. They employed Word2Vec and MFCC to extract textual and audio features, respectively, and applied the Google Speech-to-Text API to transcribe audio speech into text scripts.

Studies on English and some other resource-rich languages explored image datasets and utilized computer vision techniques to identify images that contain discriminatory, offensive, or harmful content and employed multi-modal models by combining textual and image features (Arango et al., 2022; Cao et al., 2022; Gomez et al., 2020; Perifanos and Goutsos, 2021; Bhat et al., 2023; Velioglu and Rose, 2020; Suryawanshi et al., 2020; Kiela et al., 2020).

Spreading hate speech using memes is becoming a common phenomenon on social media platforms that require hate speech detection tasks to employ concatenated features of memes, both the image features and extracted text features (Schmidt and Wiegand, 2017). Therefore, the aim of our study is to bridge this gap by employing multimodal hate speech detection models that utilize concatenated features from images and texts.

3. Amharic Language

Amharic is the working language of the Federal Democratic Republic of Ethiopia that holds significant linguistic and cultural importance (Woldemariam, 2020). It is the second most widely spoken Semitic language worldwide after Arabic (Woldemariam, 2020; Mossie and Wang, 2018). While Amharic serves as a working language in various regional states in Ethiopia (Debele and Woldeyohannis, 2022), it has limited language processing tools and remains low-resourced.

The writing system of Amharic, known as "Fidäl", is derived from the Ge'ez alphabet. It consists of 275 alphabets, including 34 consonants and six characters formed from vowel and consonant combinations. Amharic lacks capitalization and has its own unique script (Gezmu et al., 2018). The language is characterized by its distinct orthographic features, including numbers, punctuation marks, and other symbols (Belay et al., 2021).

Amharic poses challenges for researchers and NLP practitioners due to its morphological complexity and highly inflected languages (Yimam, 1999). Moreover, the redundancy of characters in the language and the various methods of rep-

resenting the same sound add further complexity to the identification of hate speech (Belay et al., 2021).

4. Data Collection and Annotation

This section provides a brief overview of the data sources, data collection techniques, data annotation tools, and data annotation procedures.

4.1. Data Source

The datasets were collected from three widely used social media platforms in Ethiopia, namely Telegram, Twitter, and Facebook. We have created a Telegram group called ጥላቻ ንግግሮች የሆኑ ምስሎችን መሰብሰቢያ ገጽ - 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 following group accounts that have more than 100k followers, including ሀላል (Halal) memes Facebook, ሀበሻን (Habeshan) Telegram memes , ሀበሻን (Habeshan) Facebook memes, ግቢ (Gibi) Telegram memes, ፈገግታ (Fegegita) Facebook memes, እገር ኳስ (Egir Kuwas) Facebook memes, ሸገር (Sheger) Facebook meme, ፈታ (Feta) Facebook meme, አዝግ (Azig) Facebook meme, ኢትዮ (Ethio) Facebook meme etc. Moreover, we carefully chose several public pages by considering factors such as the number of members, the language used, and the frequency of news or trending discussions pertaining to politics, ethnicity, religion, and gender. 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 1.

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

Table 1: Distribution of collected memes from different social media.

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

4.2. Annotation Tool

Due to the lack of access to meme annotation tools, we took the initiative to create a web-based annotation tool called **HateMemAnno**, tailored for labeling Amharic meme hate speech content. The annotation tool offers an interface for annotators and includes an admin dashboard with a dataset repository or database. The graphical interface of the annotation tool is presented in Figure 1. After uploading the dataset, the system administrator assigns annotators and authorizes the necessary privileges for annotation. Annotators are provided with annotation guidelines integrated into the tool before commencing the task. The annotation tool presents one meme at a time and permits annotators to review and adjust previous annotations if needed.

Figure 1: Mobile interface for **HateMemAnno** depicting a meme targeting individuals based on their personality, particularly harassing females.

Annotators received training through practical sample annotations and were given detailed explanations of the annotation guidelines before their involvement in the main annotation task. The dataset of 2k memes was annotated in four separate batches, each containing 500 memes.

Each meme underwent annotation by three native Amharic speakers, classifying them into binary categories of **hate** or **non-hate**, resulting in a Cohen's kappa score of 0.50 for inter-annotator agreement. A majority voting scheme was utilized to establish the definitive gold labels. As shown in Table 2, out of the 2k annotated memes, 919 were labeled as **hate** while 1,088 were labeled as **non-hate**.

Batch	Annotator	HS	NHS
Batch 1	Annotator 1	289	211
	Annotator 2	299	301
	Annotator 3	373	127
	Majority Voting	307	193
Batch 2	Annotator 1	321	179
	Annotator 2	332	168
	Annotator 3	351	149
	Majority Voting	319	181
Batch 3	Annotator 1	163	337
	Annotator 2	170	330
	Annotator 3	140	360
	Majority Voting	127	373
Batch 4	Annotator 1	181	326
	Annotator 2	163	344
	Annotator 3	168	339
	Majority voting	166	341
Total	Majority Voting	919	1088

Table 2: Summary of annotated dataset statistics: **HS** column indicates hate speech labels, while **NHS** corresponds to non-hate speech.

5. Experimentation

This section presents the preprocessing methodologies and classification techniques employed in our research. It encompasses the data preparation steps, covering text and image preprocessing, and explained the array of machine learning algorithms and models built for the detection of hate speech within Amharic memes.

5.1. Optical Character Recognition

We employed Tesseract, 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. Following these preprocessing steps, text extraction from the preprocessed memes was conducted using Tesseract.

We retain Amharic sentences with mixed English content to account for users who frequently switch between languages in their message compositions. This approach prevents unintended changes in meaning that might occur if we were to remove English content from mixed sentences. For instance, the meme **GENOCIDERS ዕብዕብ**, which translates to "a group of genociders," would lose its intended meaning if we removed the English term "GENOCIDERS." Instead, we employed Python language detection and translation libraries to identify and translate mixed English terms into their corresponding Amharic equivalents.

The meme images are standardized to uniform dimensions, and their pixel values are rescaled to a range of 0 to 1. Additionally, data augmentation techniques are employed to mitigate the challenges posed by limited training data and to alleviate overfitting concerns.

To facilitate effective model training and testing, it is imperative to preprocess the text extracted from the memes into an appropriate format. This text preprocessing encompasses several steps, such as dataset cleaning, normalization, translating specific English words into their Amharic counterparts, expanding abbreviations, eliminating stop words, and tokenization.

5.2. Feature Extraction

We utilized word embedding techniques to process the textual data, while the pre-trained VGG16 was employed for the extraction of image features as depicted in Figure 2. VGG16, a convolutional neural network architecture, has been extensively trained on a substantial image dataset, endowing 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 them to serve as input for our model.

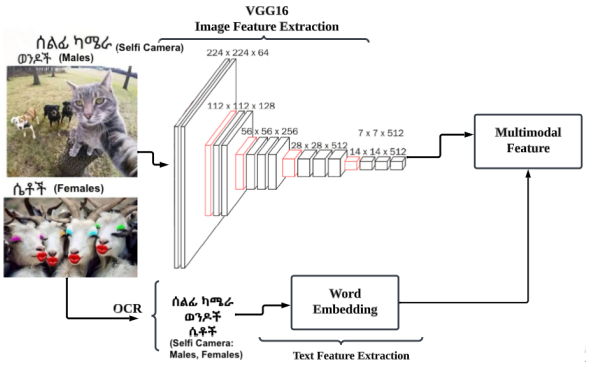

Figure 2: Text and image features concatenation.

5.3. Classification Models

We leveraged several deep-learning algorithms, including LSTM, BiLSTM, and CNN, selected for their 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). LSTM and BiLSTM have demonstrated their effectiveness in hate speech detection from textual data, mainly due to their capacity to capture contextual information and temporal dependencies between words. In contrast, CNN has exhibited superior performance in the detection of hate speech within images. This is attributed to its capability to extract spatial features and intricate patterns inherent in image data, rendering it as a robust choice for this specific task.

Unimodal Textual Experiments

We implemented three distinct deep learning models - LSTM, BiLSTM, and CNN - for the purpose of detecting hate speech independently from unimodal textual or image inputs. In this section, we delve into the specifics of our approach for unimodal Amharic hate speech detection, concentrating on three deep learning techniques.

This experiment was designed to assess the model's proficiency in identifying hate speech solely based on the text content within memes. Given the intricate and subjective nature of hate speech, pattern recognition presented a significant challenge. To address this, we have developed a Keras deep learning model incorporating both the *BatchNormalization layer* and *Dropout layer*. These components play a pivotal role in *normalizing activations* from previous layers, thus substantially mitigating **overfitting** and enhancing the stability of the learning process.

This textual experiment was conducted to ascertain the extent to which text contributes to meme-based hate speech detection. The outcomes of this experiment, detailing the accuracy of each algorithm, are summarized in Table 3.

	Parameters				
Dropout		0.10	0.10	0.20	0.50
Epochs		32	64	32	32
Batch		32	32	32	64
BiLSTM	acc	62%	62%	**63%**	61%
LSTM	acc	61%	62%	62%	60%
CNN	acc	57%	58%	57%	56%

Table 3: Hyperparameters and performance measures for text-based unimodal experiments.

Unimodal Image Experiments

After obtaining features from Amharic memes through the VGG16 model, the image data undergoes a similar hate speech detection process as the textual dataset. In this image-based analysis, an input shape of **(7, 7, 512)** is utilized, followed by a dense layer consisting of 64 neurons. To enhance model performance and mitigate overfitting, *ReLU activation* is applied, complemented by batch normalization and dropout techniques. These additional layers normalize preceding layer activations and reduce the risk of overfitting, ensuring more stable learning. The final classification is executed using the softmax activation function. For a comprehensive overview of the outcomes derived from the unimodal image dataset, please refer to Table 4.

	Parameters				
Dropout		0.10	0.10	0.20	0.50
Epochs		32	64	32	32
Batch		32	32	32	64
BiLSTM	acc	62%	65%	64%	62%
LSTM	acc	63%	62%	64%	65%
CNN	acc	67%	**69%**	67%	66.6%

Table 4: Hyperparameters and performance measures for image-based unimodal experiments

Multimodal Model Experiments

We utilize the embedding matrix feature vectors obtained from both the textual data and VGG16 image features, combining them within the model's input layer. Word2vec is utilized from (Yimam et al., 2021) to properly build the required feature vectors for the textual model. This fusion of features enables us to employ a multimodal training strategy for our model, harnessing the power of both textual and image information to enhance its overall performance and capabilities. This multimodal approach provides the opportunity to capture more complex relationships between the various modalities and facilitates improved identification and classification of hateful content. The results of this multimodal approach experimentation can be seen in Table 5.

Figure 3 presents the confusion matrix of the BiLSTM model, as described in Table 5, which achieved the best performance.

6. Results and Discussion

In this section, we provide a comprehensive overview of the results obtained from our experiments, which encompass both unimodal and multimodal approaches. These experiments were de-

	Parameters				
Dropout		0.10	0.10	0.20	0.50
Epochs		32	64	32	32
Batch		32	32	32	64
LSTM	acc	71%	71%	69%	68%
BiLSTM	acc	73%	**75%**	72%	68%
CNN	acc	68%	69%	69%	71%

Table 5: Hyperparameters and performance measures for multimodal experiments.

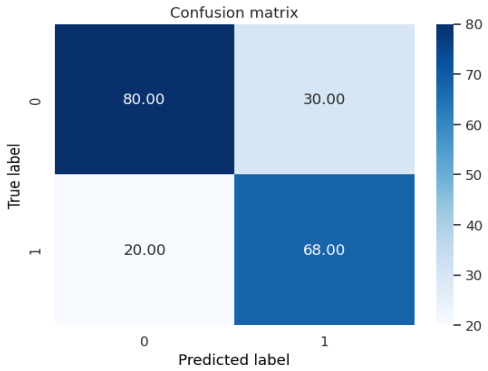

Figure 3: Confusion matrix of multimodal model results using BiLSTM.

signed to address the challenge of hate speech detection in the Amharic meme dataset, and we employed three distinct deep learning algorithms: LSTM, BiLSTM, and CNN.

The experiments were structured into three distinct categories, each focusing on a specific modality: Textual, Image, and Multimodal models. The primary objective of these experiments was to evaluate the effectiveness of these deep learning algorithms in identifying hate speech within the Amharic meme dataset. By systematically examining the performance of each model under these different modalities, we aimed to gain insights into their strengths and weaknesses in handling the unique challenges presented by meme-based hate speech detection.

To ensure the effectiveness of our model, we fine-tuned the models with several hyperparameters configurations. These included configuring the batch size, dropout rate, and embedding dimensions as can be seen in Tables 3, 4, and 5. We set the embedding dimensions to 300. For the loss function, we utilized **binary cross-entropy**, and we specified the number of training epochs that range from 32 to 64. Additionally, we employed the **Adam optimizer** with a **learning rate** of 0.001.

To evaluate the performance of our model, we employed a range of metrics, including Precision, Recall, F1 scores, and accuracy. These metrics provided a comprehensive assessment of the model's ability to correctly classify memes as hate or non-hate, allowing us to gauge its effectiveness in hate speech detection within the Amharic meme dataset.

As depicted in Table 3, our experimental results revealed that the BiLSTM model outperformed both the LSTM and Convolutional Neural Network (CNN) models in terms of accuracy. Specifically, the BiLSTM achieved an impressive accuracy rate of 63%, surpassing the LSTM, which achieved an accuracy rate of 62%, and the CNN, which achieved an accuracy rate of 57%.

The better performance of the BiLSTM model can be attributed to its unique ability to analyze sequential data in both forward and backward directions. This bidirectional processing capability allows the BiLSTM model to capture deeper contextual information from the input data. In the context of our hate speech detection task, this deeper contextual understanding proved to be advantageous in identifying and classifying hateful content within Amharic memes. Consequently, the BiLSTM emerged as the most effective choice among the three deep learning models, showcasing its potential for improving the accuracy of hate speech detection in meme-based datasets.

Our dataset exhibits considerable variability in the lengths of the textual sequences it contains, encompassing sequences that range from very short, consisting of a single word, to longer phrases. To illustrate this diversity, it is important to note that within our dataset, there are 273 instances with sequences of less than two words. Among these instances, a significant portion, precisely 130 of them, consist of only a single word.

These single-word sentences exemplify the brevity and conciseness found in our dataset. Some illustrative examples of these single-word sentences include words and phrases such as **ጅቦች፣ ብአዴን፣ ፍኖ፣ ያዘዋል፣ ትግራይ፣ አፍርሳት፣ (Hyenas, ANDM, Fano, Yazewal, Tigray, Break her)** and **ፍትህ (Justice)**. This diversity in the length of textual sequences poses a unique challenge for natural language processing tasks, as the model must effectively process and understand both very short and longer textual inputs to accurately classify hate speech within Amharic memes.

In the context of the image-based experiment, CNN outperformed LSTM and BiLSTM in terms of accuracy (see Table 4). CNN achieved an accuracy of 69%, surpassing BiLSTM with an accuracy of 65% and LSTM with an accuracy of 62%. This performance difference can be attributed to CNN's inherent strength in extracting features from two-dimensional data, especially images. CNNs are specifically designed to work well with 2D data, making them highly effective in image-based tasks. Conversely, LSTM and BiL-

STM models excel in scenarios involving sequential and time-dependent datasets. The evaluation of the multimodal experiment, as presented in Table 5, involved testing various parameters to identify the configuration that resulted in the highest accuracy. Significantly, the BiLSTM model outperformed both the LSTM and CNN models, achieving a testing accuracy of 75%. In contrast, both CNN and LSTM achieved an accuracy of 71% each. The superior performance of BiLSTM in this context can be attributed to its unique characteristics. BiLSTM can capture both forward and backward dependencies in the input data, which allows it to consider contextual information from both directions. Additionally, BiLSTM can dynamically adjust the size of its hidden layer to match the length of the input text sequences, providing flexibility in handling varying text lengths. These qualities make BiLSTM particularly effective in capturing complex relationships within multimodal data, resulting in the highest accuracy among the tested models in the multimodal experiment. The findings of our study indicate that a multimodal model outperforms unimodal models, primarily due to the synergistic interaction between text and image features. The utilization of multiple modalities leads to improved accuracy in the detection of hate speech.

7. Error Analysis of the Experiment

To evaluate the unimodal and multimodal models' performance, we assessed using the golden labels to identify any inconsistencies. During this evaluation, we encountered inconsistencies in testing accuracy with our proposed model. This challenge was influenced by several factors, including errors from the Tesseract OCR model, mistakes by annotators, the location of the meme (regions in Ethiopia), missing context, and the model itself.

7.1. Text Unimodal Error Analysis

The textual model correctly labeled 125 instances out of the total test dataset, indicating that 73 instances were incorrectly labeled. In order to comprehensively grasp the causes of errors, we conducted an in-depth error analysis on 50% of the mislabeled datasets, taking into account various influencing factors. Our investigation revealed that 61.1% of the errors originated from the mistakes done by the model, while 13.89% were linked to the Tesseract OCR extraction. Missing context, especially when image and text were separated, contributed to 8.33% of the errors. Annotator errors were responsible for 5.56% of the mistakes. Surprisingly, geographical location (the region where the meme was generated) played a role in 2.7%

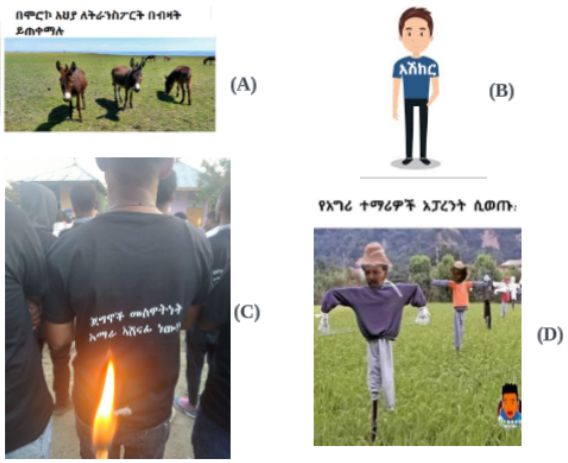

Figure 4: Model errors: Samples of wrongly predicted Memes against the gold labels. **English** translations of the meme texts on images (A, B, C and D) are presented in Table 6.

of the errors. Lastly, 8.33% of the errors were challenging to categorize into specific categories, falling into the ambiguous memes group.

7.2. Image Unimodal Error Analysis

The image classification model demonstrated that out of the total test dataset, 137 instances were correctly predicted, indicating that 61 instances (30.8%) were wrongly predicted. To gain a comprehensive insight into the factors contributing to these inaccuracies, we conducted a detailed error analysis on 50% of the incorrectly predicted instances. Upon examination, it became clear that 60% of the errors stemmed from inaccuracies in the model. About 13.3% of the errors were attributed to missing context when image and text were separated, with 10% attributed to annotator errors. The remaining 16.67% of the errors proved to be ambiguous, posing a challenge even for human categorization.

7.3. Multimodal Error Analysis

Similarly, the multimodal model also exhibited errors in its predictions. The multimodal model is able to catch 148 out of the total test instances properly, which accounts for 74.7% accurate prediction. After careful review, it was clear that 56% of the errors stemmed from model inaccuracies. Another 12% were attributed to Tesseract OCR effects, while 8% were caused by annotator errors while the location of the meme was attributed for 4% of the errors. The remaining 20% of errors were difficult to categorize into specific groups.

As illustrated in Table 6 and Figure 4 (B), it is evident that the labeling of the word is inconsis-

Meme	Tesseract OCR	Correct Texts on Memes	English	Gold	Pred.
Figure 4(A)	በሞሮኮ እህያ ለትራንሰፖርት በብዛት ይጠቀማሉ	በሞሮኮ እህያ ለትራንሰፖርት 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: Model errors: Samples of wrongly predicted memes against the gold labels

tent and varies in meaning across different regions. For instance, in **Gojjam**[2] and **Wollo**[3], it represents **slave** or **servant** for men, whereas in **Gondar**[4], it signifies a **Young boy or girl**. In the context of Table 6 and as depicted in Figure 4 (C), it is evident that the incorrect labeling arises from a failure of the Tesseract OCR to accurately extract the text from the memes. The reason is that Tesseract OCR may encounter difficulties in extracting text due to the non-straight line nature of the text arrangement within the meme images. The text on the image is **intentionally distorted** and **curved**. This departure from standard, linear text presentation can pose challenges for Tesseract OCR. In Figure 4 (A), the model classified it as hate speech, likely because the word **donkey** has been used as a derogatory term in Ethiopia. In contrast, Figure 4 (D) was labeled as "normal" by the model, possibly as the text is a sarcastic expression, specifically directed at agricultural students.

8. Conclusion and Future Work

This paper introduced the Amharic meme dataset and conducted multimodal classification experiments. We successfully collected a dataset comprising 2k memes sourced from prominent social media platforms, including Facebook, Twitter, and Telegram. A dedicated web-based annotation tool called **HateMemAnno** was designed to facilitate the annotation of Amharic memes within a multimodal context. Furthermore, we harnessed OCR technology, specifically the Tesseract library, to extract textual content from meme images. We employed a preprocessing technique to generate text and image features and feed both these input vectors to the model. In summary, we divided the dataset into training, validation, and testing subsets. We efficiently harnessed the *Concatenate* method of Keras to fuse unimodal features. Employing BiLSTM, LSTM, and CNN algorithms, we conducted multiple experiments for each modality, analyzing their performance. The findings revealed that multimodal features, particularly the inclusion of image data, significantly enhanced model performance. Notably, the BiLSTM model with multimodal inputs outperformed all other models, regardless of modality.

In the future, there is room for dataset expansion to bolster the hate speech detection model's capabilities. Although existing deep neural network models exhibit strong performance, we are presently investigating the potential of utilizing multimodal transformer models to harness multimodal features for the prediction of hate speech in Amharic memes. We also recommend enlarging categories to encompass various forms of hate speech, such as racism, sexism, religion, and political hostility. Additionally, exploring new modalities like audio and emojis could enhance the model. To facilitate further research in multimodal hate speech detection for low-resource languages like Amharic, we released our dataset, annotation tool, guidelines, top-performing models, and source code under a permissive license[5].

Limitations

One of the main limitations of this study is the relatively small size of the dataset and its coverage across different domains. Our dataset does not encompass every aspect of memes that are prevalent in the current social media landscape in Ethiopia. With additional budget and resources, it would be possible to collect more data and develop a more robust scraping technology to gather a more extensive dataset. The utilization of APIs from platforms like Facebook, Telegram, and Twitter could also enhance data collection. Another limitation pertains to the performance of the

[2] https://en.wikipedia.org/wiki/Gojjam

[3] https://en.wikipedia.org/wiki/Wollo_Province

[4] https://en.wikipedia.org/wiki/Gondar

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

Tesseract OCR tool. Improvements in this aspect could lead to more accurate text recognition and extraction from images. Additionally, considering alternative OCR technologies might mitigate errors in data extraction. Moreover, while prior studies such as D'hondt et al. (2017) have suggested the utilization of language models for post-OCR processing and error correction, the scope of this study did not allow for an in-depth exploration of this approach. It is crucial to conduct further research to assess the suitability and effectiveness of specific language models designed to address Amharic text errors. Overcoming these challenges holds promise for strengthening the reliability of research outcomes and, consequently, advancing the field of hate speech detection in the context of Amharic memes and social media.

9. References

Zeleke Abebaw, Andreas Rauber, and Solomon Atnafu. 2022. 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)*, pages 603–618, Bahir Dar, Ethiopia. Springer.

Ayme Arango, Jesus Perez-Martin, and Arniel Labrada. 2022. HateU at SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 581–584, Seattle, WA, USA. Association for Computational Linguistics.

Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Tegegne Asfaw, and Chris Biemann. 2022a. Challenges of Amharic hate speech data Annotation using Yandex Toloka crowdsourcing platform. In *Proceedings of the sixth Widening NLP Workshop (WiNLP)*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022b. The 5Js in Ethiopia: Amharic hate speech data annotation using Toloka Crowdsourcing Platform. In *Proceedings of the 4th International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120, Bahir Dar, Ethiopia. IEEE.

Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023a. 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., Shoumen, Bulgaria.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023b. Exploring Amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP2023)*, pages 59–59, Varna, Bulgaria. Association for Computational Linguistics.

Tadesse Destaw Belay, Abinew Ali Ayele, Getie Gelaye, Seid Muhie Yimam, and Chris Biemann. 2021. Impacts of homophone normalization on semantic models for Amharic. In *Proceedings of the 3rd International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 101–106, Bahir Dar, Ethiopia. IEEE.

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)*, pages 1137–1142, Salem, India. IEEE.

João Bran and Adeline Hulin. 2023. *Social Media 4 Peace: local lessons for global practices*. Countering hate speech. the United Nations Educational, Scientific and Cultural Organization (UNESCO.

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*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing different supervised approaches to hate speech detection. In *Proceedings of The Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, pages 230–234. European Language Resources Association (ELRA), Turin, Italy.

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, pages 25–35, Florence, Italy. Association for Computational Linguistics.

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)*, pages 102–107, Bahir Dar, Ethiopia. IEEE.

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 Technology*, 14(31):2567–2578.

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*.

Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2017. Generating a training corpus for OCR post-correction using encoder-decoder model. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1006–1014, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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*, pages 65–70, Santa Fe, NM, USA. Association for Computational Linguistics.

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)*, pages 1459–1467, Snowmass Village, CO, USA. IEEE.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online Only. Association for Computational Linguistics.

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*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

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*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Md. 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*, pages 293–308, Beijing, China. Springer International Publishing.

Simon Kemp. 2023. Digital 2023: Global overview report. Accessed Oct. 20, 2023, URL: https://datareportal.com/reports/digital-2023-global-overview-report.

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)*, pages 2611–2624, Vancouver, Canada.

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*, pages 150–161, Online. Association for Computational Linguistics.

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*, pages 14867–14875, Palo Alto, CA, USA. Association for the Advancement of Artificial Intelligence.

Zewdie Mossie and Jenq-Haur Wang. 2018. Social network hate speech detection for Amharic language. In *4th International Conference on Natural Language Computing (NATL2018)*, pages 41–55, Dubai, United Arab Emirates. AIRCC Publishing.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):1–16.

Konstantinos Perifanos and Dionysis Goutsos. 2021. Multimodal hate speech detection in Greek social media. *Multimodal Technologies and Interaction*, 5(7):1–10.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. 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*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *ArXiv*, abs/2202.06218.

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)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Siva Sai, Naman Deep Srivastava, and Yashvardhan Sharma. 2022. Explorative application of fusion techniques for multimodal hate speech detection. *SN Computer Science*, 3(2):1–13.

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*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

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*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Surafel Getachew Tesfaye and Kula Kakeba. 2020. Automated Amharic hate speech posts and comments detection model using recurrent neural network. *Preprint*. Version 1.

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*.

Getachew Assefa Woldemariam. 2020. The Language policy of federal Ethiopia: A case for reform. *J. Ethiopian L.*, 32:83.

Baye Yimam. 1999. The verb to say in Amharic. *Journal of Ethiopian Studies*, 32(1):1–50.

Seid Muhie Yimam, Abinew Ali Ayele, and Chris Biemann. 2019. Analysis of the Ethiopic Twitter dataset for abusive speech in Amharic. In *In Proceedings of International Conference On Language Technologies For All: Enabling Linguistic Diversity And Multilingualism Worldwide (LT4ALL 2019)*, pages 210v–214, Paris, France.

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).

Yanling Zhou, Yanyan Yang, Han Liu, Xiufeng Liu, and Nick Savage. 2020. Deep learning based fusion approach for hate speech detection. *IEEE Access*, 8(1):128923–128929.

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*, pages 53–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.