ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Amharic Hateful Memes Detection on Social Media

By:

Abebe Goshime

Advisor:

Dr. Yalemzewud Negash

A thesis submitted in partial fulfillment of the requirements for the degree of ***Master of Science*** in **Computer Engineering**

February, 2024

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Amharic Hateful Memes Detection on Social Media

By:Abebe Goshime

Dr. Bisrat Derebssa	
Dean, SECE, AAiT	Signature
Dr. Yalemzewud	
Thesis Advisor	Signature
Dr. Bisrat Derebssa	
External Examiner	Signature
Dr. Fitsum Assamnew	
Internal Examiner	Signature

Declaration

In order to meet the criteria for a Master's degree in Computer Engineering, I thus certify that the work provided in this paper, named Amharic Hateful Meme Detection on Social Media, was submitted to Addis Ababa University. Every attempt is made to explicitly state along with the appropriate citation of sources anywhere else that other people have contributed. Declared by the undersigned, this thesis has not been submitted for consideration for a degree at any other university.

Abebe Goshime

Acknowledgment

Before all else, I want to express my gratitude to God for all of his blessings, grace, and strength throughout my life. It was with his help that I conducted this research. For her love, prayers, support, and sacrifices made on my behalf to further my studies, my mother Yeshi.G has my sincere gratitude. My study adviser, Dr. Yalemzewud.N, has my profound gratitude for his invaluable time, advice, helpful suggestions, and remarks that helped me have a positive research experience. My instructor Dr. Fitsum.A deserves special recognition for his invaluable help in getting my thesis completed so quickly. They have been a crucial part of my thesis and without them we would not be where we are now. Heartfelt and big thanks to my friends for their constant guidance and inputs that helped me in completing this research work.

Abstract

Hateful meme is defined as any expression that disparages an individual or a group on the basis of characteristics like race, ethnicity, gender, sexual orientation, country, religion, or other characteristics. It has grown to be a significant issue for all social media platforms. Ethiopia's government has increasingly relied on the temporary closure of social media sites but such kind of activity couldn't be permanent solution so design automatic system. These days, there are plenty of ways to communicate and make conversation in chat spaces and on social media such as , text, image, audio, text with image, and image with audio information. Memes are new and exponentially growing trend of data on social media, that blend words and images to convey ideas. The audience can become dubious if one of them is absent. Previous research on the identification of hate speech in Amharic has been primarily focused on textual content.

We should design deep learning modal which automatically filter hateful memes in order to reduce hate content on social media. The basis of our model consists of two fundamental components. one is for textual features and the other is for visual features. For textual features, we need to extract text from memes using optical character recognition (OCR). The extracted text through the OCR system is pixel-wise, and the morphological complex nature of Amharic language will affect the performance of the system to extract incomplete or misspelled words. This could result in the limited detection of hateful memes. In order to work effectively with an OCR extracted text, we employed a word embedding method that can capture the syntactic and semantic meaning of a word. LSTM is used for learning long-distance dependency between word sequence in short texts. The visual data was encoded using an ImageNet-trained VGG-16 convolutional neural network. In the studies, the input for the Amharic hateful meme detection classifier combines textual and visual data. The maximum precision was 80.01 percent. When compared to state-of-the-art approaches using memes as a feature on CNN-LSTM, an average F-score improvement of 2.9% was attained.

Key words: social media, Memes, hate speech, word embedding, OCR, VGG-16, LSTM

Contents

List of Figures

List of Tables

List of Abbreviations

OCR	Optical Character Recognition
RNN	Recurrent Neural Network
CBOW	Continuous Bag-of-Words
GAP	Global Average Pooling
CNN	Convolutional Neural Network
VGG	Visual Geometry Group
FDRE	Federal Democratic Republic of Ethiopia
SVM	Support Vector Machine
RF	Random Forest
NB	Naïve Bayes
FC	Fully-Connected
ILSVRC	Image Net Large Scale Visual Recognition Challenge
LSTM	Long-Short-term-Memory
ReLU	Rectified Linear Unit
VQA	Visual Question Answers
GRU	Gated Recurrent Unit
NLP	Natural Language Processing

Chapter 1

Introduction

1.1 Background

Social media are computer-based technologies or form of communication on the internet used to facilitate the sharing of opinions, ideas, thoughts, and information through virtual networks and communities[8].Social media provides quick electronic communication with feature of user generated content and personalized user profiles. Interactions among users on social network platforms are usually positive, constructive and insightful. However, sometimes people also get exposed to objectionable content such as hateful meme, bullying, and verbal abuse etc. Contents on social media can be personal information, documents, videos, photos and entertainments. conversations, shared information, and created online content on social medias can be in different mode, such as text, audio or video in combination way.

One of the essential aspects of modern democratic countries around the world is the right to speech [9]. This means that people have inherent freedom to express themselves on whatever subject they like. Social media are good place for freedom of speech, but it has drawbacks, such as promoting hate and dehumanizing messages directed at specific individuals or groups based on their political affiliation, religious views, ethnicity, sexual orientation, and other factors [10]. In January 2020, there were 21.14 million internet users in Ethiopia, up 2.6 percent from 2019 and 2020 [11].

Ethiopia has been utilizing social media more frequently, which is largely due to the growth of various mobile technologies and the number of individuals accessing the internet. In study [12], social media usage in Ethiopia is changing gradually from posting, tweeting, or sharing matters to significantly debating political, financial, and social situation. There is growing evidence that hate speech on social media in Ethiopia is encouraged and contributes to violence [11].The rise of reckless social media behavior and the propagation of hatred have come under fire in recent years, especially in relation to the ethnic conflict in Ethiopia[13].

Hate speech is content that criticizes a person or a group based on traits such as race, ethnicity, gender, sexual orientation, nationality, religion, or other factors [14].Hateful meme propagation is highly destructive in human social life, as it cause individual to become more biased and discriminate against others, affect societies in many aspects, such as mental health of targeted audiences, social interaction, and at last it leads to violence and distraction of properties. Hateful meme detection is a vital tool for preventing the spread of hateful memes.

Different countries have their own laws regarding to hate speech controlling and preventing mechanism. Since social media intensifies ethnic tensions among all ethnic groups, the Ethiopian government has enacted laws prohibiting hate speech on social media and other electronic devices [15]. The proclamation states that it is strictly forbidden and illegal to post anything on social media that could possibly offend someone based on their color, age, gender, sexual orientation, national origin, religion, political beliefs, disability, or any other attribute [16].In an effort to control hate speech that incites ethnic violence through intermittent disruptions of internet connections, the Ethiopian government regulates and keeps an eye on data uploaded to social media [11]. Reports indicate that there is an increasing number of state sensors and shutdowns targeting mobile Internet services for political or security reasons all over the country[12].This opposes freedom of speech. Mechanisms should be conceived to protect such communities from banning of Internet services and at the same time, the risk of online hate speech.

Social media platforms have tried different techniques to defend such ill-suited behaviors, that the one who constantly change strategies to evade the current hate speech detection methods of social networks platforms. Researchers are working harder to identify hate speech on social media, but hatemongers are able to adapt their tactics to avoid the systems in place for detecting hate speech. Hateful memes are also spread hatred through social networks. It would be beneficial to reduce the negative societal impact of unpleasant memes by automatically identifying them. Tasks that employ the multi-modal analysis techniques are image captioning [17], multimodal image retrieval

[18], visual question answering [19] and multimodal publications [20].

Unlike traditional multimodal tasks, where textual and visual information are semantically linked, the issue of detecting violent memes is on the multimodal information [21] and only multi-modal models have demonstrated success in learning, while unimodal models collapse under pressure. For hateful meme there might be an alternative image or caption that cause to change its label in the datasets. For example in figure 1.1 (c) and (d) because of the visual information the meme changes from hateful to non-hateful and figure 1.1 (a) and (b),(i) and (j) and (g) and (h) also flip its label with the only changes of caption.

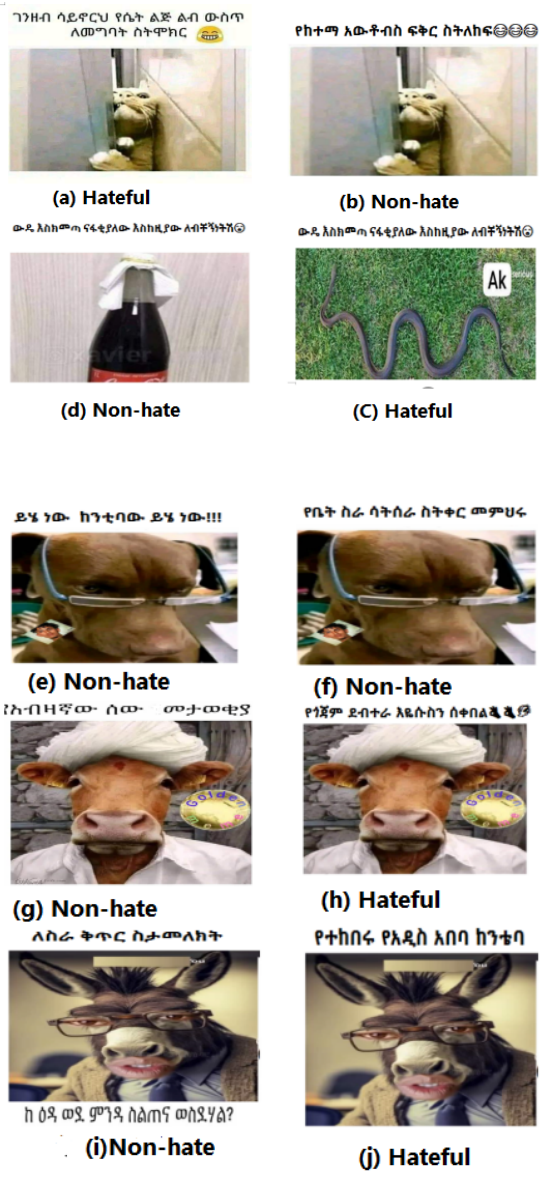

Figure 1.1: Examples of Amharic memes

While working with Amharic memes extracting text out of the image is also another challenge. as Amharic is morphological complex and the extracted text is in pixel-wise that would occurs in the OCR system. This could be cause to wrong predication of one or two letters in a word and result in limited detection of hateful memes. So it needs to modeling very precise correlations between the image and texts embedded on it as it need to combine the visual and textual information together[21, 22].

Several works like [23, 24, 21, 25, 26, 27] have have been done to hateful memes detection, most of which were for the English language. Unfortunately, no significant studies have been conducted on hateful memes regarding low-resource languages, especially Amharic. As hateful meme is distractive and it affect mental and nation co-existence by creating tension between and many other related issues, social media contents has to be detected for Amharic language using the state-of-the-art.

1.2 Problem Statement

Social media now contains a huge variety of significant data types. Only textual content analysis has been used in prior research on Amharic hateful detection. Memes are multimodal, so identifying hateful meme within them is a challenging but crucial task that requires combining the textual and visual features. We need to extract text from memes in order to process them separately by using OCR. Due to Amharic's morphological complexity, which can make OCR text extraction more challenging, this could result in limited detection of hateful memes. In order to effectively work with an OCR text extractor, we should employ a word embedding method that can capture the syntactic and semantic meaning of a word. Therefore, the input for the Amharic hateful meme detection classifier combines textual and visual data. Nevertheless, no prior study has examined the effects of nasty memes in Amharic. For this reason, the following research questions are the focus of the current study.

- RQ1.Can visual information be used to enhance automatic detection of Amharic hateful memes on social media?
- RQ2. How much does visual and textual information contribute to Amharic hate-

ful meme detection?

1.3 Objective

1.3.1 General objective

The general objective of this research work is to investigate the effect of integrating visual information with textual information to Amharic hateful meme detection.

1.3.2 Specific objective

- Collecting the datasets from different social media.
- Create a combination of textual and visual information for Amharic hateful meme detection classifier.
- Examine how visual features affect the identification of harmful memes on social media.
- Examine the role that textual and visual information plays in the identification of harmful Amharic memes.
- To analysis the result and gives future research direction on Amharic memes.

1.4 Contribution

The foremost contributions of this research can be listed as below:

- The first Amharic memes has been gathered and annotated; it now serves as a corpus for future studies.
- We have substituted a very important feature extractor that extracts both textual and visual information to joint feature representation that is necessary for automatic design.
- We developed an automated system called Amharic hateful meme detection that replaces human labor with machines capable of directly processing, analyzing, and manipulating data related to Amharic memes.

1.5 Scope and Limitation of the study

The aim of this study is to design system that can detecte hateful memes for Amharic memes.This study does not take in consideration other form of data contents of social media such as audio, video and other multimodal data.

1.6 Methodology

- In order to conduct objective of this work, the following procedure and activities will be done.
- Literature Review: Related works on hate speech, hateful meme detection and multimodal problems with the algorithms and techniques used to feature extraction, representation and classification including fusion techniques in multimodal features will be reviewed.
- Data Collection and Preparation: - To this research work, we used manually collected meme datasets. For better performance of the system, different preprocessing tasks are performed on the datasets we gathered such as resize, gray scale, thinning and skeletonization and remove noise for image will perform.
- Data Annotation: - The preprocessed data will be annotated manually for memes to investigate the effect of using memes on the automatic labeling system.
- Feature Extraction:- After reviewing various papers and investigating existing feature extraction mechanisms, we chose a concatenation of textual features and visual features as automatic feature extraction and appropriate selection for our thesis.
- Model Development: - Deep neural network models are used to construct the Amharic violent meme identification model after the automatically generated features and vector results are selected as input.
- Evaluation: Experiment will be conducted to test use of labeled memes on Amharic

hateful memes detection. The performance of the system will be evaluated in terms of precision, recall, accuracy, and F score.

1.7 Organization

The remains of this work are structured as follows:

In Chapter 2 describes theoretical background, definition of hateful meme, Amharic language and its challenges to hateful meme detection process, detection approaches, models and algorithms.

In chapter 3 is the literature review that discuses about existing hateful meme detection approaches. In Chapter 4 explains how the study was designed and carried out, data preparation, and system architecture, feature engineering and fusion techniques with performance measurement ways.In Chapter 5 explains the outcomes and findings regarding the experiment. Finally, In Chapter 6 describes the conclusion and recommendation of the work.

Chapter 2

Theoretical Background

2.1 Introduction

This chapter deals with the theoretical concepts of hateful meme definition, the Amharic language and its challenge in hateful meme detection, definition of hateful meme according to the FDRE constitution, deep learning algorithms and social media discourse analysis approaches.

2.2 Amharic Language

Amharic is a Semitic family of the Afro-Asiatic language group that related to Hebrew, Arabic, and Syrian. Which is written from left-to-right by using unique geez script, and the federal working language of Ethiopia. It has 275 alphabets with the writing script known as "Fidel". Speaking after Arabic, Amharic is the second most spoken Semitic language globally, and its script, known as Ethiopic, comes from the Geez alphabet [28].It is one of the less-resourced languages, with few tools and methods for processing language. However, it may be able to assist scholars studying computational linguistics in going deeper and creating more beneficial computer- or Internet-based models and applications.

Although Amharic texts have traditionally been transliterated into Latin characters by social media users to express their opinions, the advent of the Amharic Unicode font and its integration into various technologies has made it possible for many online publishers and users to communicate using the native Ethiopic script. The increasing usage of Amharic as a medium of internet communication among speakers is enhancing networks with Amharic resources and creating new possibilities for language study. In Amharic, there are several words that are considered vulgar, rude, or derogatory in society. These words mainly relate to sexual racial, ethnic, religious, or social class, and they should be used with control.There are many ways where their use can be hurtful and upsetting.

2.3 Challenges of Amharic Hateful Memes Detection

Amharic's multimodal content poses a challenge for hate meme detection, in contrast to traditional multimodal tasks where textual and visual information are semantically linked and there is insufficient data set and annotation. Amharic, as a language, is morphologically complex, with characters having similar sounds with different shapes, the usage of semantically the same words interchangeably, compound word formation using hyphens or spaces, and sometimes they may merge, but they have different lexicons when they are put separately. Especially using Amharic texts embedded in images has many challenges, such as the highest similarity between some characters, noise over the image, font style, stroke, and spelling variation that might cause the OCR system to extract incomplete or misspelled words from the image, and a high number of non-Amharic characters and irrelevant symbols.

2.4 Hate Speech Definition

One of the greatest inventions in human history, the Internet has united people of all racial, religious, and national backgrounds. Billions of people have been connected by social media platforms like Facebook and Twitter, which enable them to instantly share their thoughts and opinions. In addition to the online, platforms give an environment to discussions that are damaging to specific groups of people. The FDRE Constitution has provided proclamation with respect to hate speech and false information. The 2005 Criminal Code of Ethiopia, and Information and Network Security Agency Re-establishment Proclamation No.808/2013, Computer crime proclamation 958/2016, Draft proclamation to prevent the dissemination of hate speech and false information[15] and Ethiopian disinformation control and hate speech protection [16]. Proclamation [15] categorizes the computer crimes in to three. Those are, First, committed against computer, computer system, computer data or computer network. Second, conventional crime committed by means of a computer. Third, illegal computer content data disseminated through a computer, computer system, or computer network. According to Section 3 of the Proclamation, hate speech on the internet is essentially

defined as prohibited content. In particular, posting anything online that can be seen as intimidating is prohibited by Proclamation Article 13(1), which entails criminal consequences.Additionally, publishing anything that incites disorder, terror, violence, or conflict is prohibited by Article 14 and is punishable by law. The proclamation also introduces the idea of crimes against people's reputations and liberties. As per the computer crime proclamation's articles 13 and 14, an individual who intentionally causes serious harm or danger to another person or their family by sharing any produced information, audio, video, or image via a computer system may be sentenced to three years in simple imprisonment or, in severe cases, five years in rigorous imprisonment. Second, anyone who sends or repeatedly transmits information about the victim or his family through a computer system, or who keeps the victim's computer communication under surveillance, and who does so in a way that incites fear, threat, or psychological strain on another person, faces a maximum sentence of five years in simple imprisonment or, in more serious cases, ten years in rigorous imprisonment.

Thirdly, sharing any text, picture, audio, video, or other content via a computer system that disparages the dignity or reputation of another individual is punished, upon complaint, by a maximum of three years in simple imprisonment, a maximum fine of Birr 30,000, or both. According to the Ethiopian hate speech and disinformation Prevention and Suppression Proclamation [16], Page 1233/9 under Proclamation No. 1185 /2020, "Hate speech is the speech that intentionally promotes discrimination, hatred, or attack against a discernible group of identity or person, based on race, ethnicity, gender, religion or disability", this shows that the definition of hate speech is any expression that, in a way, targets, disparages, discriminates against, or encourages violence against another individual on the basis of that individual's race, religion, color, sex, handicap, nationality, immigration, language, appearance, or any combination of these characteristics.

In our research context, content in memes considered as hate if it is one of the following categories. The text, image and image -text combined information on social media is defined as hate if its intent is any of the following.

1. If a post is rude or exhibiting lack of respect toward certain individuals or group

of individuals (ethnic groups):

- Slurs: phrases that try to attack a culture or ethnicity in some way,
- Racism: phrases that intimidate race or ethnicity of individuals,
- Crude language: expressions that embarrass people, mostly because it refers to sexual matters or excrement,
- Taboo: expressions, which are forbidden within a certain society/community. Many expressions are forbidden because of what they refer to, not necessarily, there is some particular taboo words used in the expression.
- Unrefined languages: some expressions that lack polite manners and the speaker or writer are harsh and rude.

2. To cause harm (to oneself or others): harms including physical, psychological, etc.
    - References to handicaps: These phrases attack the reader using his or her shortcomings (i.e., IQ challenged).
3. Related to an activity that is illegal as per the laws of the country: The rise of irresponsible social media activism and fake news in recent times is being blamed as the catalyst especially for ethnic related violence in various parts of the country.
4. Has cause of extreme violence:
    - Extremism: These phrases target some religion or ideologies, Provocative language (expressions that may cause anger or violence).
    - Social media posts that make society to get angry at one another.
5. When the images are sexually explicit, nudity, materials which being cause for violence and suicide.
6. When concept of the combination of text with image together contains hate and harmful contents such as slur, racism, crude language and dehumanized are considered as hateful meme.

Based on the above definitions, when we say hateful meme, implicitly we are talking about every context that falls into one or more of the defined cases. Characterizing, certain expression as offensive has an important role in advancing the values of dignity and equality.

2.5 Hate Speech Detection Approaches

The strategy already known in text mining and sentiment analysis are employed to the specific problem of hate speech detection. The frequently used hate speech detection approaches can be grouped into lexicon-based, machine learning and deep-learning approach.

2.5.1 Lexicon Based Approach

By using a previously created vocabulary of words with varying levels, the lexicon method essentially matches each token with a word that is already available. The model then takes the sentence as input and tokenizes it, calculating an average or sum to determine the sentiment and context of the sentence [29, 30, 31].

2.5.2 Machine Learning Approach

Machine learning algorithms like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) are used for learning [32, 33, 34].In order to create classifiers, machine learning models use samples of labeled text; however, these feature representation techniques are unable to capture the multi-view aspects of hate speech. A test data set is used to determine the classifier's performance, and a training set is used to teach the machine learning classifier how to distinguish between different types of documents. IF-IDF, BOW, unigram, n-gram trigram, and word2vec are well-known feature representation techniques in hate speech detection. To overcome the challenges posed by machine learning, researchers must turn to deep learning techniques, such as RNN and CNN models with word embedding technique for feature representations like

fast-text and word2vec, can capture both the syntactic and semantic meaning of words [35, 23].

2.6 Feature Extraction

These are the most critical steps in learning rich and informative representations from raw input data to produce accurate and reliable outcomes. Generating features for use in prediction and classification tasks is known as feature extraction [36]. However, not all features are equally important for a prediction task, and some features might even introduce noise in the model. To identify or classify user-generated content, textual and visual features indicating hate must be extracted. Machine learning models require input features that are relevant and important to predict the outcome. Since the extraction step converts written human language and image into a form that a computer can understand. A collection of features that together make up a feature vector—a representation of the datasets—is the end result of the extraction process.

2.6.1 Text Feature Extraction

Feature extraction for textual data can be done by different techniques such as TF-IDF, n-grams, Bag of words, word embedding, combination word embedding, and character n-gram embedding are a few examples.

Bag of Words: - Word occurrences in a document are described by a representation of text called a "bag of words." Information retrieval and natural language processing are made easier by the bag-of-words model, which is a textual representation of words. To extract features for each sentence, a distribution of different words in the sentence must be computed, i.e. how many times in the phrase does each word from the vocabulary appear?. The text is represented in this typical feature extractor by a bag of its words. Pay no attention to grammar or even word order while maintaining multiplicity. Bag of Word ignores word order to the benefit of syntactic and semantic content [30]. Therefore, it will be able to result in miss classification if the words are uses in different contexts.

N-gram: - An extension of Bow is lead to the incoming of the bag of n-grams, which replaces the unit of interest in Bow from words to N connecting tokens. A token is usually a word or a character in the text, giving rise to word n-gram and character n-gram models. The disadvantages of N gram are it ignores syntactic and semantic content. Therefore, it can bring miss classification when the words are practices in different contexts [30].

Neural Word Embedding: - In recent years, deep learning methods have been used in Amharic hate speech detection[37] and automatic feature extraction. Word embedding is one of the methods used as an automatic feature extractor because it allows for the discovery of both semantic and syntactic relationships between words. In order to encode and embed words into numeric vectors that may be utilized for arithmetic operations, a two-layer neural network known as the word embedding model was developed. it operates on two basic models: the Continuous Bag-of-Words (CBOW) and the Continuous Skip-Gram. The CBOW model uses a continuous, distributed representation of the verbal context to predict the value of the current word while the Continuous Skip-Gram model predicts the verbal context using the current word. This allows capturing more refined attributes and contextual cues inherent in human language [1].

Both the continuous bag-of-words and skip gram are consists of an input layer, a projection layer, and an output layer to predict nearby words. The model iterates over the words in a given set of sentences, trying to forecast the next word or utilize the present word as a predictor[1]. Continuous bag-of-words, predicts the present word grounded on the context, whereas skip gram predicts the neighboring words given the current word. The output is the probability distribution over all words in the vocabulary, which defines the likelihood of a word being selectness as the input word's context.

2.6.2 Visual Feature Extraction

A sort of dimensional reduction known as feature extraction involves efficiently representing a significant number of the image's pixels so as to effectively capture the interesting areas of the image. During the feature extraction phase, various features

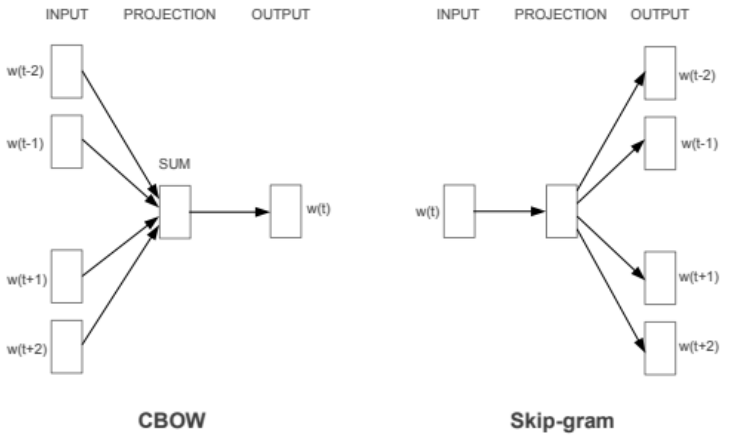

Figure 2.1: CBOW and the Skip-gram architecture (picture credit ([1])

are determined and then extracted to support the learning process of different machine learning models.

Color, texture, and shape are examples of general features, which are characteristics that are independent of application.Depending on the abstraction level, they can be divided even further into pixel-level features: characteristics like color, position, and local features that are computed at each pixel: characteristics derived from image segmentation or edge detection as well as the outcomes of the image's subdivision. Worldwide characteristics are characteristics that are computed for a regular subset of a picture or the complete image. Features unique to a domain: characteristics specific to a given application, such as faces, fingerprints, and conceptual traits.
For image processing, feature extraction techniques include edge detection, texture analysis, and color-based methods. Some features related to images are:-

Gray-scale features: It is useful in situations where color information is not relevant or important for the task at hand such as text recognition (OCR) reading text from an image and edge detection detecting boundaries or edges in an image.

Edge Features: It can be useful for certain image processing tasks that involve detecting and analyzing edges or boundaries in an image like Object detection, Motion detec-

tion, Image segmentation, Medical imaging and Optical character recognition(OCR). Edge features can be used to identify variations in pixel intensity and to detect sharp transitions between regions in the image. It can be effective in situations where the goal is to perform simple image processing operations, such as edge detection or boundary extraction. Edge features can also be useful in situations where the image data is relatively simple and the edges are well-defined.

Histogram of Oriented Gradients (HOG): HOG is a feature descriptor that computes histograms of directed gradients in specific areas of a picture to extract gradient information. Computer vision tasks like object, face, and pedestrian detection, text classification, and optical character recognition (OCR) frequently use HOG.

Local Binary Patterns (LBP): Local Binary Patterns (LBP) is a widely used feature extraction method for analyzing texture information in images. LBP captures the local structure of an image by comparing the gray values of a pixel to its surrounding neighbors and encoding the result into a binary pattern. LBP features are often used where there is limited data and in machine learning algorithms for various computer vision tasks such as face recognition, texture analysis, and object recognition. LBP is computationally efficient and can operate in real-time applications, such as surveillance and tracking systems, where the processing speed is crucial.

Color-based features: Color-based features are a type of feature extraction technique that capture the color distribution and statistics of an image and also used for image retrieval. These features are often used in computer vision applications such as object recognition, image segmentation, and content-based image retrieval. It is computationally efficient than CNNs and used where there is limited data.

Auto encoders: Auto encoders can be used for feature extraction in scenarios where the CNN architecture is too computationally expensive or where the amount of labeled data is limited [28]. Feature extraction involves training an auto-encoder on an image dataset in order to extract features that can be utilized in subsequent machine learning

models. Auto encoders are commonly employed in unsupervised learning tasks, where the objective is to learn a compact representation of the input data, such as anomaly detection, dimensionality reduction, picture denoising, compression, and the creation of new data points. They can assist in removing noise and extracting essential characteristics, which makes them especially helpful when the input data is noisy and high dimensional. If the task requires supervised learning, such as image classification or object detection, VGG 16 are the better choice. There are cases where both auto encoders and VGG 16 can be used together for feature extraction.

VGG-16 The designs of the VGG-N models are extremely similar; however, they differ in having N numbers of layers. The VGG-16 network architecture is one of the most well-known CNN architectures. The Visual Geometry Group at the University of Oxford first proposed for use in image classification[38]. In the 2014 ILSVRC competition, the VGG group at the University of Oxford achieved incredible outcomes using this network structure.

2.7 Features Fusion Techniques

Fusion is a vital aspect as it outputs the multi-modal features by combining the visual features and text features. Some of the fusion techniques include element wise multiplication, bilinear pooling, concatenation, attention-based pooling, compositional approach, Bayesian-based methods, Element wise multiplication and concatenation are used when both features are of the same dimension and bilinear pooling is used when the dimensions are different. Feature fusion is the processes of constructing a joint vector representation from multiple modalities to perform a classification task. Exploring hateful memes detection involves vision and language tasks. To do that deep learning model such as CNN and RNN will trained jointly to learn a a joint vector representation from aligned multi-modal data [22]. A typical task in multimodal visual and textual analysis is to learn an alignment between feature spaces.

To constructing a joint representation of feature vectors, we need to have feature fusion techniques. Feature fusion techniques broadly divided into three types.

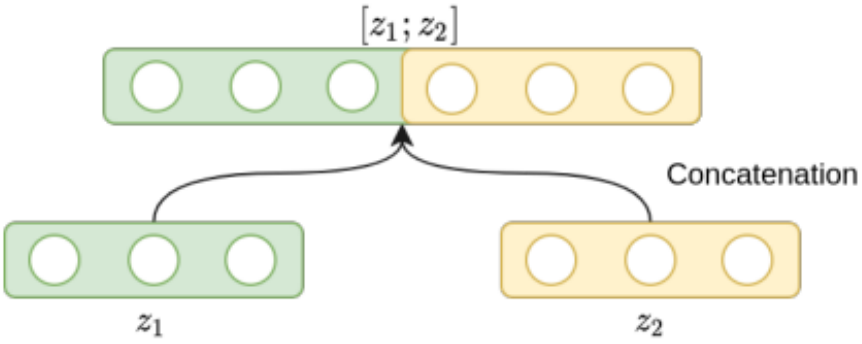

Figure 2.2: Simple Fusion (picture credit ([2]))

Feature-level fusion (i.e. early fusion): - To make a choice in early fusion, the output vectors from the learners of each modality are combined. Initially, each modality's directly extracted low-level features will be fused before being categorized. It provides an abundance of information from various sources[3].A single, significant representation that can result from this fusion process can be predicted.This fusion process can generate a single large representation that can lead to prediction. Categorizing features

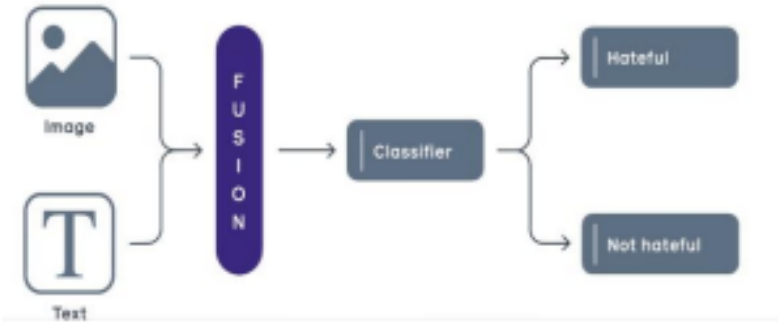

Figure 2.3: Early fusion (picture credit ([3]))

extracted from different modalities and fused independently is called decision-level fusion, often referred to as late fusion. With it, the final prediction scores from several classifiers were merged. This may reduce the overall effectiveness of the integration process since decisions are made independently by each modality.

The process of combining the multi-modal characteristics of early and late fusion before making a choice is called hybrid fusion, also referred to as "intermediate fusion." When combining the intermediate representations of various data streams spatially, the results

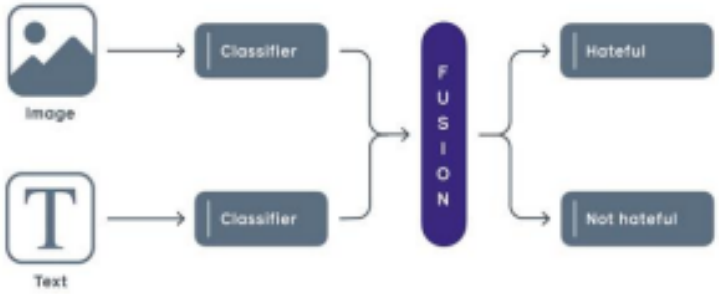

Figure 2.4: Late fusions (picture credit ([3]))

typically have different scales and dimensions, which makes the merging process more difficult. This is known as intermediate fusion.

2.8 OCR text extraction

Text on images can be machine-encoded using a tool called optical character recognition (OCR). The written content may consist of handwritten text, numbers, traffic signals, billboards, invoices, bank records, and signs. Text extraction can be accomplished in two steps: either by utilizing a single deep neural network model for both text detection and identification, or by combining text detection and recognition into one phase. Our approach uses an image for visual analysis and extracted text for textual analysis as its two inputs. To process each meme independently, we need to employ Tesseract-OCR to extract text from them.

Tessaract-OCR is an LSTM [6] network based OCR engine which helps in detection and recognition of texts embedded in memes image. OCR systems may extract incomplete or misspelled words from images, producing a large number of non-Amharic characters and irrelevant symbols. Other factors that could contribute to this process include noise in the image, the morphological complexity of the Amharic language, some scripts' visual similarity, inconsistent character formation, strokes, and font variations.Rather than manually correcting any incomplete or misspelled words in the extracted text from the image [39, 40, 41], Adding an extra embedding layer with a fixed embedding for simplicity allowed us to extract the required features using the built-in Fast-Text module.

2.9 Deep Learning Approaches

A subset of machine learning techniques called deep learning makes an effort to understand the layered model of inputs. It allows data representations with multiple abstraction levels to be learned by computational models made up of various processing layers. Neural networks are used in deep learning techniques to automatically extract multiple layers of features from the provided data. Various deep learning algorithms are applied to transfer learning tasks, such as text classification and hateful meme detection. A network's internal workings require an interface in order to use transfer learning to enhance a model. Users first feed new data with previously unidentified classifications into the network that already exists. Once the network has been adjusted, new tasks can be completed with improved classification abilities. Compared to other methods, this one has the advantage of requiring significantly less data, which cuts down on computation time to minutes or hours.

2.9.1 CNN Model

CNNs are generally the preferred choice for feature extraction from images because CNNs are specifically designed to processing images [38], with the advantages of built on ImageNet dataset and resolves the disappearing gradient and inflating gradient issues, performs better on additional datasets and tasks, and highlights the significance of the deepest model in visual representation. It also excels sophisticated recognition tasks using less detailed images. It is capable of carrying out increasingly difficult tasks and demonstrating exceptional performance in a variety of applications, including visual question answering [42, 19, 43] and hateful memes detection [23, 24, 21, 25, 26, 27] and extract complex and descriptive features with any variations such as lighting conditions, scale, and other factors in the image. They all start out as convolutional layers and finish up as fully linked networks for classification. In order to minimize the size of the activations along layers, they frequently also employ extra processes like max-pooling. The activation is often flattened into a one-dimensional layer, known as a feature vector, after a few convolutional layers. This one-dimensional layer is then used for other

tasks, such classification using any kind of model.

Compared to a conventional, fully linked network, these convolutions offer the benefit of having many fewer parameters because they frequently run with small kernels. When it comes to images or any other type of data containing exploitable spatial information, they work particularly well. These networks are mostly known in different applications such as image classification [4], image recognition [38]and face recognition [44].

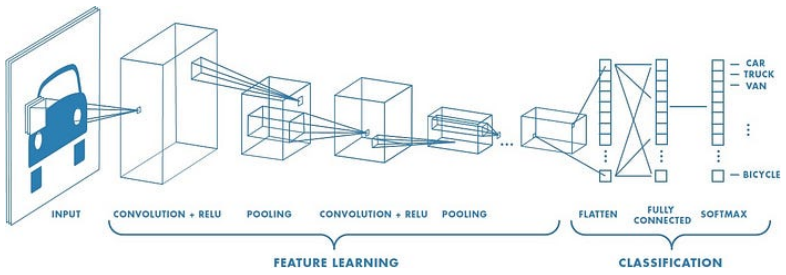

Figure 2.5: Architecture of Convolutional Neural Network ([4]))

Convolutional Layer: Convolutional Layer: The most significant component of the CNN design is the convolutional layer. It is composed of many convolutional filters. The input image, which is expressed as N-dimensional metrics, is convolved using these filters to create the feature map that is output.

Activation Function (non-linearity) Mapping the input to the output is the basic function of all activation functions in all varieties of neural networks. This involves altering the input information in a nonlinear manner. The rectified linear unit activation function, or ReLU, is a piecewise linear function that outputs zero if the input is negative and positive otherwise [45].

Pooling Layer: Pooling Layer: The main job of the pooling layer is to subsample the feature maps. Convolutional operations are used to create these maps. Stated differently, this technique produces smaller feature maps by downsizing huge feature maps. In addition, it maintains the majority of the salient features or data during the entire pooling procedure. Similar to the convolutional procedure, the size of the kernel and the stride are assigned before the pooling operation is performed. Diverse pooling ap-

proaches can be employed by distinct pooling layers. Gated pooling, average pooling, max pooling, global average pooling (GAP), and global max pooling are some of these strategies.[4].

Fully Connected Layer: Fully Connected Layer: In most CNN architectures, this layer is found at the very end. This layer employs the fully connected (FC) method, in which each neuron in the layers above it is coupled to every other neuron. It performs the role of CNN classifier. The FC layer receives its input from the final pooling or convolutional layer. An input vector is created once the feature maps have been fattened. The output of the FC layer represents the CNN's ultimate output.

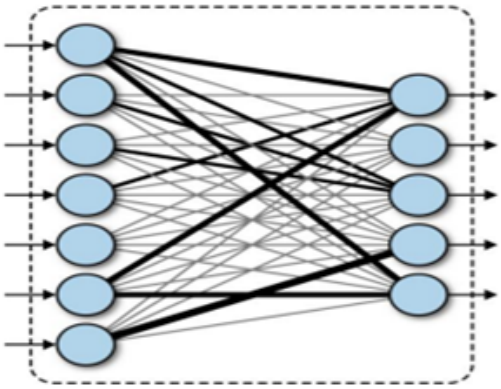

Figure 2.6: Architecture of Fully Connected Layer ([4]))

The convolutional neural network (CNN) accuracy is being continually increased with the development of new techniques. VGG16 is one of the CNN architectures that shows great accuracy and won the ILSVR in 2014[4]. The 16 in VGG16 refers to that it has 16 layers that have weight.it has been proved to provide relevant information of the image, for ImageNet classification, and other tasks such as visual tracking.

2.9.2 RNN Model

When a recurrent neural network (RNN) is used, the output from the previous step is used as the input for the current step. Conventional neural networks have independent inputs and outputs; hence, they do not require past word recall when predicting a sen-

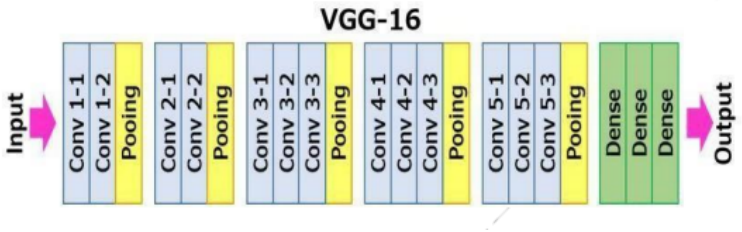

Figure 2.7: Architecture of VGG-16 network ([5]))

tence's next word. RNN was developed as a result, and it used a hidden layer to tackle this issue. An RNN's hidden state is that it uses the same job on all inputs, or hidden layers, with the same parameters for each input in order to produce the output. This means that the RNN maintains some information about a sequence, including its main and most important feature. This reduces the parameter complexity compared to other neural networks.

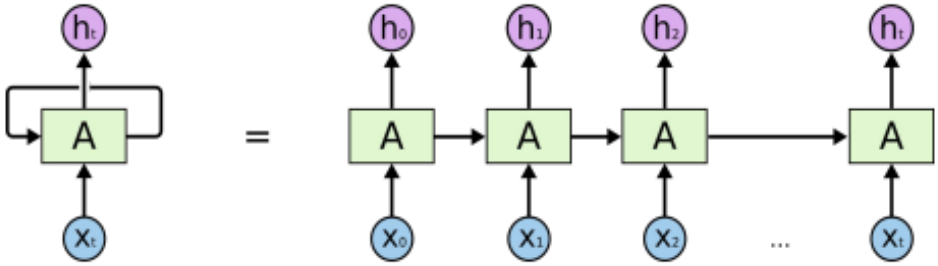

Figure 2.8: Architecture of RNN (picture credit ([6]))

Recurrent units, which are present in the hidden layer of an RNN, enable the algorithm to process sequence data. It continuously shifts a concealed state from one time step to the next and combines it with an input from the current one in order to do this. A feedback loop within the cell is how RNNs achieve memory, and this is the primary distinction between an RNN and a conventional neural network. Unlike feed-forward neural networks, which only allow information to flow between layers, the feedback loop permits information to flow within a layer. Then, RNNs need to specify what data is pertinent enough to be stored in memory. Long-short-term memory recurrent neural networks (LSTM), gated recurrent unit recurrent neural networks (GRU), and classic recurrent neural networks (RNN) are the three different forms of RNN.

Long-term dependencies can be challenging for the network to learn in a standard RNN

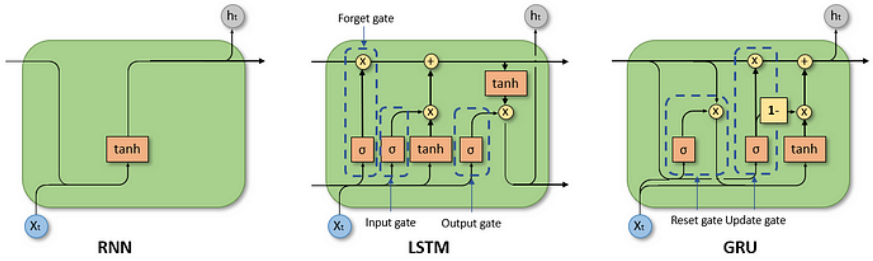

Figure 2.9: Architecture of RNN types (picture credit ([7]))

because it only has one hidden state that is transferred through time.An input gate, an output gate, and a forget gate make up the gating mechanism utilized by the enhanced recurrent neural network (RNN) design known as LSTM[6]. These gates help with deciding which data from the previous state should be forgotten or kept in the current state.

Therefore, the gating mechanism helps the LSTM solve the long-term information preservation and vanishing gradient problem that traditional RNNs encounter [46]. The input gate controls the data that is added to the memory cell. The forget gate controls what data is removed from the memory cell. The output gate also controls the data that is transmitted from the memory cell. This allows LSTM networks to selectively store or discard information as it goes through the network in order to understand long-term dependencies. Speech recognition and language translation are two tasks that heavily rely on the capacity of long-term memory to manage sequential information and resolve long-term dependency problems. For the activity at hand, longer-term context information is required.

LSTM offers several advantages such as its ability to tackle complex sequence learning tasks in speech and handwriting recognition; its function in optimizing the performance of LSTM structures, including coupling inputs; and its capacity to simplify LSTM structures by removing forget gates, which reduces computational expenses and the number of parameters without significantly impacting performance.

The highest measured interaction between hyper parameters is also relatively small. [47]. In order to learn more intricate patterns from sequential data, LSTMs are stacked to produce deep LSTM networks. These networks are then combined with other neural

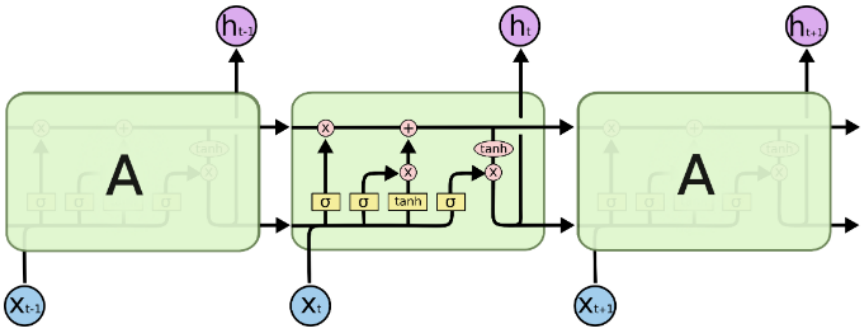

Figure 2.10: Architecture of LSTM (picture credited ([6]))

network designs, such as convolutional neural networks (CNNs) for image captioning [17] and VQA [42, 19, 43].

Chapter 3

Literature Review

3.1 Introduction

This section covers the research, journals, and papers that have been done to identify hateful content in textual, visual, and multi-modal data in general, and for the Amharic language in particular. Our review also focuses on research and discussion in the field of multi-modal problems, including methods or techniques and their accomplishments.

3.2 Unimodal Hateful meme Detection

Multimodal, textual, and visual data are all built together on social media platforms. Hate speech detection methods that are unimodal in nature can only take into account textual, visual, or other unimodal data. On the other hand, multi-modal hate speech detection combine multiple unimodal data. There has been a lot of work in recent years on detecting hate speech [32, 48, 37, 33, 34, 49] and natural language processing. Several text datasets have been released, mostly based on face book and various architectures have been proposed for classifiers. Almost all proposed hate speech detection approaches could be lexicon-based, machine learning and deep learning.

The lexicon-based approach to hate speech identification relies on the use of pre-existing lexicons or keywords. To improve the identification accuracy of hate speech, the author in [32] enriches the original dataset with textual emotional information and employs lexical baselines. Using the semantic and subjectivity features for hate speech classification along with a feed-to-classifier model that employs sentiment analysis techniques, the work of [29] investigates the idea of developing a classifier that is used to detect the presence of hate speech in online discourses such as web forums and blogs. Hate speech is categorized into three thematic areas: race, nationality, and religion. The words of a statement are compared to definitions found in dictionaries to ascertain whether or not they contain hate.

Subjectivity and semantic aspects of hate speech were considered in the creation of a lexicon that is used to construct a classifier for hate speech identification. Results employing a hatred corpus show great relevance for actual online conversation. Similarly the work in [30] use dictionary which contains the information about the polarity of each word done by incorporating an offensive lexicon composed of implicit and explicit offensive and swearing expressions that annotated with binary classes for European and Brazilian Portuguese. The study in [31] also use a dictionary-based approach to classify racism and not racism in the Dutch language, through 6,375 data sets for train and test; they trained multiple Support Vector Machines, using the distribution of words over the different categories in the dictionaries as features.

To solve the limitation of traditional hate speech detection approaches with respect to the ever increasing volume of social media data, and it becomes difficult to collect, store and analyze through traditional detection methods, researchers find alternatives. The use of an Apache spark open-source application model [33] was on of technique with the benefit of suitability on big data and it includes module for feature selection and machine learning. The work in [33] uses random Forest (RF) and Naïve Bayes (NB) for learning and Word2Vec, and TF-IDF for feature selection. The finding shows that word2vec embedding can outperform the best with 79.83% accuracy. However, expanding the classification category with different aspects of hate and including other sources, therefore, increase the information gain to improve performance of the model. Research work [34] has been done to determine which algorithms for machine learning and feature extraction methods work best together to create an Amharic detection model. The dataset collected from Facebook public page are manually annotated into three classes, and then transformed into a binary class in order to construct binary and ternary datasets, and an experimental strategy was used to ascertain the optimal configuration of the machine learning algorithm and feature extraction for models.Using word unigram, bigram, trigram, combined n-grams, TF-IDF, and combined n-grams weighted by TF-IDF and word2vec for both datasets, the SVM, NB, and RF models are trained on the entire dataset. For both binary and ternary models, the SVM plus

word2vec models outperform the NB and RF models by a small margin.The classification performance result indicates that ternary models outperform binary models in terms of achieving less confusion between hatred and non-hate. The performance of binary class detection models is generally inferior than that of hate speech detection using machine learning and text feature extraction techniques based on a multi-class dataset.
In addition to this the work in [47] has been done to determine the optimal fusion between the notion of sentiment analysis and the machine-learning algorithm, features extraction models, and current comparative studies. Applying the Word2Vec based SVM classifier to a sample of four classes yields good accuracy (0.72). Since deep learning models were introduced and demonstrated their capacity to manage long-term dependencies, many researchers have turned to them to address the issue of vanishing gradients that sometimes arises during the training of standard RNNs.
Computers can now learn and carry out tasks that are inherently human thanks to a sort of machine learning called deep learning. They can be learned in supervised, unsupervised, or semi-supervised ways, just like machine learning algorithms. Because deep learning models rely on neural network classifiers with extensive knowledge, they perform exceptionally well in analytics related to text and tasks that detect hate speech now days.
It makes an actual effort to recognize patterns in the text that is given and aims to replicate the event in layers of neurons. The optimal neural network algorithm, its hyper-parameters, and feature representation methods determine how well deep learning models perform. Some other works such as [48, 37, 49] are also done through deep learning algorithms with a textual dataset and can achieved remarkable results. The work [37] studies the significance of sub-word and context information for Amharic hate speech identification on social media platforms. It does this by using deep recurrent neural networks to extract the context of the social media comment and quick text word embedding to extract the sub-word information..

The findings demonstrate that, in comparison to using the word2vec feature, the pre-

cision of hate speech detection in Amharic was increased from 81.58% to 84.78% by utilizing a feature like Fast Text that can record sub-word information. Additionally, that incorporating context information also improves the accuracy of hate speech detection system than using just the target comments. The work[48] had done by using huge labeled Amharic dataset by collecting posts and comments from the selected Facebook pages of activists that participated actively. These Facebook data sets have been manually classified as free and hateful. Deep neural network models for automated hate speech identification are built in this research using Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) with word n-grams for feature extraction and word2vec to represent each unique word by vector representation. In order to train the model and determine the ideal combination of hyper-parameters for automated hate post detection, experiments are carried out on those two models, using 80% of the data set for training and 10% for validation. As a result, LSTM based RNN achieves better accuracy of detect posts as hateful or free.

Works like [49] also done to see the advantage of deep learning over traditional techniques.Here, word2vec embedding is used to build fresh Amharic hate speech datasets from Facebook and Twitter social media that are divided into four classes as a feature extraction technique for deep learning models. Using the Word2vec embedding feature and the embedding layer to automatically generate features, CNN, LSTM, Bi-LSTM, GRU, and combined CNN-LSTM models were trained on the entire dataset. The models were then evaluated using an 80–20 train-test split. A slightly higher F1-score of 90% is achieved by the BILSTM-based word2vec model compared to the other models, while an F1-score of 89% is achieved by the CNN classifier.

3.3 Multi-modal Hateful Meme Classification

Social media users have been gravitating toward multimodal data in recent times due to the appeal of sharing ideas, knowledge, and opinions about specific topics, issues, events, or products while maintaining the interest of their audience. Similarly, hate-mongers use this new development to take advantage of the strategies and tools that researchers have created to identify hate speech. Many studies have been done to use text datasets for natural language processing and to assess hate speech on social media. Few studies in recent years have examined multi-modal information for identifying hate speech, in contrast to text-based analysis.

Language and vision impairments have become well-known in recent years. There has been a surge of interest in multimodal problems since 2015 such as cyber bullying in photo sharing networks[18], visual question answering (VQA) [19], memes sentimental analysis [50, 51], and recently hateful meme detection [3, 24, 21, 20, 25, 26, 27]. However studies in [52, 18] have shown that textual features that combined with image can significantly improve the image classification task. Unfortunately, they are still lagging on multimodal problems like visual question answering and hateful memes classification. The work [52] also done on multi-modal classification combined textual embedding and image features with an emphasis on natural language understanding to increase classification accuracy through the use of related metadata. Compared to benchmark results, there is a 1.56% improvement. Therefore, it is possible to enhance image classification through the use of external text features.

Another work in [18] also done to examine how posted photos with captions are used to identify instances of bullying in reaction to shared content. In order to do this, more than 3000 photos and user-generated comments from Instagram are used. A convolutional neural network that has been trained on image pixels and themes extracted from image captions are two of the novel features that are used. Finding out how helpful these innovative features are at identifying cyberbullying in comments that have been posted is the goal. Similarly the work in [52] created a multi-modal dataset by introducing a dataset of Instagram photos paired with captions.

To comprehend the meaning, context, and semantics of each of the 1.3K Instagram posts, an annotation utilizing three orthogonal taxonomies are applied. The work in [50] also done to analyze sentiment in memes using VGG19, a language model that has been pre-trained on ImageNet datasets, to learn the combined visual and textual features to the OCR-extracted text in order to generate predictions. The proposed approach performs better than the baseline multi-modal and independent uni-modals based on either text or images. Similarly work [51] also done to categorize memes using textual features—extracted through optical character recognition (OCR)—and visual features. Memes are cultural units of style and information that spread among social media users with minor alterations. These features are then used to classify memes as negative, positive, and neutral. Hate speech on social media has been extensively studied.

Only a few articles have included both imagine and text modes, which is unexpected given a lack of study on multimodal hate speech. A multimodal visual-linguistic problem is hateful meme detection [27]. The research works like[3, 24, 21, 20, 25, 26, 27] has been conducted to analysis the problem of hateful meme detection by considering the joint features of textual features and visual features.

The work in [3] has been done by using torch vision model to extract visual features and fast text to extract the textual features from the meme and concatenate them to form a multimodal hateful meme detector. The work in [25] carried out to examine the impact of combining text and image embedding information. Various approaches, such as basic concatenation, bilinear transformation, attention, and gated summation, were used, and the results demonstrated how quickly performance could be improved by combining text and image embedding information. The author in [24] proposed an approach to identify hateful memes on the internet by taking into account both integrated use of textual and visual data. The suggested method attains an accuracy of 0.765 on the challenge test set and an AUROC of 0.811 on the test set. However, since most multi-modal baselines benefit the hate speechmore than one another modality, the uni-modal priors present a significant classification challenge in multi-modal hate speech detection approaches.

In order to address these issues, it has been decided to include the uni-modal sentiment to enrich the features rather than relying solely on multi-modal representations derived from neural networks that have already been trained [21]. The work in [20] has been done by preparing biggest hate speech datasets, composed of multimodal data, formed by image and text and trained different textual, visual and multimodal models to explore how every one of the inputs contributes to the classification and to prove that the proposed model can learn concurrences between visual and textual data useful to improve the hate speech classification results on multimodal data and despite the fact that images are useful for hate speech detection; the multimodal models do not outperform the textual models.

On the other hand, work in [26] compiled a dataset of 5,020 memes to evaluate and train the proposed modal over the individual or combined language and visual representations. The findings show that language in memes can be considerably less informative for identifying hate speech than visual modalities. This work involves assumptions about the reasonableness of the computing cost of encoding the retrieved text and performing optical character recognition (OCR). Study[27] has been achieved by combining created visual features, object tags, and text features of memes that are retrieved using optical character recognition (OCR) technology with a Vision-Language. The model can attain an average accuracy of 0.684 and an AUROC of 0.768 once it has been modified and linked to the classifier.

When it comes to fusion-based methods, several of them have employed various strategies based on early, late, and hybrid fusing of multimodal characteristics from text and images. [24, 20, 26]. Most of these VL features are extracted from unimodal pretrained models such as: BERT [27], VisualBERT [21] for the linguistic features, and VGG-16 [26], CNN [20] and ResNet-50 [52] for the visual features. On the other hand, alternative strategies have chosen to employ multimodal models that have already been trained, utilizing joint multimodal information (text and image) [52, 25], since their training takes into account both modalities (text and image) concurrently, allowing for a better alignment between them, the latter have demonstrated superior performance in

the detection of hateful multimodal[24, 25].

To extract text from image Google Vision API Text Detection module [20] and Tesseract 4.0.0 OCR [26] can be employed. Generally hateful meme detection is in infancy stage especially for under resourced language such Amharic. There different research works which conducted to Amharic language they use text as datasets for their study. Using Amharic memes is new and exponentially growing trend. Unlike the previous works on Amharic hate speech detection Amharic hateful meme detection is complex task as it required being deeper and complex approaches. In addition to this unlike to resourceful languages such as English, Amharic is morphological complex and have high similarity between characters this by itself another challenge to use extracted texts from memes image.

Chapter 4

Proposed Approach

4.1 Introduction

In this section, we describe proposed methodology architecture for Amharic hateful memes detection; the proposed approach includes seven basic processes. Each step in the proposed approach is mentioned in detailed below.

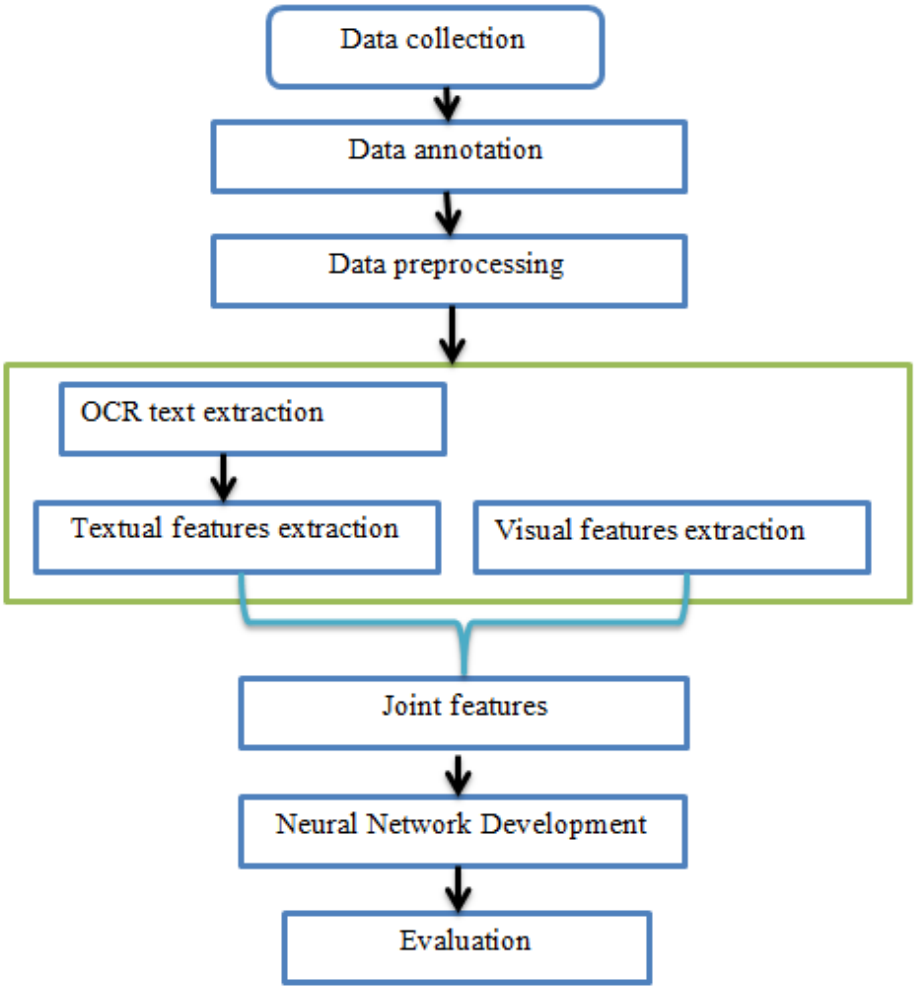

Figure 4.1: Overall structure of Proposed Methodology.

4.2 Data Collection

The first and most important step in developing the detection of hateful memes in Amharic is data collection, which includes gathering user-generated content from social

media platforms like Facebook, Twitter, Telegram, and others. Therefore, we manually gathered 6590 memes from Facebook and Telegram for our experiment.

Total number of memes	Hateful Memes	Non-hateful Memes
6590	3322	3268

Table 4.1: Collected datasets

4.3 Data Annotation

Data annotation is a process of labeling any type of data with in different class. Beside data collection, data annotation is one of the most important elements for the development of hateful meme detection. There isn't any standardized annotated data for Amharic hateful meme detection because the field of hateful meme detection is still in its infancy. Memes in Amharic can be found on various social media sites, but nobody utilizes them or labels them properly for instruction. Our methodology is based on two core components: text retrieved from memes for textual analysis and images for visual analysis. Every meme that is associated with a class needs to be labeled in order for our meme-modal to be trained. Making ready labeled meme datasets is therefore one of the most important responsibilities. In this study, five individuals annotated the collected material into two categories: hate and non-hate, reflecting a diversity of genders, races, and religious beliefs.

4.4 Data Pre-processing

Image preprocessing is quite useful and major step to improve the quality of images to analyze more effectively. It allows eliminating unwanted distortions and improving specific qualities that are essential for the application we are working on. We do some preprocessing activities to feed memes image to our models such as image scaling, noise removal and normalization. To maximize the quality of images to OCR system, the following image preprocessing steps are done for our study.

Noise Reduction: Noise in an image can be caused by various factors such as low light, sensor noise, and compression artifacts. Noise reduction techniques aim to remove noise from the image while preserving its essential features [28].

Normalization: is an essential preprocessing step that helps to segment the text from the background and increase the contrast between the characters and the background[82]. It transforms the input image into a binary format that enhances the visibility of the characters and makes them more easily recognized by the OCR system.

Image resize: Image resize in deep learning refers to the process of changing the size of an image [53]. This is done in order to ensure that the image is a suitable size for the deep learning model being used, while maintaining the integrity of the images original content and it helps to improve the quality and consistency of data.

4.5 Jointed features

Our classification model takes meme as input, analysis and extracts the visual and textual features. The output vectors from the two separate modalities are combined using concatenation techniques specifically early fusion method with the advantages of better task accomplishment [54] by utilizing the correlation between multiple features from different modalities at an early stage, provides a richness of information from heterogeneous data [3], and requires only one learning phase on the combined feature vector to obtain the required meme feature representations.

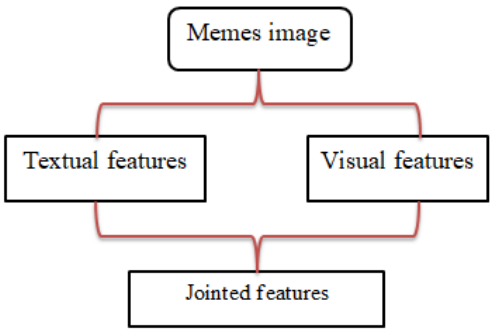

Figure 4.2: Jointed features

4.6 Neural Network Model Development

In this step we create deep neural network combination of LSTM and VGG 16 approaches. These are an improved version of the recurrent neural network and convolution neural network respectively. It is capable of tackling a variety of problems and providing robust solutions with combination of one another in different visual-linguistic tasks, such as visual question answering[42, 19, 43], image captioning[5].

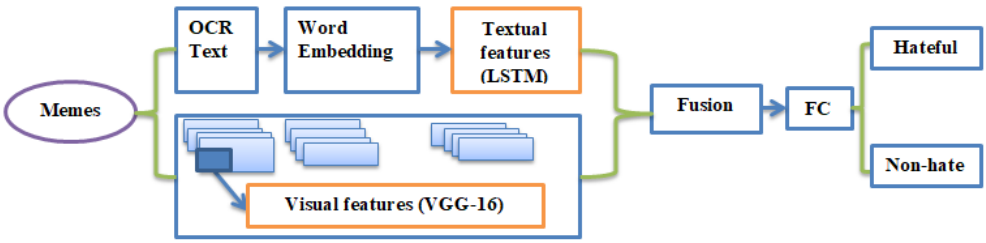

Figure 4.3: Memes Classification Model

4.6.1 Image text extraction

Our model works on two features, one is image for visual analysis and the other is extracted-text form image for textual analysis. To extract texts from the memes we used tesseract-OCR. Tessaract-OCR is an LSTM [6] network based OCR engine which helps in detection and recognition of texts embedded in memes image.

4.6.2 Text Feature Extraction

As word input representations, the OCR-extracted text is represented using word embedding [1], an advanced method that has demonstrated exceptional performance across a range of tasks. The Amharic language's morphological complexity, the noise surrounding the meme image, the visual resemblance of the script, the inconsistent letter formation, the stroke, and the font variation could all lead to the OCR system extracting words that are incorrect or incomplete from the image. Instead of editing the text by hand to include missing or incorrect words, the received text data [39, 40, 41], By adding an extra embedding layer, whose embedding is kept fixed for simplicity, we

were able to extract the required features using the built-in fast-text library. One word embedding technology called Fast-text is utilized for word representation. Numerous natural language processing applications, including sentiment analysis, have seen successful application of the method [46]. In the word embedding model, a fixed number of dimensions can help to permit more efficient computations to better reflect restricted content in short texts. Long- and short-term dependencies in sequential data have been successfully captured by LSTM, a unique kind of deep neural network[6]. The word sequence is fed into the LSTM after each word is represented by a corresponding vector that was trained by the word embedding model. Next, brief texts' long-term word dependencies are trained into LSTM classifiers. The word embedding model encodes the text, and the LSTM, trained for hostile meme categorization, retrieves the most relevant information for this job.

4.6.3 Visual Feature Extraction

Since CNNs are made expressly to process images and carry out more complicated operations like segmentation, object identification, and image classification, they are typically the best option for extracting features from images. Convolutional layers make up this system, which uses features in the input space to be automatically identified. We used the VGG-16 architecture, which is family of CNN, pre-trained on the ImageNet classification dataset [4]. Then we used the activations of a hidden layer as feature vectors for the image, which has been proved to provide relevant information of the image, not only for ImageNet classification, but it can also be used for other tasks such as visual tracking.

4.6.4 Fusion

Concatenation is used to combine the features obtained from the textual and visual models. A one-dimensional prediction output from the concatenation layer is sent to the fully connected layer.

4.6.5 Fully Connected Layer

At the end, the single joint vectors from the concatenation layer are fed to the fully connected layer, which classifies the memes as hateful or non-hate.

4.7 Model Evaluation

After classification has finished, the performance of the classifier is evaluated using different performance metrics like recall, precision, F score and accuracy. Training and validation losses and accuracy are used for choosing appropriate epoch number to avoid occurrence of over fitting. The learning ability of different deep neural learning models trained on the Amharic hateful meme datasets was investigated and evaluated. The datasets gathered from various social media platforms would be fed into the proposed models, as well as various performance evaluation metrics selected for models.

TP (True Positive): The number of occurrences that are hateful and correctly predicted as hateful. FP (False hateful): The number of instances that are non-hate but are incorrectly predicted as hateful. FN (False Negative): The number of instances that are hateful but are incorrectly predicted as non-hate. TN (True Negative): The number instances those are non-hate and truly predicted as non-hate.

4.7.1 Precision

It is the percentage of the number of items labeled as a particular desired class, true positives (TP) to a total number of items labeled as that class.

$$Precision = \frac{TP}{TP+FP} \tag{4.1}$$

4.7.2 Recall

It is the total number of true positives distributed by total number of items that are known to belong to that class.

$$Recall = \frac{TP}{TP+FN} \tag{4.2}$$

4.7.3 F-score

The F-score is a metric used to evaluate the performance of a Machine Learning model. It combines precision and recall into a single score.

$$F - score = \frac{2Rp}{R+p} \tag{4.3}$$

F-score is a better and well-known measure to use if we need a balance between precision and Recall.

4.7.4 Accuracy

The additional performance measure of a model is accuracy. Accuracy measures how much accurately the model learns to classify the data.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.4}$$

Chapter 5

Experiments and Results

5.1 Introduction

In order to answer our research questions, we investigated the impact of using Amharic memes as features for automatic Amharic hate speech detection. We ran an exploratory experiment to compare the hate speech classifier performance of a labeled memes. To evaluate the effect of incorporating labeled memes, we used the labeled memes, as an input for CNN-LSTM. These experiments answer RQ1. Can Amharic hateful meme be used to enhance Amharic hateful meme detection on social media?. The second group of experiments aims to identify the contribution of each modal within Amharic hateful meme detection process. These experiments are conducted to see each modality contribution within Amharic hateful meme detection?

5.2 Development Tools and Packages

We use many development tools in this research. These are Tensor flow deep learning library, Keras deep learning library, Scikit learn machine learning library, OCR engine and python with Anaconda navigator.

Python: A programming language has been used for preprocessing the data and develops the model. We used Keras to construct and train the designed neural network model through the Tensor flow backend engine. Hence, Tensor Flow is an end-to-end platform that makes it easy for building and deploys deep learning models, used as a backend for Keras API.

Tesseract –OCR: ALSTM network based OCR engine, which helps in detection and recognition of texts embedded in an image. Tesseract –OCR extracts the embedded texts from the meme dataset and store the text in a comma separated value files for further operations.

A panda: An open-source library that provides high performance, easy to use data structure and data analysis tools for python programming language. A panda is used to read CSV files and perform different operations on the CSV files.

Keras : a library for developing deep neural networks in python that can run using TensorFlow as a back-end to process's data, to create, evaluate, optimize, fit and test a model.

Tensor Flow: is an end-to-end open-source platform that has flexible tools and resources for machine learning. Many states of the art NLP applications are developed using TensorFlow as a back-end. Popular organizations like Google, Intel, and others also use TensorFlow to develop systems. Generally, for doing all tasks we used Jupyter Notebook which is a web-based interactive computing notebook environment for edit and run python codes.

5.3 Experimental Setup

We use one personal computer for all experiments. Table 5.1 below shows the hardware and software specification of the machine used in all experiments.

Manufacturer	Intel Core i5-4300U
Model	HP ProBook
Processor	IntelCore i5-3320MCPU 2.60GHz,2601MHz
Memory	8 GB
Operating System	Ubuntu 20.04 LTS

Table 5.1: Material requirement for experimental setup

5.4 Data set description

The table 5.2 shown the dataset with the label of class. we created meme datasets written in Amharic language, which contain 6590 memes data that labeled into two classes as hate and non-hate.

Class	Number of memes
hateful	3322
Non-hate	3268
total	6590

Table 5.2: Class distribution of meme datasets

5.5 Data Splitting

One of the most important phases in creating and evaluating a neural network classifier model is data splitting. In this stage, 90% of the annotated data are used for training and 10% are used to test the neural network.

5.6 Evaluation Metrics

A popular method of assessing machine-learning models is to compare the model's anticipated outputs with human-labeled data. A model's performance can be assessed using a variety of assessment metrics, including F score, accuracy, precision, and recall. Following classification, the classifier's performance is assessed using various performance metrics, including accuracy, recall, precision, and F score.

5.7 Model Configuration

To create hateful meme detection models, we need to configure the deep learning network parameters in addition to the hardware and software tools and their requirements. In order to choosing the ideal parameters that the learning algorithm will use to discover the ideal parameters that accurately map the input features to the labels or targets is the process of training a model. The dropout number is employed to choose the suitable era. Over-fitting happens when a model learns training datasets too thoroughly, and it is crucial to identify this phenomenon in neural networks during training. About the training data that needs to be remembered, it produces noise. It cannot, therefore, forecast an output for an input that has never been observed before. A model is said to be over-fit when it fits the training set too closely.

Under-fit refers to a model that is unable to accurately represent the training set or extrapolate to novel data. When a machine learning model performs poorly on training data, it is under-fit and will be readily apparent. Smaller datasets are more likely than larger datasets to exhibit under-fitting.

Figure 5.1: Training vs validation accuracy

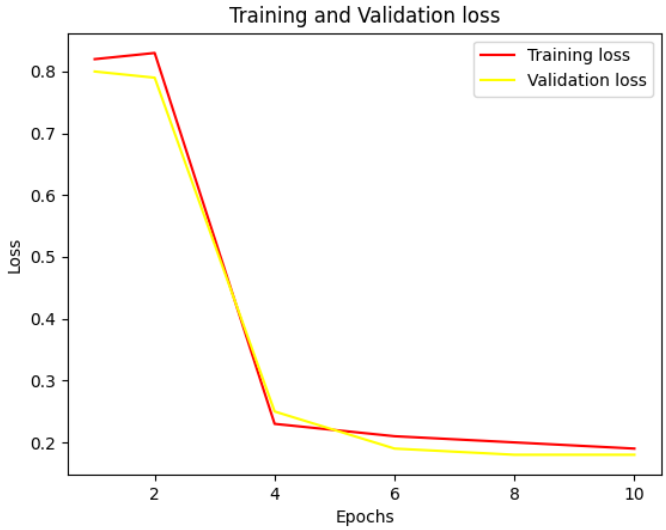

Figure 5.2: Training vs validation loss

The graph shows that the model's training accuracy rises and its validation accuracy falls as it is trained past epoch eight. We can conclude that the model is over-fitted as a result, and to see the impact of this, we will set our epoch number to eight. When

both training and validation loss simultaneously decreases. However, at some point of time validation loss increases while training loss decreases. When both training and validation accuracy starts increasing and at some point when validation accuracy increases over training accuracy. This is one of the conditions to say the model is over fitted.

5.8 Results and Discussions

We compare our model with the uni modal based on textual and multi modal (visual and textual) baseline techniques for a thorough assessment of Amharic meme detection.

5.8.1 Contribution of each Modality to Amharic Hateful Meme Detection

One of the interesting aspect of the Amharic hateful meme detection is to measure and analyze how much does each modality contribute to the overall of solving the problem. In this part, we examine how Amharic Hateful Meme Detection model performs within the contribution of each Modality. We used the image version of text terms datasets for hateful meme detection and we didn't change the class polarity on the datasets to show the contribution of each modality. Our findings suggest that visual signals are significantly more important than linguistic ones in the identification of hostile memes. The reason for the better performance of the vision-only model compared to the text-only model could be a discrepancy in capacity for encoding data from datasets and there may be a visual bias in the datasets. Since, it benefits from the CNN's ability on the way to extract features with image. CNNs might have convolving filters over each input layer to generate the best features form visual than text generate form OCR, and it has been proven that CNN is a powerful tool for selecting feats for visual. Memes frequently contain heavily distorted, highly compressed images, this could have an impact on the accuracy of OCR detection and, consequently, language encoding.

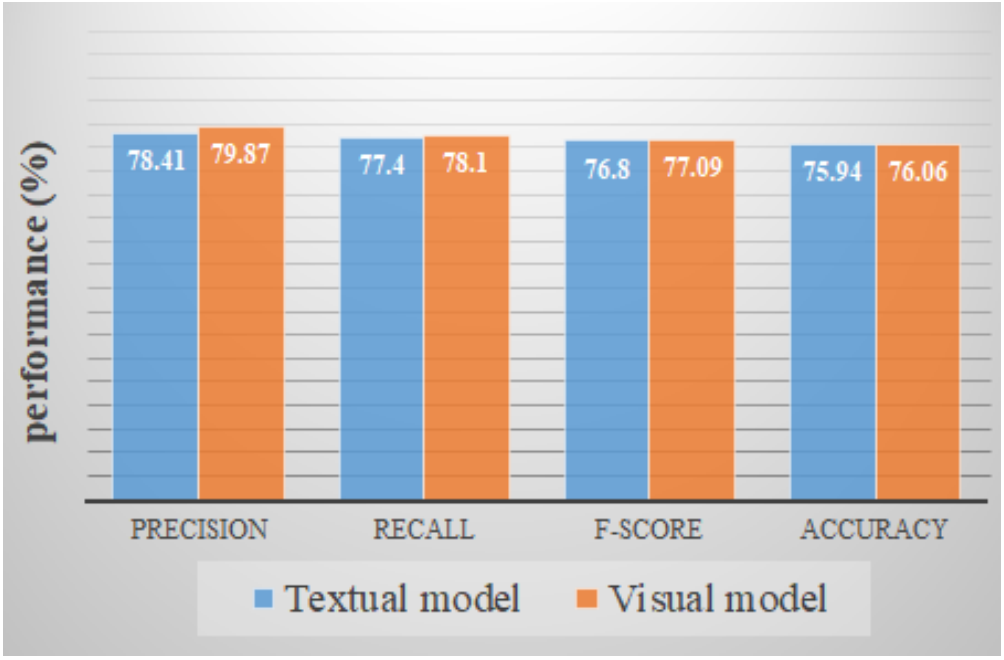

Figure 5.3: Performance of Contribution of textual and visual features to Amharic hateful memes classifier.

Therefore, it will be crucial to enhance the textual model and the text's contribution to the multi-modal model.

5.8.2 Effect of using Memes for Amharic Hateful Meme Detection

In order to answer the research question, we use Amharic memes as features to the combined Visual Geometric Group (VGG-16) and Long Short-Term Memory (LSTM) model using 6590 data size. To answering RQ1: Can Amharic hateful meme be used to enhance Amharic hateful meme detection on social media?. The input for the Amharic hateful meme detection classifier is combination of textual and visual features. The 6590 memes collection has been divided into test, validation, and training datasets. Data statistics are displayed in table 5.2. A model combining the two targeted modalities has been implemented using early fusion method. The outcomes of the meme categorization results are displayed in figure 5.4. The maximum precision in the cruel classification of hateful meme offered by the meme shows its ability to recover the hateful memes. However, the precision of 1.6 indicates that a large number of memes are correctly classified as hateful.

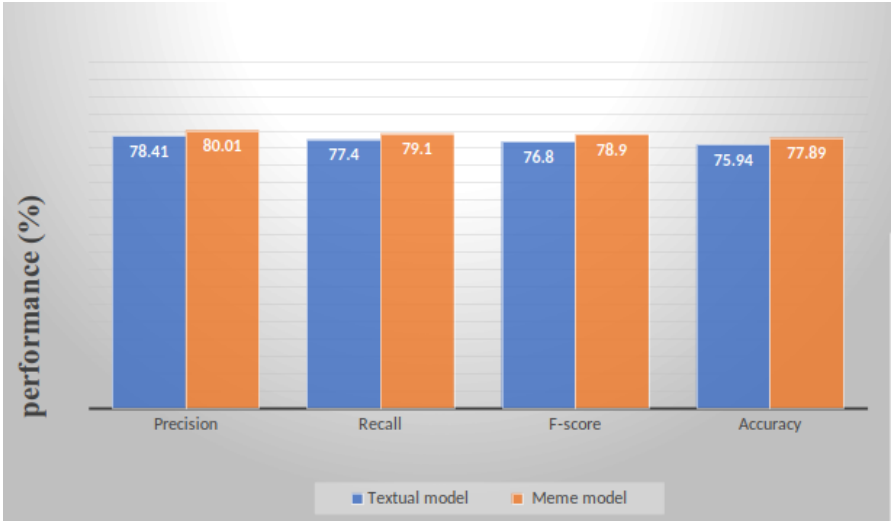

Figure 5.4: Performance of CNN-LSTM hateful memes classifier using combination of textual and visual features

It returns better results for memes in this domain. Figure 5.4 demonstrates that when text and image features are taken into account, the meme model based classifier performs better in terms of f-score 78.9%. The outcome in a balanced F1-score that preserves the propensity to increase precision without decreasing recall. A fascinating feature regarding the multi-modal classifiers is displayed in figure 5.4. Using the combined power of several classifiers, an ensemble model might be constructed to detect hateful content. In the meantime, we have seen the impact of optimizer, learning rate, dropout, and activation function for the meme-model classifier after determining the model. In the end, 0.001 learning rate, 8 batch size, Adam optimizer, 0.25 dropout, and softmax activation register 75.94% accuracy for the text model and 77.89% accuracy for the meme model. In most cases, we used a social media meme for this work, but a variety of memes from many domains can be used to eliminate the biases brought about by the use of a certain domain. Fusing the text and image features for meme representations might be more efficient. Since hostile material is difficult to express, more training data will help us better understand the abstract properties that underlie hateful content recognition.

The latest papers combine these two distinct modes to address not only the gap between two feature extractor mechanisms, but also the challenge of integrating the noisy nature of OCR extracted text data due to its importance into a supervised learning framework for the task of Amharic hateful meme detection. This fact might have an impact on the level of language encoding and OCR recognition quality. Therefore, in order to create the word vector representation, the learning algorithm will need to create words from the vocabulary. We use the built-in fast-text library to combine the extracted text with the embedding layer in order to extract the necessary features.

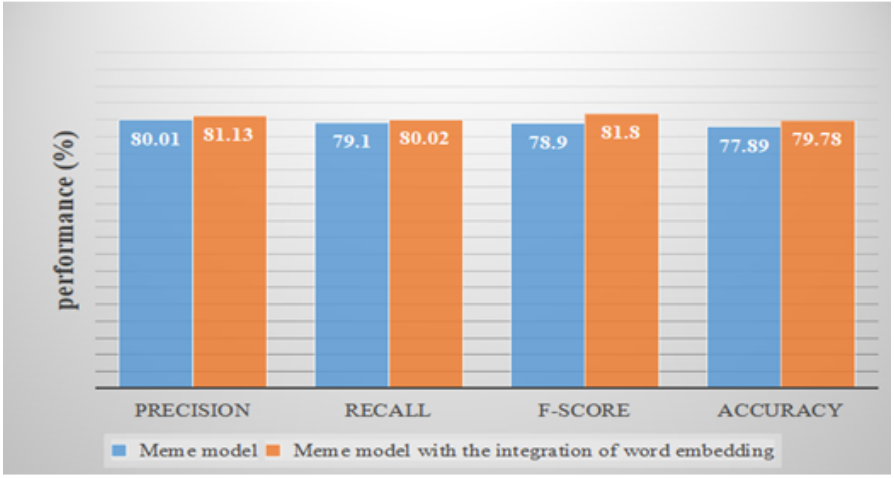

Figure 5.5: Performance of CNN-LSTM hateful memes classifier with the integration of word embedding.

As can be seen from the results in figure 5.5 above, Amharic texts now have the feature of word embedding based on similarity. Additionally, the experiments demonstrate that when identifying hateful memes, memes with word embeddings are far more significant than meme model. Although multi modal approach yields the best results, word embedding gain is achieved through integration. In order to store the morphological information in the embedding space, we employed the automatic feature extractor for word embedding, which slides on a word's variable n-gram character. From the figure 5.5, it is quite clear that the meme model with the integration of word embedding performed better (around 2.9%) on F-score than the meme modal.

Chapter 6

Conclusions and Future Works

6.1 Conclusion

We aimed to demonstrate how the addition of textual and visual elements to the classifier for hate speech classification impacts the development of an improved hateful meme detection system for Amharic. In this work, we used OCR text based on the Amharic language to generate memes and neural embedding to capture the morphology and semantics of specific words in a given challenge. The use of neural embedding of text contains as a feature for automatic extraction from OCR and a hate speech detection classifier that combines textual and visual features has a good. We describe the outcome of a model that combines convolution with recurrent neural networks. This model performs well because it takes advantage of CNN's feature extraction capabilities. On the other hand, long-term analysis of the text's directional dependencies, both past and present, is distinctive of LSTM. In conclusion, the experiments conducted demonstrate that the performance of Amharic Hateful Meme detection is positively impacted by the combination of textual, visual, and embedding elements in Amharic hate speech detection.

6.2 Recommendation

The studies we have conducted on public opinion participation in hate speech identification have some limitations. Our findings also presented the following problems, which could be useful for guiding future study.

1. It would be very helpful to look closely at how combining text with related audio and video could improve the analysis of hate speech identification performances.

2. It would be quite helpful to thoroughly examine the effects of the textual alignment with horizontal and vertical.

3. For more difficult tasks, enormous amounts of labeled data as well as fresh unlabeled datasets are needed, particularly for morphological complex languages with memory-efficient models that require less resources.

4. It is necessary to develop further hybrid and ensemble ways to increase performance.

Bibliography

[1] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, and J. W. Kim, "Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism," *Applied Sciences*, vol. 10, no. 17, p. 5841, 2020.

[2] G. Sahu, "Adaptive fusion techniques for effective multimodal deep learning," Master's thesis, University of Waterloo, 2020.

[3] G. Darshan, K. Deepak, and G. Suresh, "Hateful meme detection for social media applications," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, pp. 1606–1609, 2021.

[4] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, pp. 143–150, 2019.

[5] LSTM, "Lstm-vgg-16: A novel and modular model for image captioning using deep learning approaches," *Journal Engineering Science*, 2018.

[6] L. S.-T. Memory, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 2010.

[7] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa, "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets," *The Visual Computer*, pp. 1–32, 2021.

[8] M. Šimpachová Pechrová and V. Lohr, "Efficiency of social media communication of czech universities and faculties," 06 2016.

[9] S. Neshkovska and Z. Trajkova, "The essentials of hate speech," *Teacher*, vol. 14, no. 1, pp. 71–80, 2017.

[10] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the fifth international workshop on natural language processing for social media*, 2017, pp. 1–10.

[11] I. Gagliardone, M. Pohjonen, Z. Beyene, A. Zerai, G. Aynekulu, M. Bekalu, J. Bright, M. A. Moges, M. Seifu, N. Stremlau, *et al.*, "Mechachal: Online debates and elections in ethiopia-from hate speech to engagement in social media," *Available at SSRN 2831369*, 2016.

[12] M. A. Adamu, "Role of social media in ethiopias recent political transition," *J. Media Commun. Stud*, vol. 12, no. 2, pp. 13–22, 2020.

[13] EIP, "Fake news misinformation and hate speech in ethiopia: A vulnerability assessment," *European Institute of Peace* |, 2021.

[14] Z. Mossie and J.-H. Wang, "Vulnerable community identification using hate speech detection on social media," *Information Processing & Management*, vol. 57, no. 3, p. 102087, 2020.

[15] K. M. Yilma, "Cybercrime lawmaking and human rights in ethiopia," *Mizan Law Review*, vol. 15, no. 1, pp. 73–106, 2021.

[16] Nagaret, "Hate speech and disinformation prevention andsuppression proclamation," *FEDERAL NEGARIT GAZETTEOF THE FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA*.

[17] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023.

[18] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, and C. Caragea, "Content-driven detection of cyberbullying on the instagram social network." in *IJCAI*, vol. 16, 2016, pp. 3952–3958.

[19] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, "Visual question answering using deep learning: A survey and performance analysis," in *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*. Springer, 2021, pp. 75–86.

[20] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1470–1478.

[21] A. Das, J. S. Wahi, and S. Li, "Detecting hate speech in multi-modal memes," *arXiv preprint arXiv:2012.14891*, 2020.

[22] Y. Zhou, Z. Chen, and H. Yang, "Multimodal learning for hateful memes detection," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[23] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? the challenging case of long tail on twitter," *Semantic Web*, vol. 10, no. 5, pp. 925–945, 2019.

[24] R. Velioglu and J. Rose, "Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge," *arXiv preprint arXiv:2012.12975*, 2020.

[25] F. Yang, X. Peng, G. Ghosh, R. Shilon, H. Ma, E. Moore, and G. Predovic, "Exploring deep multimodal fusion of text and photo for hate speech classification," in *Proceedings of the third workshop on abusive language online*, 2019, pp. 11–18.

[26] B. O. Sabat, C. C. Ferrer, and X. Giro-i Nieto, "Hate speech in pixels: Detection of offensive memes towards automatic moderation," *arXiv preprint arXiv:1910.02334*, 2019.

[27] Y. Chen and F. Pan, "Multimodal detection of hateful memes by applying a vision-language pre-training model," *Plos one*, vol. 17, no. 9, p. e0274300, 2022.

[28] A. T. Birhanu, "Amharic character recognition system for printed real-life documents," Ph.D. dissertation, Addis Ababa University, 2008.

[29] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.

[30] F. Alves Vargas, F. Rodrigues de Góes, I. Carvalho, F. Benevenuto, and T. Alexandre Salgueiro Pardo, "Contextual lexicon-based approach for hate speech and offensive language detection," *arXiv e-prints*, pp. arXiv–2104, 2021.

[31] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "A dictionary-based approach to racism detection in dutch social media," *arXiv preprint arXiv:1608.08738*, 2016.

[32] R. Martins, M. Gomes, J. J. Almeida, P. Novais, and P. Henriques, "Hate speech classification in social media using emotional analysis," in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 2018, pp. 61–66.

[33] Z. Mossie, J.-H. Wang, *et al.*, "Social network hate speech detection for amharic language," *Computer Science & Information Technology*, pp. 41–55, 2018.

[34] K. Yonas, "Hate speech detection for amharic language on social media using machine learning techniques," Ph.D. dissertation, ASTU, 2019.

[35] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[36] M. E. Nogales, Rubén E. Benalcázar, "nalysis and evaluation of feature selection and feature extraction methods," *International Journal of Computational Intelligence Systems*, 2023.

[37] B. Girma, "Hate speech detection using deep recurrent neuralnetworks for amharic text," *URI: https://repository.ju.edu.et//handle/123456789/6826Date: 2021-01-22*.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[39] K. Kukich, "Techniques for automatically correcting words in text," vol. 24, no. 4, 1992. [Online]. Available: https://doi.org/10.1145/146370.146380

[40] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of post-ocr processing approaches," vol. 54, no. 6, 2021. [Online]. Available: https://doi.org/10.1145/3453476

[41] Y. Hu, X. Jing, Y. Ko, and J. T. Rayz, "Misspelling correction with pre-trained contextual language model," in *2020 ieee 19th international conference on cognitive informatics & cognitive computing (icci* cc)*. IEEE, 2020, pp. 144–149.

[42] S. S. N. Mohamed and K. Srinivasan, "Imageclef 2020: An approach for visual question answering using vgg-lstm for different datasets." in *CLEF (Working Notes)*, 2020.

[43] R. Nisar, D. Bhuva, and P. Chawan, "Visual question answering using combination of lstm and cnn: a survey," *IRJET e-ISSN*, pp. 2395–0056, 2019.

[44] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

[45] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[46] J.-H. Wang, T.-W. Liu, X. Luo, and L. Wang, "An lstm approach to short text sentiment classification with word embeddings," in *Proceedings of the 30th conference on computational linguistics and speech processing (ROCLING 2018)*, 2018, pp. 214–223.

[47] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[48] S. G. Tesfaye and K. Kakeba, "Automated amharic hate speech posts and comments detection model using recurrent neural network," 2020.

[49] B. EMUYE, "Amharic text hate speech detection in social media using deep learning approach," Ph.D. dissertation, 2020.

[50] R. R. Pranesh and A. Shekhar, "Memesem: a multi-modal framework for sentimental analysis of meme via transfer learning," in *4th Lifelong Machine Learning Workshop at ICML 2020*, 2020.

[51] E. Smitha, S. Sendhilkumar, and G. Mahalaksmi, "Meme classification using textual and visual features," in *Computational Vision and Bio Inspired Computing*. Springer, 2018, pp. 1015–1031.

[52] S. J. Miller, J. Howard, P. Adams, M. Schwan, and R. Slater, "Multi-modal classification using images and text," *SMU Data Science Review*, vol. 3, no. 3, p. 6, 2020.

[53] M. A.-M. Khan, S.-H. Kee, A.-S. K. Pathan, and A.-A. Nahid, "Image processing techniques for concrete crack detection: A scientometrics literature review," *Remote Sensing*, vol. 15, no. 9, p. 2400, 2023.

[54] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.

Appendix

ይሄ ነው ከንቲባው ይሄ ነው!!!

(e) Non-hate

የቤት ስራ ሳትሰራ ስትቀር መምህሩ

(f) Non-hate

የአብዛኛው ሰው መታወቂያ

(g) Non-hate

የጎጃም ደብተራ እዬሱስን ሰቀበል

(h) Hateful

ለስራ ቅጥር ስታመለክት

ከ ዕዳ ወደ ምንዳ ስልጠና መስደሃል?

(i)Non-hate

የተከበሩ የአዲስ አበባ ከንቲባ

(j) Hateful


አቢቹ ጭምብሉን ሲያወልቅ
FOSAR
የጀሞ ቺክ ጠብሰህ baby ስትላት
eyobvb
ማንን ነው ህፃን አንተ
በአንድ ማትስ በሁለት ፊዚክስ ከሆነ
@Neba
ትምህርት ቤት የምሄደው አካሄድ
father ኢሳት እያየ በተመሳሳይ ሰአት
ETV ላይ የአብይ መግለጫ ሲጀምር
በመብራት ሀይል ሰራተኞች ላይ
ሙሉ እምነት ሲኖርክ:-
ZOL
ለመጨረሻ ጊዜ ልጠይቅሽ
ትክክለኛ እድሜሽን ንገሪኝ
ግ...ደ...ለ...ኝ!!


መንግስ የኑሮ ውድነቱን እያረጋጋበት
ሁኔታ
30 ያለፋትን ቺክ ለጥምቀት ሎሚ
ይሻልሻል ብርቱካን ስትላት
Bina
ምሳ ሰዓት ደርሷል ማለት ነው
ዛሬ ደሞ እንዴ ከወለድ ነፃ ባንክ ብሩን
ከቶ አልወልድ ሲለው ምን ቢል ጥሩ
ነው...
ባንኩንም ወሊድ መከላከያ አስዋጡት
እንዴ😂
አረብ ሀገር ውስጥ የሆነ ሬስቱራንት
ገብተ ኮካ ስታዝ
Coca-Cola
ትኩስ ድንች ወጥ እንደጎረስክ
እንኪ ጁስ እያቴ...
ገንዘብ ሳይኖርህ የሴት ልጅ ልብ ውስጥ
ለመግባት ስትሞክር 😂

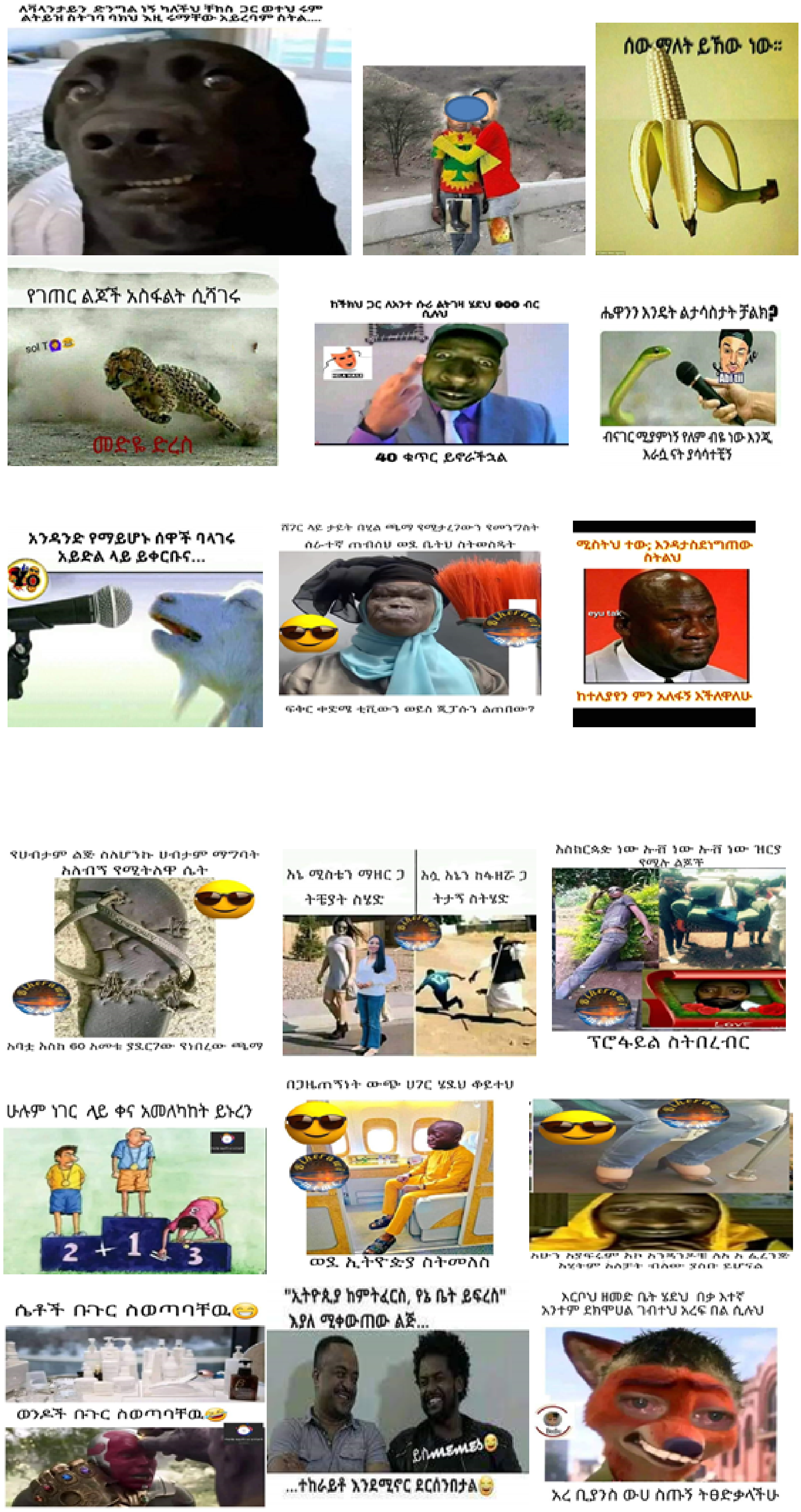
ሰው ማለት ይኸው ነው።
የገጠር ልጆች አስፋልት ሲሻገሩ
40 ቁጥር ይኖራችኋል
ሔዋንን እንዴት ልታሳስታት ቻልክ?
ፕሮፋይል ስትበረብር
ሁሉም ነገር ላይ ቀና አመለካከት ይኑረን
"ኢትዮጲያ ከምትፈርስ, የኔ ቤት ይፍረስ"

Figure 6.1: Examples of Amharic memes