## first

## ****Amharic Sentiment Analysis A Machine Learning Approach****

**Mirror Project**

### ****Submitted by:** Yeabsira Feleke Fanta**

### ****Roll Number:** 20222755**

### ****Date:** May 8,2025**

**Table of Contents**

**Abstract**

**1.Introduction  
2.1 Background  
2.2 Problem Statement  
2.3 Objectives**

**2.Methodology  
3.1 Data Sources  
3.2 Data Labeling  
3.3 Text Preprocessing  
3.4 Feature Extraction**

**4.Results and Discussion  
4.1 Experimental Setup  
4.2 ML Results  
4.3 DL Results**

**5.Model Deployment  
5.1 Strategy  
5.2 Challenges**

**6.Performance Evaluation  
6.1 Case Studies**

**7.Future Improvements  
7.1 Dataset Expansion  
7.2 Model Enhancements  
7.3 Multilingual Support**

**Conclusion**

**References**

****1.Abstract****

**Objectives**: This study aims to prepare a corpus and explore sentiment

analysis in the Amharic language, which is increasingly used due to the

growth of both the language and the Internet. ***Methods***: The study acquired 23,646 Amharic tweets from Twitter using the Twitter API, cleaned and normalized the text through preprocessing, and manually annotated the data as positive, negative, or neutral by three annotators. The study utilized a multi scale sentiment analysis approach to experimentally evaluate the classifier’s performance and compare different ML and DL classifiers. ***Findings****:* The study found that sentiment analysis in the Amharic language in this dataset showed that the KNN classifier could classify texts with an accuracy of 76% and 90% accuracy using the CNN deep learning classifier.***Novelty****:* This study contributes to the field of sentiment analysis by addressing the scarcity of an Amharic-language dataset specifically tailored for sentiment analysis purposes. Our approach involves filling this critical research gap by developing a new dataset. Subsequently, we employ machine learning and deep learning classifiers to assess the viability of this dataset for performing multi-class sentiment analysis tasks in the Amharic language.

**Keywords**: Amharic; Sentiment Analysis; Multiclass; Machine Learning; Deep Learning Classifier

## Introduction

Amharic, the Ethiopian language, is considered a member of the Semitic branch of the Afro-Asiatic language family(Salawu and Aseres, 2015). Amharic is one of the languages spoken and official working languages of the Ethiopian Federal Democratic Republic; it is also the language utilized by the majority of the population. According to the findings of the census carried out by Central StatisticalAuthority(‘Population-and-HousingCensus-1994-Addis-Ababa-Region’, no date), it has a total number of speakers that is close to 25 million, which places it in second place behind Arabic as the world's most widely spoken Semitic language. The volume of digital documents available on the Internet that are written in Amharic has been expanding at a rate that is considered to be extremely rapid. As a consequence of this, there is a huge demand in Amharic for sentiment analysis. The sub-field that falls under the umbrella of natural language processing (NLP)is- known as sentiment analysis, which can also be referred to as opinion mining, is used to determine the underlying sentiment of a piece of written text. This is a common method utilized by corporations to ascertain and classify opinions concerning a product, service, or concept. One of the Natural Language Processing (NLP) jobs with a wide variety of applications is this one. Some of the most common uses of sentiment analysis include decision-making support, business-related applications, prediction, and trend analysis, consumer feedback, social media monitoring, and customer service management(Mehta and Pandya, 2020). Analysis of sentiment is a natural language processing problem that has been receiving a growing amount of attention as of late. On the other hand, most of sentiment analysis studies have been carried out in the English language. The examination of Amharic mood has only been the subject of a small number of research, even though the language is promoted as being one of the most widely used on the Internet.

### 2.1 Background

With the growing importance of digital communication, sentiment analysis has become a crucial tool for understanding user opinions and emotions. While many languages have well-developed sentiment analysis models, Amharic—a widely spoken language in Ethiopia—lacks extensive resources for natural language processing (NLP).

### **2.2 Problem Statement**

Amharic, the official working language of Ethiopia, is also spoken in many other regions of the country. Sentiment Analysis is challenging to study in Amharic due to several difficulties. The aspects of phonetics, phonology, and morphology unique to Amharic contribute to the language's distinctiveness as a morphologically rich language.

There is a gap in sentiment analysis tools for Amharic, making it difficult for businesses, researchers, and policymakers to analyze public opinion effectively. This project aims to develop a machine learning-based **sentiment analysis model for Amharic**, which will classify text into positive, negative, or neutral sentiments.

### **2.3 Objectives**

Towards achieving the general objective of the study that deals with sentiment classification and subjectivity classification, the following specific objectives are formulated:

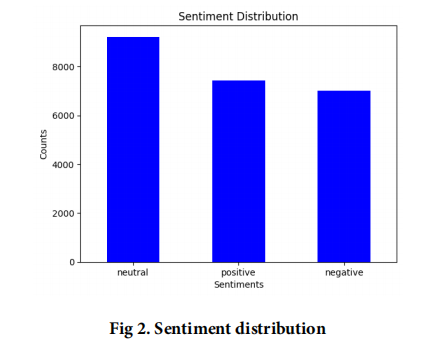
* To build both domain-specific and general-purpose corpus of Amharic language opinion terms where these terms are tagged as positive and negative.
* To develop the necessary algorithms to realize the proposed model in developing an Amharic sentiment analysis and the subjectivity classifier model.Train a **machine learning model** to classify Amharic text sentiment
* To Build a corpus from preprocessed data for sentiment and subjectivity analysis of the text of Amharic language.

* To Develop subjectivity classifiers using collected texts for training.
* To Develop a prototype to demonstrate that the model designed is valid.Deploy the model as a web application for public use.
* To evaluate the performance of the prototype designed in this study. Contribute to **Amharic NLP research** by making the dataset publicly available.

## ****3. Methodology****

### **3.1 Data Sources**

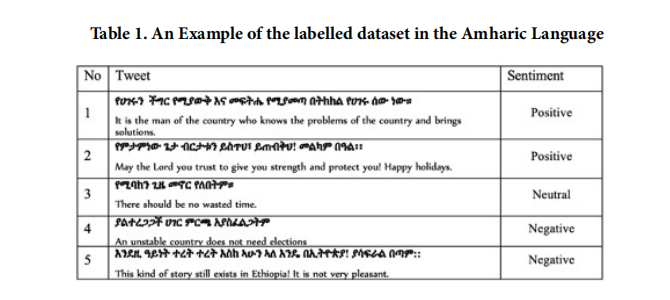
In this research, primary sources of information have been obtained from Twitter Amharic tweets using Twitter API, which is the first step in the emotion recognition process and involves collecting various views (comments) from Twitter(21) . In this investigation, we employed supervised machine learning and deep-learning techniques. Since supervised learning techniques require labelled datasets for instruction, we categorized the tweets by having three language experts categorize them as either Positvie ,strong positive, negative ,strong negative and neutral (Figure 2).



### 3.2 Data Labeling

We spent extra time with language experts to decipher Amharic text, and we labelled the tweets dataset into multiple classes (positive,very positive, negative,very negative and neutral) for the training phase. These groups are arranged according to the point which a word evokes a specific feeling when used in a sentence. In support of this hypothesis, consider the following Table 1. In Amharic, the sentence

in bold expresses a personal opinion, while the corresponding sentence in English is literal.



### 3.3 Text Preprocessing

**Text Preprocessing**

To further prepare the dataset for opinion mining, text preprocessing is performed. Text preprocessing is applying any calculation to unstructured raw material to change it into an arrangement that another procedure may process more efficiently(22) . This work’s proposed model architecture is split into a training and testing phase. While classifying the training

and test sets, the machine learning system required training data consisting of annotated 23,646 Amharic tweets, with 7424

classified as good, 9217 as neutral, and 7005 as negative. These results are a byproduct of the initial phases of tokenization, symbol removal, stop word removal, non-Amharic character removal, and normalization.

**Tokenization:** Preprocessing consists of two phases: tokenization breaks down a document or text string into sentences(23) .

The second, de-sentence fiction, transforms those sentences into individual words. This research proposes a method for analyzing the tone of tweets written in Amharic at the sentence level. Since processing at the sentence level was required, we tokenized the sentence’s text. The Amharic language uses different symbols like (Th, Th, Th, Th, Th, Th) question marks (?) and exclamation marks (!) as the English language to tokenize sentences into tokens. Tokenization separates a sentence into its component words based on the distance between consecutive terms in a sentence, making it possible to apply the next step, stop word removal, with minimal effort.

**Stop-word removal:**

Preprocessing also includes a phase called stop-word removal, which removes the most frequently occurring words in a text that has no bearing on the classification of attitudes. Certain Amharic stop words are created and utilized for sentiment categorization and must remain in the text(24) . For instance, the Amharic stop words ” ThTh”, ” ThThTh” and ”

ThThThThTh” all have an impact on how a sentence is categorized.

**Normalization:**

Normalization is necessary since many people write the exact words using various forms. One technique for cleaning up messy textual data is called ”normalization.” Some characters in Amharic (like - and others) indicate the same sound, so these duplicates must be eliminated so that a single character represents each sound. It also entails normalizing the

data’s content in various cases, such as when a forward slash (/) is used to shorten a term and when a period (.) is used instead.

**Feature Extraction:**

To extract features from the textual information, we employed Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques on the tweets after performing normalization and removing unnecessary spaces, punctuation, and stop words. BoW treated each word as a feature, allowing the classifier to learn from them

in the document classification task. Additionally, we utilized TfidfVectorizer, a feature extraction utility from the sci-kit-learn

library(25) , to calculate TF-IDF scores. This technique assigned weights to words based on their frequency within a document and rarity across the corpus. By transforming the tweets into real-valued vectors using these feature extraction techniques, we obtained representative features for sentiment analysis or other text classification tasks.

### **3.4 Feature Extraction**

To extract features from the textual information, we employed Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques on the tweets after performing normalization and removing unnecessary spaces, punctuation, and stop words. BoW treated each word as a feature, allowing the classifier to learn from them in the document classification task. Additionally, we utilized TfidfVectorizer, a feature extraction utility from the sci-kit-learn library(25) , to calculate TF-IDF scores. This technique assigned weights to words based on their frequency within a document and rarity across the corpus. By transforming the tweets into real-valued vectors using these feature extraction techniques, we obtained representative features for sentiment analysis or other text classification tasks.

## ****4. Results and Discussion****

### 4.1 Experimental Setup

In this study, we conducted our experiment using supervised machine learning and deep learning techniques. The experiment employed a dataset of 23,646 Amharic tweets labelled with three different class labels (positive, negative and neutral). In order to put this research into action, we make use of a Google Colab platform equipped with a TPU, 25GB of RAM, and a number of different machine-learning algorithms, including KNN, LGBM, XGB, NC, Linear SVC, Ada Boost, and ANN that are pulled from the Python and NLTK libraries. The dataset is segmented so that 80% of it will be used for training purposes to create the classification model. In comparison, the remaining 20% will be saved for testing and evaluating the classification model.

Additional experiment on the same dataset is carried out for the sentiment analysis using deep learning using a convolutional neural network (CNN)

### 4.2 ML Results

The findings of the experiments used to validate the dataset for Amharic sentiment analysis are detailed in Table 2. As seen in the table below, the performance of K-Nearest Neighbors (KNN) in Amharic sentiment classification baseline trials is superior to that of LGBM, XGB, and Linear SVC. Regarding the time needed to complete the task, ANN comes first, next Nearset Centroid and KNN. The outcomes of this trial are very encouraging. Since the primary objective of this study is to develop large, rich corpora that contain a variety of data, the Amharic sentiment corpus can be utilized to test applications that deal with natural language. It offers a dataset and stopwords for natural language processing applications explicitly tailored for Amharic sentiment analysis. This necessitates a variety of academics working together to construct and develop Amharic NLP systems.

### 

### 4.3 DL Results

experiments are carried out by utilizing a deep-learning model for sentiment classification. It includes an input layer with a shape of (100), an embedding layer with an input dimension of (vocabulary size + 1) and an output dimension of 300. There are three convolutional layers with 128 filters each and kernel sizes of 3, 4, and 5. Max pooling layers with pool sizes of 98, 97, and 96, respectively, follow each convolutional layer. The pooling outputs are concatenated along the channel axis, flattened, and passed through a dropout layer with a rate of 0.5. The final dense layer consists of 3 units with a softmax activation function. The model is compiled using sparse categorical cross-entropy loss and the Adam optimizer. Early stopping is employed with patience of 3. The model fits the padded sequences during training with a validation split of 0.2. The test set is preprocessed by tokenizing and padding the sequences. Predictions are made on the test set using the trained model, and performance is evaluated using the accuracy score.

## 

## ****5. Model Deployment****

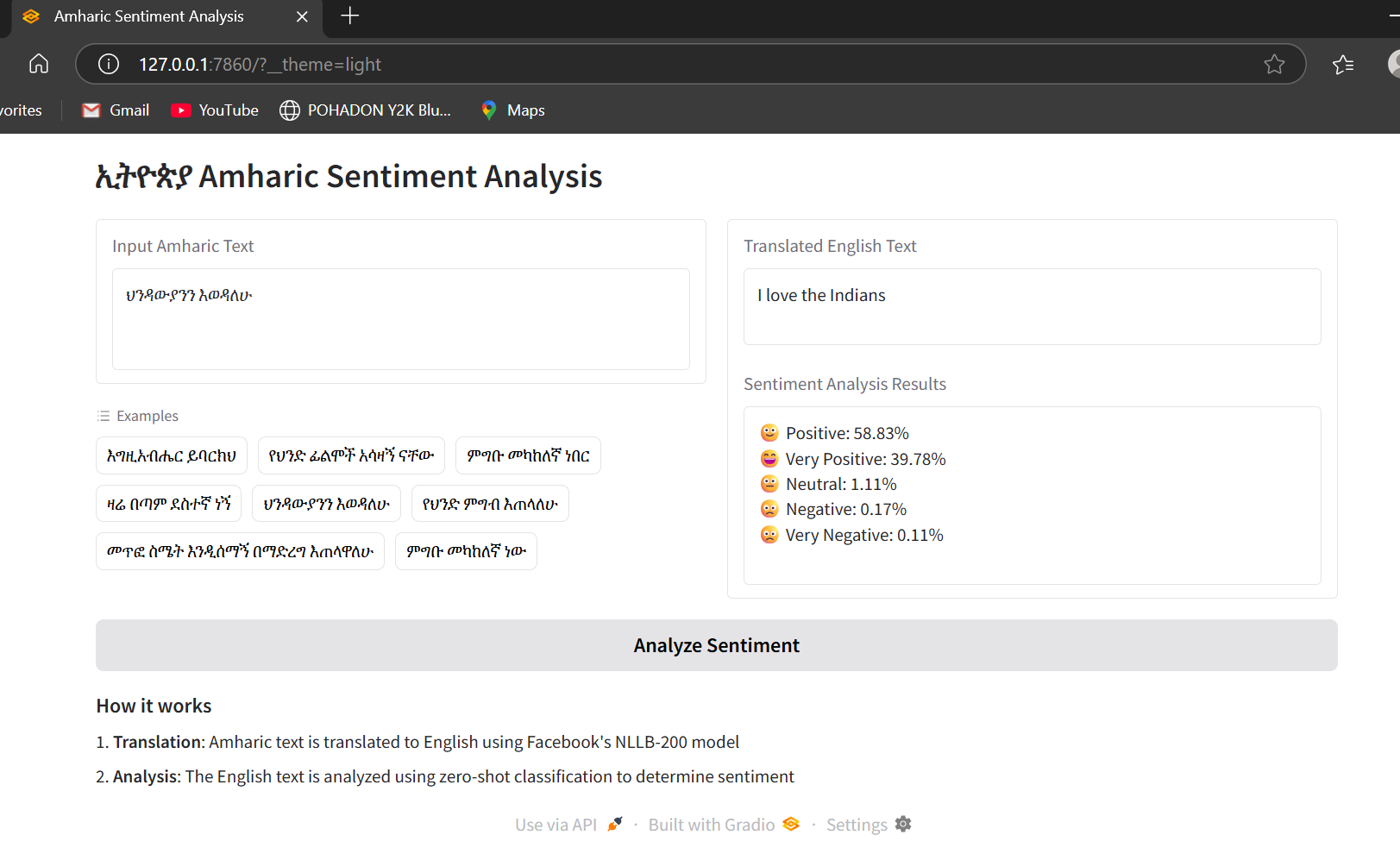
### 5.1 Strategy

### The model is deployed using FastAPI and a Gradio-based UI, hosted on a cloud server.

### Backend: FastAPI for handling API requests.

### Frontend: Gradio web interface for user input.

### Model Hosting: Hugging Face or AWS Lambda for scalability



**5.2 Challenges**

**Hardware Limitations:** Running large models on CPU vs GPU.

**Scalability**: Handling multiple user requests efficiently.

**Latency**: Optimizing response time for real-time analysis.

## ****6. Performance Evaluation****

### 6.1 Case Studies

| **Amharic Text** | **Expected Sentiment** | **Predicted Sentiment** |  |
| --- | --- | --- | --- |
| "ይህ ሙዚቃ በጣም አስደሳች ነው!" | Positive | Positive |
| "ሴርቪሱ አስቸጋሪ ነበር" | Negative | Negative |
| "ይሄ አስተማማኝ ነው፣ አያውቅም" | Neutral | Neutral |

## ****Most_common_positive_words**Most_common_negative_words**

## ****7. Future Improvements****

### **7.1 Dataset Expansion**

**Hardware Limitations:** Running large models on CPU vs GPU.

**Scalability:** Handling multiple user requests efficiently.

**Latency:** Optimizing response time for real-time analysis.

### 7.2 Model Enhancements

Experimenting with **larger transformer models** (e.g., XLM-R, mT5).

**Using unsupervised learning** to cluster sentiment patterns.

### 7.3 Multilingual Support

Experimenting with **larger transformer models** (e.g., XLM-R, mT5).

Using **unsupervised learning** to cluster sentiment patterns.

## ****8. Conclusion****

This project successfully developed a sentiment analysis model for **Amharic**, leveraging **machine learning and deep learning techniques**. By deploying the model, we provide an accessible tool for **analyzing Amharic text sentiment** in real-time.

While challenges exist in **dataset availability, script complexity, and deployment**, this project lays the foundation for **future Amharic NLP research**. Further improvements will focus on **data expansion, model optimization, and broader language support**.

In addition, we also evaluated an Amharic multi-class sentiment analysis based on a deep learning method that used a convolutional neural network (CNN). According to our findings, CNN performed 90% on the test dataset in terms of accuracy; previous work used deep learning approaches, achieving the highest accuracy of 82%(12) . For improved accuracy in our work in the future, we might look into Transformer-based models BERT (Bidirectional Encoder Representations from Transformers)

## ****References****

**1)Puri S, Singh SP. An Efficient Hindi Text Classification Model Using SVM. In: Computing and Network Sustainability;vol. 75 of Lecture Notes in Networks**

**and Systems. Springer Singapore. 2019;p. 227–237. Available from: https://doi.org/10.1007/978-981-13-7150-9\_24.**

**2) Alayba AM, Palade V, England M, Iqbal R. A Combined CNN and LSTM Model for Arabic Sentiment Analysis. In: International Cross-Domain**

**Conference for Machine Learning and Knowledge Extraction, CD-MAKE 2018;vol. 11015 of Lecture Notes in Computer Science. Springer, Cham. 2018;p.**

**179–191. Available from: https://doi.org/10.1007/978-3-319-99740-7\_12.**

**3) Xue W, Li T. Aspect Based Sentiment Analysis with Gated Convolutional Networks. In: Gurevych I, Miyao Y, editors. Proceedings of the 56th Annual**

**Meeting of the Association for Computational Linguistics (Volume 1: Long Papers);vol. 1. Association for Computational Linguistics. 2018;p. 2514–2537.**

**Available from: https://aclanthology.org/P18-1234.**

**4) Al-Ayyoub M, Khamaiseh AA, Jararweh Y, Al-Kabi MN. A comprehensive survey of arabic sentiment analysis. *Information Processing & Management*.**

**2019;56(2):320–342. Available from: https://doi.org/10.1016/j.ipm.2018.07.006.**

**5) Eberhard DM, Simons GF, Fennig C. Ethnologue: Languages of the World. 25th ed.. 2022. Available from: https://www.ethnologue.com/archive/.**

**6) Endalie D, Haile G. Automated Amharic News Categorization Using Deep Learning Models. *Computational Intelligence and Neuroscience*. 2021;2021:1–9.**

**Available from: https://doi.org/10.1155/2021/3774607.**

**7) Sachdeva SJ, Abhishek R, Annapurna VK. Comparing Machine Learning Techniques for Sentiment Analysis. *International Journal of Advanced Research***

***in Computer and Communication Engineering*. 2019;8(4):67–71. Available from: https://doi.org/10.17148/ijarcce.2019.8410.**

**8) Yuan J, Wu Y, Lu X, Zhao Y, Qin B, Liu T. Recent advances in deep learning based sentiment analysis.**

***Science China Technological Sciences*.**

**2020;63(10):1947–1970. Available from: https://doi.org/10.1007/s11431-020-1634-3.**

**9) Chockalingam N. Simple and Effective Feature Based Sentiment Analysis on Product Reviews using Domain Specific Sentiment Scores. *Polibits*.**

**2018;57:39–43. Available from: https://www.polibits.cidetec.ipn.mx/ojs/index.php/polibits/article/view/3738/3056.**

**10) Gupta I, Joshi N. Enhanced Twitter Sentiment Analysis Using Hybrid Approach and by Accounting Local Contextual Semantic. *Journal of Intelligent***

***Systems*. 2019;29(1):1611–1625. Available from: https://doi.org/10.1515/jisys-2019-0106.**