HiLCoE
School of Computer Science & Technology

Tittle:

An Amharic News Text classification

Group member

1. **Tilahun Tadios**

NLP _CS724

Date September 21, 2023

Submitted to Dr. Wondosen Muluta

Table of Contents

1.0 Introduction

A news article discusses recent news of either general interest (i.e., daily newspapers) or on a specific topic (i.e. political or trade news magazines, club newsletters, or technology news websites). A news article can include accounts of eyewitnesses to the happening event. We must have seen the news divided into categories when we go to a news website. Some of the popular categories that you'll see on almost any news website are tech, entertainment, sports, etc. If you want to know how to classify news categories using machine learning, this article is belonging to the readers. Every news website classifies the news article before publishing it so that every time visitors visit their website can easily click on the type of news that interests them. For example, I like to read the latest technology updates, so every time I visit a news website, I click on the technology section. But you may or may not like to read about technology, you may be interested in politics, business, entertainment, or maybe sports. Currently, the news articles are classified by hand by the content managers of news websites. But to save time, they can also implement a machine learning model on their websites that read the news headline or the content of the news and classifies the category of the news. In this project will implement NLP machine learning approaches for classifying Amharic text news.

Amharic is the second most spoken Semitic language. It is the official working language of 100 million people that reside in the Federal Democratic Republic of Ethiopia. The language uses its unique alphabet called Fi-del. Amharic alphabet consists of punctuation and numbers in addition to its 231 primary letters [1].

Amharic is considered as a low resource language [1]This is not due to the lack of raw data, rather it is due to the scarcity of labeled data. Most of the time re- searchers prepare data for their use but fail to make the dataset available Text classification or text categorization is a task of assigning a sentence, paragraphs or documents into one of n classes we have on our dataset. This task is one of the core NLP tasks that needs manually annotated data as an input[6]Tasks like Sentiment analysis, News categorization, Topic Analysis and more are promi- nent application of classification task [2]we usually use languages like English for NLP tasks especially in academia for education and we don’t study the effect of different algorithms in languages which have different structure than English. This does not consider characteristics of low resource languages while developing new algorithms. In NLP, text classification is one of the primary problems we try to solve and its uses in language analyses are indisputable. The lack of labeled training data made it harder to do these tasks in low resource lan-guages like Amharic. The task of collect- ing, labeling, annotating, and making valuable this kind of data will encourage juniorresearchers, schools, and machine learning practitioners to implement existing classifi- cation models in their language. In this short paper, we aim to introduce the Amharic text classification dataset that consists of more than 50k news articles that were categorized into 6 classes. This dataset is made available with easy base-line performances to encourage studies and better performance experiments.

2.0 Problem

Text classification datasets are used to categorize natural language texts according to content. For example, think classifying news articles by topic, or classifying book reviews based on a positive or negative response. Text classification is also helpful for language detection, organizing customer feedback, and fraud detection. while this process is time-consuming when done manually, it can be automated with machine learning models. Category classification, for news, is a multi-label text classification problem.

3.0 Objective

Text classification is a common NLP task used to solve business problems in various fields. The goal of text classification is to categorize or predict a class of unseen text documents, often with the help of supervised machine learning . automated classification of texts has been flourishing in the last decade or so due to incredible increase in electronic documents on the Internet; this renewed the need for automated text classification. When Amharic is considered, electronic documents are increasing that needs automatic classification. This project describes how to organize massively available Amharic news items into meaningful way by undergoing automatic classification. Therefore, the main objective of this project is to classify Amharic document in predefined class based on bag of words.

4.0 Previous works

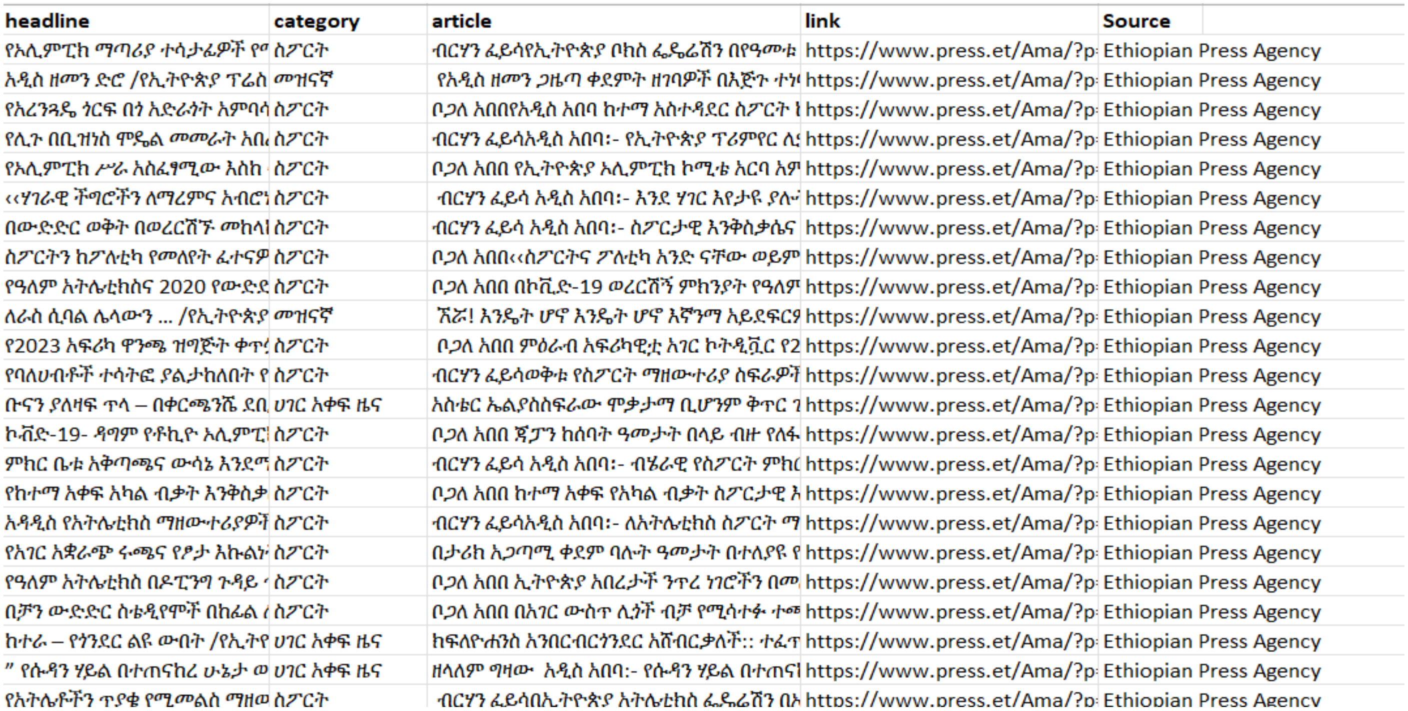
Text classification task is one of the core NLPtasks that needs manually annotated data as an input [2]There are some works done by [3], [4]), [5], [6] and others. We have found that all of them have used a very small dataset which ranges from 200 - 15,000 articles from a single data source. Some researches also talk about lack of standard Amharic text classifi- cation corpus [4]

5.0 Methodology.

For doing this project we use anaconda, Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment.

5.1 Data Cleaning and Data Preprocessing

Data preprocessing is the process of transforming raw data into an understandable format. Lowercasing the data, Removing Punctuations, Removing Numbers, Removing extra space, Replacing the repetitions of punctations, Removing Emojis, Removing emoticons and Removing Contractions. The quality of the data should be checked before applying machine learning or data mining algorithms.



headline	category	article	link	Source
የኦሊምፒክ ማጣሪያ ተሳታፊዎች የ	ስፖርት	ብርሃን ፈይሳየኢትዮጵያ ቦክስ ፌዴሬሽን በየዓመቱ	https://www.press.et/Ama/?p	Ethiopian Press Agency
አዲስ ዘመን ድሮ /የኢትዮጵያ ፕሬስ	መዝናኛ	የአዲስ ዘመን ጋዜጣ ቀደምት ዘገባዎች በእጅጉ ተነ	https://www.press.et/Ama/?p	Ethiopian Press Agency
የአረንጓዴ ጎርፍ በጎ አድራጎት አምባሳ	ስፖርት	ቦጋለ አበበየአዲስ አበባ ከተማ አስተዳደር ስፖርት	https://www.press.et/Ama/?p	Ethiopian Press Agency
የሊጉ በቢዝነስ ሞዴል መመራት አበ	ስፖርት	ብርሃን ፈይሳአዲስ አበባ:- የኢትዮጵያ ፕሪምየር ሊ	https://www.press.et/Ama/?p	Ethiopian Press Agency
የኦሊምፒክ ሥራ አስፈፃሚው እስከ	ስፖርት	ቦጋለ አበበ የኢትዮጵያ ኦሊምፒክ ኮሚቴ አርባ አም	https://www.press.et/Ama/?p	Ethiopian Press Agency
‹‹ሃገራዊ ችግሮችን ለማረምና አብሮ	ስፖርት	ብርሃን ፈይሳ አዲስ አበባ:- እንደ ሃገር እየታዩ ያሉ	https://www.press.et/Ama/?p	Ethiopian Press Agency
በውድድር ወቅት በወረርሽኙ መከላ	ስፖርት	ብርሃን ፈይሳ አዲስ አበባ:- ስፖርታዊ እንቅስቃሴና	https://www.press.et/Ama/?p	Ethiopian Press Agency
ስፖርትን ከፖለቲካ የመለየት ፈተናዎ	ስፖርት	ቦጋለ አበበ‹‹ስፖርትና ፖለቲካ አንድ ናቸው ወይም	https://www.press.et/Ama/?p	Ethiopian Press Agency
የዓለም አትሌቲክስና 2020 የውድድ	ስፖርት	ቦጋለ አበበ በኮቪድ-19 ወረርሽኝ ምክንያት የዓለም	https://www.press.et/Ama/?p	Ethiopian Press Agency
ለራስ ሲባል ሌላውን … /የኢትዮጵያ	መዝናኛ	ኸረ! እንዴት ሆኖ እንዴት ሆኖ እኛንማ አይደፍርን	https://www.press.et/Ama/?p	Ethiopian Press Agency
የ2023 አፍሪካ ዋንጫ ዝግጅት ቀጥ	ስፖርት	ቦጋለ አበበ ምዕራብ አፍሪካዊቷ አገር ኮትዲቯር የ2	https://www.press.et/Ama/?p	Ethiopian Press Agency
የባለሀብቶች ተሳትፎ ያልታከለበት የ	ስፖርት	ብርሃን ፈይሳወቅቱ የስፖርት ማዘውተሪያ ስፍራዎች	https://www.press.et/Ama/?p	Ethiopian Press Agency
ቡናን ያለዛፍ ጥላ – በቀርጫንሼ ደበ	ሀገር አቀፍ ዜና	አስቴር ኤልያስስፍራው ሞቃታማ ቢሆንም ቅጥር ገ	https://www.press.et/Ama/?p	Ethiopian Press Agency
ኮቪድ-19- ዳግም የቶኪዮ ኦሊምፒ	ስፖርት	ቦጋለ አበበ ጃፓን ከሰባት ዓመታት በላይ ብዙ የለፋ	https://www.press.et/Ama/?p	Ethiopian Press Agency
ምክር ቤቱ አቅጣጫና ውሳኔ እንደሚ	ስፖርት	ብርሃን ፈይሳ አዲስ አበባ:- ብሄራዊ የስፖርት ምክ	https://www.press.et/Ama/?p	Ethiopian Press Agency
የከተማ አቀፍ አካል ብቃት እንቅስቃ	ስፖርት	ቦጋለ አበበ ከተማ አቀፍ የአካል ብቃት ስፖርታዊ እ	https://www.press.et/Ama/?p	Ethiopian Press Agency
አዳዲስ የአትሌቲክስ ማዘውተሪያዎች	ስፖርት	ብርሃን ፈይሳአዲስ አበባ:- ለአትሌቲክስ ስፖርት ማ	https://www.press.et/Ama/?p	Ethiopian Press Agency
የአገር አቋራጭ ሩጫና የዖታ እኩልነ	ስፖርት	በታሪክ አጋጣሚ ቀደም ባሉት ዓመታት በተለያዩ የ	https://www.press.et/Ama/?p	Ethiopian Press Agency
የዓለም አትሌቲክስ በዶፒንግ ጉዳይ	ስፖርት	ቦጋለ አበበ ኢትዮጵያ አበረታች ንጥረ ነገሮችን በመ	https://www.press.et/Ama/?p	Ethiopian Press Agency
በቻን ውድድር ስቴዲየሞች በከፊል	ስፖርት	ቦጋለ አበበ በአገር ውስጥ ሊጎች ብቻ የሚሳተፉ ተጫ	https://www.press.et/Ama/?p	Ethiopian Press Agency
ከተራ – የጎንደር ልዩ ውበት /የኢትዮ	ሀገር አቀፍ ዜና	ክፍለዮሐንስ አንበርብርጎንደር አሸብርቃለች:: ተፈጥ	https://www.press.et/Ama/?p	Ethiopian Press Agency
” የሱዳን ሃይል በተጠናከረ ሁኔታ ወ	ሀገር አቀፍ ዜና	ዘላለም ግዛው አዲስ አበባ:- የሱዳን ሃይል በተጠና	https://www.press.et/Ama/?p	Ethiopian Press Agency
የአትሌቶችን ጥያቄ የሚመልስ ማዘወ	ስፖርት	ብርሃን ፈይሳበኢትዮጵያ አትሌቲክስ ፌዴሬሽን በኦ	https://www.press.et/Ama/?p	Ethiopian Press Agency

Table 1: sample collected for preprocessed CSV data

5.2 Analysis Text

Text classification is one of the important tasks in supervised machine learning (ML). It is a process of assigning tags/categories to documents helping us to automatically & quickly structure and analyze text in a cost-effective manner. It is one of the fundamental tasks in Natural Language Processing with broad applications such as sentiment-analysis, spam-detection, topic-labeling, intent-detection etc. However, it not easy to do unless the texts are well structured and standard forms some common problem in Amharic text needs to be standard are :

- ✓ Problem of Amharic Writing System: there are a number of problems associated with Amharic writing system which are challenging natural language processing of Amharic documents. Redundancy of some characters sometimes more than one character is used for similar sound in Amharic For example, the word **‘ጸሀይ’ (‘sun’)** can be represented in Amharic as **ጸሀይ, ጸሐይ, ጸኀይ, ፀሀይ, ፀሐይ, ፀኀይ, sport(ስፖርት) ሥፖርት .** etc
- ✓ Amharic Punctuation Marks: like ‘Hulet Neteb’ (‘:’)-word separator and ‘Arat Neteb’ (‘::’)-sentence separator are the major punctuation marks. But space is mostly used instead of Hulet Neteb (‘:’) specially in computer writing system.

- ✓ Amharic Number System: some time the news may be written with Geez numbering. Therefore, before doing classification, it is mandatory to normalize such type of nonstandard writings.

5.3 Classification Dataset

Our dataset consists of 6 classes. This class information was found by the tag we get from websites and we manually verify the case and removed noises in the process. **Table 3** shows the detailed description of the web pages we collected our dataset from. The web pages are local and international news sites. We have collected the datasets from different sources to increase the variety of the text. As far as we know, this is the first work with 1) data collected from different sources,2) data size is at least 5 times greater than the existing benchmark datasets. We included several details that might be useful for different purposes. It includes information like the web page the article is found from, Views it had, the title of the article, and the date it was posted at. However, the ’category’ and ’article’ metadata are very important while the other metadata might be still useful for different use-cases.In this data local news and international news refers to topics that are not included in the rest and are categorized as international or regional issue.

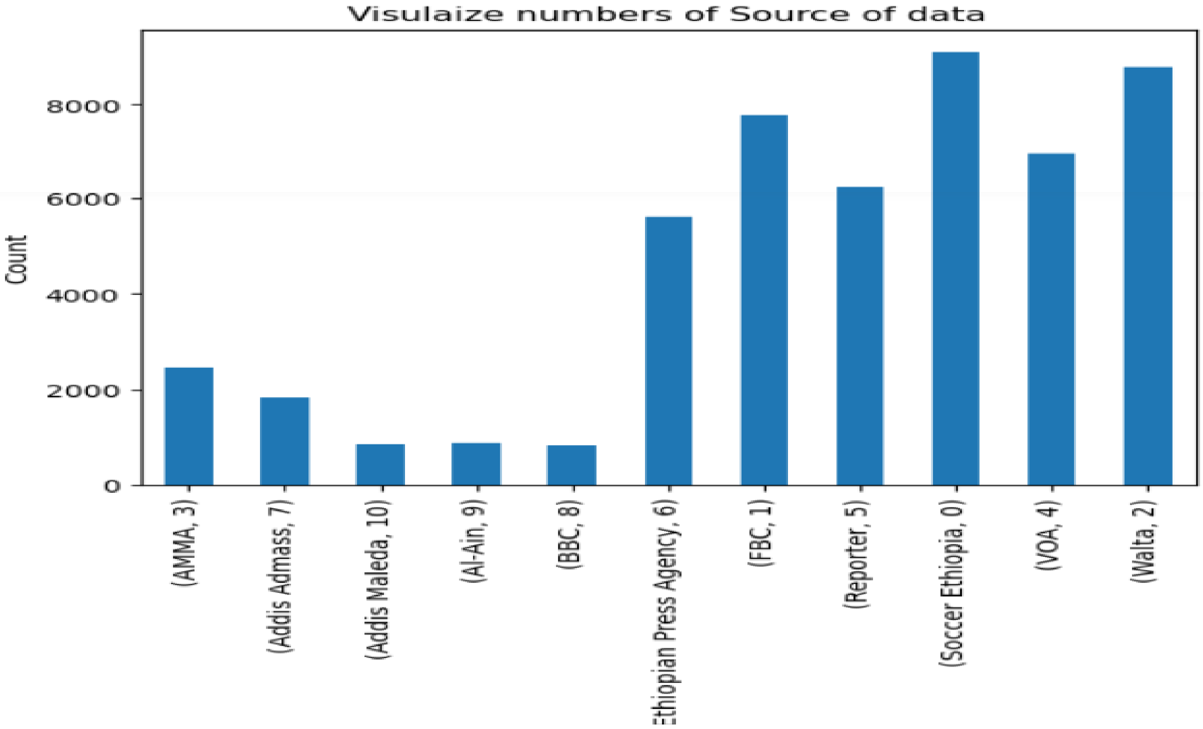
Class Name	**items**
ሀገር አቀፍ ዜና / local news	20564
ስፖርት / sport	9812
ፖለቲካ / politics	9307
ዓለም አቀፍ ዜና / international news	6515
ቢዝነስ / business	3873
መዝናኛ / entertainment	635
Total	50706

Table 2: Different class distrbution in the dataset

The source of corpus used in our project is from various Amharic news which was published in various time. present work aimed at compiling an Amharic corpus from the Web and automatically categorizing the texts. Amharic is the second most spoken Semitic language in the World (after Arabic) and used for countrywide communication in Ethiopia. It is highly inflectional and quite dialectally diversified. We discuss the issues of compiling and annotating a corpus of Amharic news articles from the Web. This corpus was then used in three sets of text classification experimen

News site	Item	url
Soccer Ethiopia	9091	www.soccerethiopia.net
Walta	8785	www.waltainfo.com/am/
FBC	7784	www.fanabc.com
VOA	6981	www.amharic.voanews.com
Reporter	6280	www.ethiopianreporter.com
Ethiopian Press Agency	5598	www.press.et/Ama/
AMMA	2442	www.amharaweb.com/
Addis Admass	1847	www.addisadmassnews.com
Al-Ain	887	https://am.al-ain.com
Addis Maleda	861	www.addisma eda.com
BBC	816	www.bbc.com/amharic/
Total	**51372**	

Table 3:Dataset Source distribution



6.Result evaluation

For implementing this model, we use 51343 data records from those Trained data (41074, 1000) Test data (10269, 1000) so different alternative algorithm are conducted to select best model .Results are obtained from Gaussian Naive Bayes, Random Forest, Logistic Regression, K-nearestneighbors, Decision Tree, and Support Vector Machine classification algorithms which gave an accuracy of as stated in table 4 below using both count vectorizer and Term Frequency-Inverse Document Frequency(TF-IDF) feature extraction techniques, and show that Logistic Regression was able to provide higher accuracy in comparison to the other five methods using count vectorizerand in case of TF-IDF, Support Vector Machine (SVM) is the higher accuracy. This comparison is shown in Table4 .

Algorithm	**Accuracy**	
	CountVectorizer	TF-IDF
Gaussian Naive Bayes	69.66%	71.21%
Random Forest	84.18%	86.2%

Logistic Regression	84.77%	85.92%
K-nearest neighbors	45.08%	36.66%
Decision Tree	70.12%	72.68%
Support Vector Machine	82.86%	86.55%

\Table 4:Baseline classification accuracy performance for each algorithm.

Future works in this dataset include trying to improve the performance of the models using advanced word embedding and trans- former models. From the table 4, you can see our dataset is imbalanced so we are thinking of using text augmentation techniques to increase dataset size. Different approaches to imbalance class problems should be explored on this dataset too. There are a lot of approaches that we know and use for data imbalance and exploring those approaches on this dataset can show us the effectiveness of the methods in Amharic language data.

7. Conclusion

In this work, we release the Amharic text classification dataset. This work contributes to the low amount of Amharic text classification dataset and aims to be a good starting point for future Amharic text classification works. The classification algorithm indicted that Naive Bayes using count vectorizer features, Naive Bayes using Tf-idf features 62.2% and 62.3% accurace.

8. References

[1] Andargachew Mekonnen Gezmu, B. Seyoum, M. Gasser, and A. Nürnberger.;, "Contemporary amharic corpus: Automatically morphosyntactically tagged amharic corpus," 2018.

[2] 2014., Dr.Vuda Sreenivasa Rao2 Seffi Gebeyehu., two step data mining approach for amharic text classfication American Journal of Engineering, 2014.

[3] Kelemework, Worku, "Automatic amharic text news classification: Aneural networks approach. Ethiop. J. Sci. & Technol. 6(2)," pp. 127-137, 2013.

[4] Mulugeta Sahlemariam, Meron; Daniel., Libsie and Yacob, "Concept-based automatic amharic document categorization," p. 116, 2009.

[5] TEGEGNIE, ALEMU KUMILACHEW, "Hierarchical amharic news text classification hierarchical amharic news text classification," 2010.

[6] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura E. Barnes, and Donald E. Brown., "Text classification algorithms: A survey. CoRR preprint arXiv:1904.08067.," 2019.