# Document Classification for the Under-resourced Amharic Language

Anonymous ACL submission

## Abstract

Natural language processing (NLP) is severely hampered by a scarcity of digital resources. This is especially true for Amharic, a language with few resources but a rich morphology. In response, a total of 67,739 Amharic news documents from 8 different categories are gathered from web sources. A baseline document categorization experiment is carried out to validate the usability of the obtained corpora from various domains. In the lack of linguistic information, the experimental results reveal that deep learning achieves 84.53% accuracy.

## 1 Introduction

People exchange information through oral, written, and visual communication. When compared to other forms of web communication, text-based information and knowledge sharing is the most widely used (Park et al., 2012). Amharic is mostly spoken in Ethiopia and is utilized as a form of communication in a variety of sectors, including the legal system, trade, communications, the military, and religion among the 7,164 registered languages Eberhard and Fennig (2020). Amharic is the world's second most spoken Semitic language, after Arabic. Amharic is the primary working language of Ethiopia's Federal Government and regional languages such as Amhara, Diredawa, and the Southern Nations and Nationalities People Regions (SNNPR). Unlike other Ethiopian languages, Amharic is also utilized for inter-regional communication.

The language employs characters derived from Geez, with some additional characters added to fill in the gaps. Amharic distinguishes itself from other resourced languages such as English, European, and Asian. These characteristics include the alphabet, numbering system, gender sensitivity, as well as the complicated morphology (Samplius, 2020). Due to these characteristics, it is known to be morphologically rich and complex language. Despite the complexity of the language, the expansion of the internet as a communication medium has resulted in an ever-increasing need for NLP, which contributes to the provision of information through a variety of applications. These applications include machine translation, speech recognition, topic modeling, sentiment analysis and text classification among others (Jurafsky and Martin, 2008). Text classification is a way of automatically classifying text into predefined categories using NLP based on its content (Sammut and Webb, 2011). The classification might be at the level of document, paragraph, sentence or sub-sentence level depending on the need.

The challenges involved in developing NLP applications in Ethiopia are subject to technological and linguistic factors such as character variations that generate same meaning from different orthographic representation. This may contribute to the complexity of NLP applications. Furthermore, the Amharic language's character shifts generate a high-dimensional feature space for machine learning, which not only results in a high level of computational complexity but is also prone to overfitting. In addition, Amharic language does not have sufficient and structured language corpora for the development of the natural language applications. To address this, we have attempted to collect and prepare text Amharic text corpora which can then be used for the development of different NLP applications. To check the usability of the developed text corpora, a document classification experiment is conducted using classical, ensemble machine learning as well as deep learning technique.

### 1.1 Data Collection and Preparation

The researcher used various ways to obtain a general purpose text corpus for natural language processing from more than 25 registered news and religious domains of various sources that give publicly accessible news in various languages and cat-

egories. The sources are selected based on the language of the news, ability to provide the news in electronic format and a minimum of three years in the area. From these news sources a total of 67,739 documents were collected to the end of October, 2020. To extract a web news item from the different websites, a web crawler is used for each article after identifying the structure of web documents (html) including the page navigation.

As part of the preprocessing, we performed Amharic characters normalization, sentence segmentation and data cleaning for the removal of unnecessary links, emoji, symbols and foreign words as well as the removal of the extra white space incorporated in the web documents. To perform the pre-processing tasks, Python scripts were used including NLTK and Regular Expressions libraries.

## 2 Experimental Results and Discussion

The Amharic document classification conducted using a total of 67,739 documents from 8 different categories. To validate the usability of the prepared corpora, a document classification experiment are conducted by selecting six machine learning algorithms. Accordingly, Naive Bayes (NB) and Support Vector machine (SVM) from classical machine learning (Zheng, 2019), and Gradient Boosting (GB) and Random Forest (RF) from ensemble learning are used (Onan et al., 2016), while Deep Neural Network (DNN) and Convolutional Neural Network (CNN) from deep-learning technique are employed (Kim, 2014; Conneau et al., 2016).

The performance registered for deep learning shows higher than any other machine learning techniques. Among the deep learning, DNN which is special type of Recurrent Neural Network has presented the highest performance than that of CNN. DNN has improved the accuracy of document classification by 1.8% than that of CNN with a relative error reduction of 10.42% (82.73% to 84.53%). One of the reason for the high performance by DNN over CNN is that, the convolutions and pooling operations works by selecting the best performers which makes the CNN to ignore the local ordering which is not in DNN (Kim, 2014).

Compared to classical machine learning, a GB classifier using ensemble techniques shows the lowest performance than any other experiment conducted in ensemble and deep learning. GB improved by relative error rate of 4.96% (74.78% to 76.03%) in using Random Forest and at most 38.66% (74.78% to 84.53%) while using DNN. Similarly, the experimental result of NB registered a lower performance from the classical than the deep learning techniques. The amount of relative error reduced by at least 3.50% (77.46% to 78.25%) and at most 31.36% (77.46% to 84.53%) by using SVM and DNN respectively. The reason why a better performance was registered in SVM than NB is that, NB treat the text data as independent features while SVM attempt to look at the interaction to certain extent. In addition to this, NB performs better in snippets than that of the full-length document (Wang and Manning, 2012).

The baseline experiment of the Amharic document classification shows that the corpus collected from news and religious source can be used with at least 74% accuracy for the document classification. Compared to the experiment conducted in Eyassu and Gambäck (2005); Asker et al. (2007); Kelemework (2013); Tegegnie et al. (2017), the result registered in this experiment is promising given the data size, data variety, data complexity, concept usage variety and content semantics of the corpus organized for different NLP applications. In general, the results obtained in this experiment are promising. As the main aim of this paper is to produce large size corpora, rich in content with a variety of data, the result shows the usability of the Amharic document corpus which can serve as a test bed for designing natural language applications.

## 3 Concluding Remarks

This paper presents an attempt to collect and prepare usable Amharic text corpora for one of the Ethiopian languages, Amharic. Corpora have been collected from eight different sources including the religious website using the SOTA Python libraries. The collected corpora are further processed for performing normalization, punctuation correction and data cleaning in the course of preparing the corpus for different NLP tasks. To check the usability of the collected, preprocessed and normalized data, the Amharic text document classification experiments have been conducted using machine learning techniques with an accuracy of 84.53%.

Beside these, the collected corpus can be used for designing and developing a number of NLP applications and resources. In general, the paper has demonstrated an approach to the development of a corpus for a resource-deficient language, Amharic, for use in NLP applications.

## References


Lars Asker, Atelach Alemu Argaw, Björn Gambäck, and Magnus Sahlgren. 2007. Applying machine learning to Amharic text classification. In *Proceedings of the 5th World Congress of African Linguistics*.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2016. Very deep convolutional networks for text classification. *arXiv preprint arXiv:1606.01781*.

Gary F. Simons Eberhard, David M. and Charles D. Fennig. 2020. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas.

Samuel Eyassu and Björn Gambäck. 2005. Classifying Amharic news text using self-organizing maps. *43rd Annual Meeting of the Association for Computational Linguistics; Workshop on Computational Approaches to Semitic Languages*.

Daniel Jurafsky and James H Martin. 2008. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*.

Worku Kelemework. 2013. Automatic Amharic text news classification: A neural networks approach. *Ethiopian Journal of Science and Technology*, 6(2):127–137.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57:232–247.

Namkee Park, Jae Eun Chung, and Seungyoon Lee. 2012. Explaining the use of text-based communication media: An examination of three theories of media use. *Cyberpsychology, Behavior, and Social Networking*, 15(7):357–363.

Claude Sammut and Geoffrey I Webb. 2011. *Encyclopedia of machine learning*. Springer Science & Business Media.

Samplius. 2020. The main characteristics of amharic language. [Online; accessed 15-April-2022].

Alemu Kumilachew Tegegnie, Adane Nega Tarekegn, and Tamir Anteneh Alemu. 2017. A comparative study of flat and hierarchical classification for Amharic news text using svm. *International Journal of Information Engineering & Electronic Business*, 9(3).

Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94.

Yuhan Zheng. 2019. An exploration on text classification with classical machine learning algorithm. In *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pages 81–85. IEEE.