International Journal of Engineering, Science and Information Technology
Volume 5, No. 4 (2025) pp. 191-197
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v5i4.987
Research Paper, Short Communication, Review, Technical Paper

Enhancing Multi-Label News Text Classification for an Understudied Language: A Comprehensive Study on CNN Performance and Pre-Trained Word Embeddings

Diriba Gichile Rundasa[1]*, Arulmurugan Ramu[2]

[1]Department of Computer Science, College of Engineering and Technology, Mattu University, Mattu, Ethiopia
[2]Department of Computer Sciences and Software Engineering, Heriot-Watt International Faculty, K. Zhubanov *University, Kazakhstan*

**Corresponding author Email: diriba.gichile@mau.edu.et*

The manuscript was received on 22 February 2025, revised on 15 May 2025, and accepted on 2 August 2025, date of publication 4 November 2025

Abstract

Today's news texts are classified using a multi-label system, which allows for the assignment of a potentially large number of labels to specific instances. The majority of earlier scholars have only looked into mutual exclusion at a single level. Nonetheless, the primary goal of this study was to categorise the news material using multiple labels. Many text documents are created these days from a variety of offline and internet sources. This generated news text is disordered state. As a result, timely access to the needed content from the sources is challenging. Compared with traditional text classification, multi-label classification is difficult and challenging because of its multi-dimensional labels. Convolutional neural networks are used in this study's tests on the problem domain for Afaan Oromo multi-label news text classification due to their ease of assimilation of pre-trained word embeddings. According to pre-trained word embedding with a train-test ratio of 10/90, the new proposed model has shown improved performance. The suggested CNN models might be helpful for labelling news articles in Afaan Oromo news text. The goal of many researchers working on Afaan Oromo classifier development is to use various learning algorithms to boost classification accuracy as the number of categories or labels increases. Using various approaches, they attempted to use basic machine learning methods to address the calculation time issue. Unfortunately, all low-resource language researchers focus on flat, hierarchical, and multi-class classification types, but we created a model for multi-label text classification and attempted to apply it using a deep learning algorithm. Over 5640 Afaan Oromo news dataset items are analysed experimentally over eight main news categories. Python served as our experimental platform for both text classification and word embedding. After the model is fully implemented, the best result of the precision, recall, F1 score and accuracy rate train test ratio of 10/90 for pertained word_ embedding is 89.7%, 88.6%, 93.3% and 96.5, respectively.

Keywords: *Text classification, Afaan Oromo, Convolutional Neural Network, Feature Extraction, Word2vec.*

1. Introduction

Text is one type of the vast amount of information that is available today in the real world. The amount of information produced both online and offline has grown exponentially every second due to the daily advancements in technology, specifically from the Internet news text rush to quickly increase news text. The disordered status of news texts increased, and several techniques of text classification have been proposed to overcome the problem. There are numerous text classifications like single-label, hierarchical and multi-label classification. For instance, one kind of classification when an input instance is allocated to only one class is called single-label classification. The process of automatically giving a single class to each input instance is known as single-label or class classification, or classical classification. In this approach, a classifier learns to correlate each unseen element with its most likely class or category. The categories or classes can be binary or multi-class. For example, binary classification has two classes, like positive or negative. News text classification uses training data to create models that assign classes to a text document, and the classes created are supervised techniques. This makes it easier for readers to swiftly access news texts [1]. In our study, we focus on multi-label classification. Multiple non-mutually exclusive labels are applied to instances from a collection of distinct classes in multi-label classification. Every instance is linked to a collection of relevant labels, with the remaining labels being meaningless [2]. Many researchers have explored Afaan Oromo news text classification using various machine learning algorithms, but as the number of classes increases, the accuracy decreases, and as the dataset size increases, the accuracy decreases. This problem can be solved with the aid of deep learning techniques. This study's objective is to use deep learning, which has numerous applications such as sentiment analysis, sentence classification [3] and document

classification to examine Afaan Oromo news text classification [4]. News text classification, which can be accomplished with machine learning and deep learning, is used to organise unstructured data. Because it can learn complicated feature representations, deep learning is the chosen method. Using this technique, news material from Afaan Oromo papers is categorised according to predetermined classifications [5].

Applications of multi-label learning are significant in a wide range of real-world issues, including bioinformatics, video annotation, scene classification, and text classification. Multi-label classification techniques are becoming more and more necessary in real-world applications these days, such as gene classification and article annotation. One common method is to reduce a multi-label problem to single-label tasks without considering the relationships between labels [6].

2. Literature Review

We reviewed the literature on classifying news texts using different machine learning techniques. Understanding the state-of-the-art in script identification was our goal. Although this topic has been studied by numerous researchers in various languages and geographical areas, we concentrated on closely related studies for our investigation. Several social media platforms contain raw data that is unlabelled and unstructured. Text classification research began in the 1950s. Then, in the early 1980s, knowledge engineering began classifying texts using rule-based systems. But, rule-based systems were time-consuming and had poor accuracy and coverage [7]. The techniques make it possible to quickly classify enormous amounts of text. Different researchers have conducted research on text classification, but most of them are focused on single-level, binary and multi-label or class-based classification. Some of them are described as follows:

A neural network model for the classification of Amharic news texts was developed by the first author. The study used LVQ to test a dataset of 1,583 items in nine categories. Using TF and TFIDF weighting schemes, the study also examined how classification accuracy was affected by adding more classes and news items. The study's average accuracy by TF and TFIDF was 75.5% and 71.6%, respectively [8]. The second author has developed a deep learning technique with CNN and word2vec algorithms for the classification of English text in online news and Twitter. According to the study, the accuracy of the CBOW model with CNN was greater (93.51%) than that of the skip-gram (91.61%). The second author has created a method for online news and Twitter English text classification using a deep learning approach with CNN and word2vec algorithms. The study found that using the CBOW model with CNN resulted in higher accuracy (93.51%) compared to skip-gram (91.61%). The study also found that because news material is more consistent than tweets, CNN was more appropriate for it [9]. CNN models are improved for the categorisation of English news texts using Word2vec and LDA. They obtained high F1 scores of 96.2%, 95.9%, and 96.4%, respectively, as well as great precision and recall [3].

To solve the problem of news text classification for Afaan Oromo, another experiment with the classification of news texts was carried out in the case of Afaan Oromo Radio Fana [10]. The study's researchers attempted to create processing tools for Afaan Oromo text classifications and assess the applicability of automatic text classifiers for Afaan Oromo news text classification tasks based on document content in order to address classification difficulties. Additionally, he made use of a corpus of 2,268 Radio Fana news texts. Ultimately, he evaluated the classifiers and found that Sequential Minimal Optimisation and Naïve Bayes performed better in terms of accuracy than other classifiers, at 95.82% and 96.58, respectively. An additional study was carried out using news-based Improved CNN models [3]. In order to create a real text feature representation, Word2vec is first combined with LDA. According to the results, the model in this work achieves a precision rate, a recall rate, and an F1 value of 96.4%, 95.9%, and 96.2%, respectively. The most current related study on news text classification was done utilising Python machine learning methods for multi-label Afaan Oromo news text categorisation. This study dealt with a number of multi-label text classification problems that were tested using several fronts, or data sets. With the knowledge we have gained, we will also be able to choose settings and methods that produce particular dataset properties. A detailed examination of the process of choosing various multilevel predictor parameters is given in these publications. While the KNN and Naïve Bayes algorithms had F1-Scores of 95% and 94% with a total precision of 94%, respectively, the greatest F1-Score was 93% with a total precision of 87%, suggesting that the latter is the best model for multilevel text categorisation [11].

3. Research Method

3.1. Data Collection

The first step and most important stage is gathering information from multiple sources to train the model in general and news text classification in particular. The corpus used to prepare the dataset for this study was gathered from the Oromia-Broadcasting Network (OBN) bureau and, Fana-broadcasting-Corporation(FBC). For this study, we collected 5640 records or instances that were annotated to different numbers of classes.

3.2. Dataset- preparation and Annotation

The second step is to proceed after data is collected; the dataset should be constructed by eliminating extraneous components from the raw data that was gathered. For example, eliminating the header section that provides information about each news text item or instances, such as the date, days and other news was produced and the news source that was transmitting, as well as the footer sections that typically include the name of the journalist. Experts then labelled the news data's remaining content with various predefined label groups. The Oromia-Broadcasting-Network (OBN) journalist in our instance labelled the data with various labels by going over the remaining news content and storing all of the sentences in a single Word document, separating or isolating the sentences with a (,). Ultimately, the labelled data is imported into Excel and saved as a CSV file. Data preparation, which includes Data preprocessing, involved removing and cleaning stop words, digits, punctuation, and symbols before the training and testing stages. All news texts are represented as such because of the multi-label dataset. Six categories or labels are (‘Education’, Technology’ ‘Economy’, ’Social’, ’Agriculture’, ’Politics’) where as eight categories or labels (Education’, ‘Technology‘ ‘Economy’,’ Social’,’ Agriculture’,’ Politics’,’ Business’,’ Sport’). When we assigned the instances into labels or classes, the dataset has a maximum of three multi-labels out of eight classes. One sentence could have one label, two labels, or three labels.

Table 1. The table shows the names of labels and the number of instances/records in each label individual

No	Labels (Class) name	Number of instances/records
1	Technology	750
2	Education	860
3	Economy	720
4	Social	680
5	Agriculture	638
6	Politics	612
7	Business	600
8	Sport	780
	Total Class: 8	5640 Instances

Table 2: The table displays the labels' names together with the number of typical instances for each label.

No	Labels (Class) name	Number of instances/records
1	Technology and Education	320
2	Social and Education	200
3	Social, Politics, and Sport	260
4	Politics and Social	220
5	Education, Technology and Economy	282
6	Economy and Business	380

3.3. Preprocessing

After the dataset is annotated we preparing the dataset for training and evaluation is known as preprocessing news. Language-dependent texts should be completed before implementing the automatic news text classification system. Since not all terms in the document are helpful for categorisation, there are numerous procedures involved in creating news text that is appropriate for the learning algorithms, including eliminating non-informative words or characters. To prepare news text for classification, there are stages involved in preprocessing it. One of the preprocessing processes is cleaning the data, which involves eliminating digits, special characters, and punctuation that are not required for categorisation.

Another preprocessing task is normalisation. It deals with issues pertaining to mixed cases, lowercase and uppercase variation cases. Additionally, it is crucial to normalise acronyms and abbreviations by having them represent a character using their expanded word forms. Furthermore, as words serve as the foundation for document representation or categorisation, identifying each word in a document is another preprocessing operation known as tokenisation. Word stemming and stop word removal are additional preprocessing duties. Not every news word is given the same weight during the text classification process. Certain words don't refer to any objects in a sentence; instead, they are used merely to complete the statement's grammatical framework. The sense of the phrases is not affected by the words that are used often in all documents.

3.4. Proposed Model

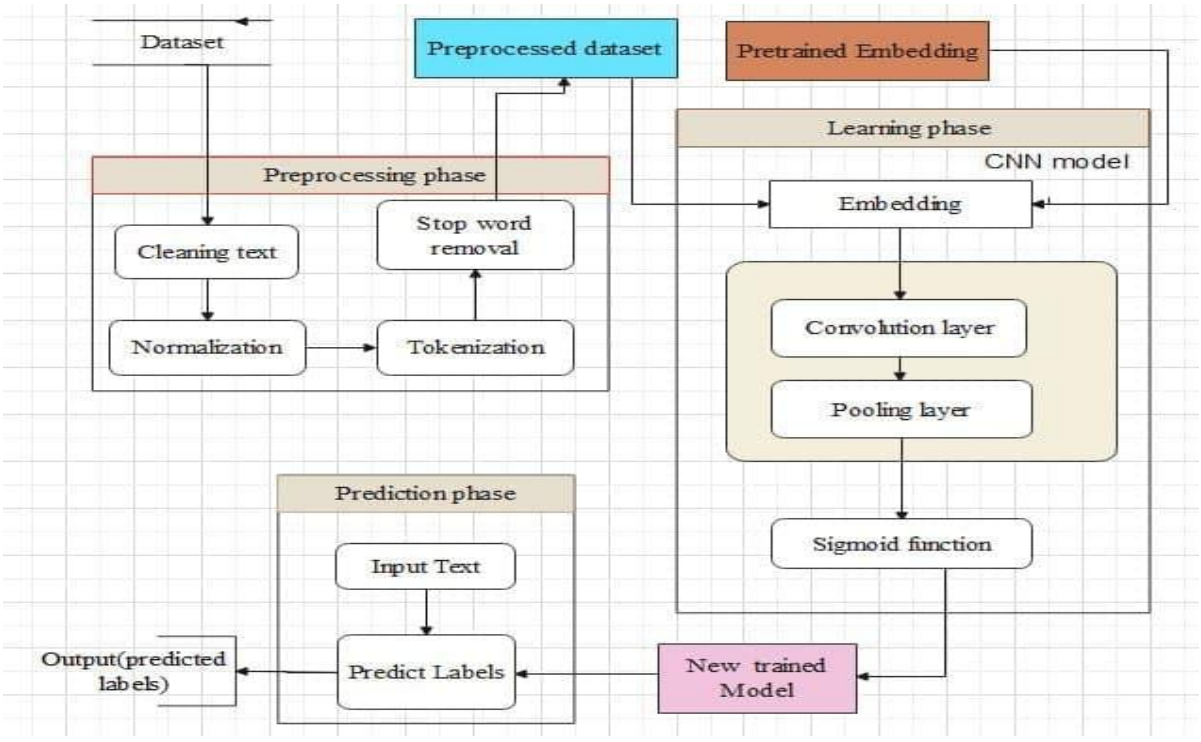

Figure 1. Architecture of the proposed System

The proposed model architecture, as shown in the above picture, is composed of a number of parts and subcomponents that can be separated into three stages. Nonetheless, the preprocessing, learning phase, such as Word-embedding, CNN-Learning model and prediction unit, are the fundamental elements that comprise the model. Normalisation, tokenisation, and stop word removal were among the subtopics covered in the preprocessing phase. The preprocessed dataset is the result of preprocessing. The next step is to convert the preprocessed dataset to Word2Vec. Each layer of the multi-label CNN has a distinct function [12]. It is made up of four layers: the sigmoid layer, the pooling layer, the convolution layer, and the embedding layer. The CNN model then receives Word2vec as input. After that, the CNN layers perform a number of tasks to create the news-trained model. Afterwards, applying what has been learnt. The prediction is made by the model, which is constructed through sample training. The prediction component is the last trained model from the training procedure. The input for prediction is the feature extractor's output, a file containing the word vector of terms in a given plain Afaan Oromo news text.

3.5. Multi-label learning algorithm

There are numerous algorithms for multi-label learning. In multi-label learning, every training set sample has several tags. Multi-label learning algorithm revision in problem-solving refers to the process of representing one instance by many labels at the same time. These algorithms can be classified as either algorithm applicability-based or problem transformation-based [13]. The problem transformation strategy alters the problem data for an existing method, whereas the algorithm application-based approach directly resolves the multi-label concerns. A support vector machine is used in a traditional machine learning model for text classification. Naïve Bayes and decision tree [14]. Furthermore, a text-training deep learning network, like CNN for sentence categorisation and RNN for news text classification, is another current approach for text multi-label text classification [15] [18]. One way to conceptualise the final output label is as the subsequent label classifier's input, given the importance of multiple labels. Recent NLP research has focused a lot of emphasis on CNN, a class of deep learning networks [16], [21], [22]. Compared to manually designed/set feature extraction, this class of deep learning is far more effective in automatically extracting features from raw input text. This type of deep learning is significantly more potent than manually designed/set feature extraction since it can automatically extract features from raw input text. For a large structure to be represented by a fixed-size vector and a CNN in NLP is used to extract local predictive elements from the structure and combine them.

CNN is frequently employed in news text categorisations, particularly in a language with a lot of resources like English, because deep learning typically needs a lot of data to train the model and provide a reliable forecast. Additionally, even in the lack of the vast amounts of data that it needs, it is used in low-resource languages like the Afaan Oromo language. CNN is an artificial neural network-based feed-forward network model structure that automatically extracts features and offers the benefits of weight sharing and local connections. The CNN model uses the produced feature vector as an input to classify news in Afaan Oromo. The model also takes advantage of our pre-trained Word2vec(W2V), which is produced from a set of domain-specific textual data that is domain-specific. CNN uses a sentence as the input of the CNN model, which is encoded in matrix form, and word vectors that have been transformed via word embedding (CBOW) as an input for text classification[19]. The matrix's rows each correspond to a single token, typically a word, when the phrase row inside the matrix matches a token. For instance, a 3X100 matrix would be the input for a text with three words utilising a 100-dimensional embedding. We generally employed the d-dimensional distributed word2vec. After that, it creates an n×d matrix input for a text of length n.

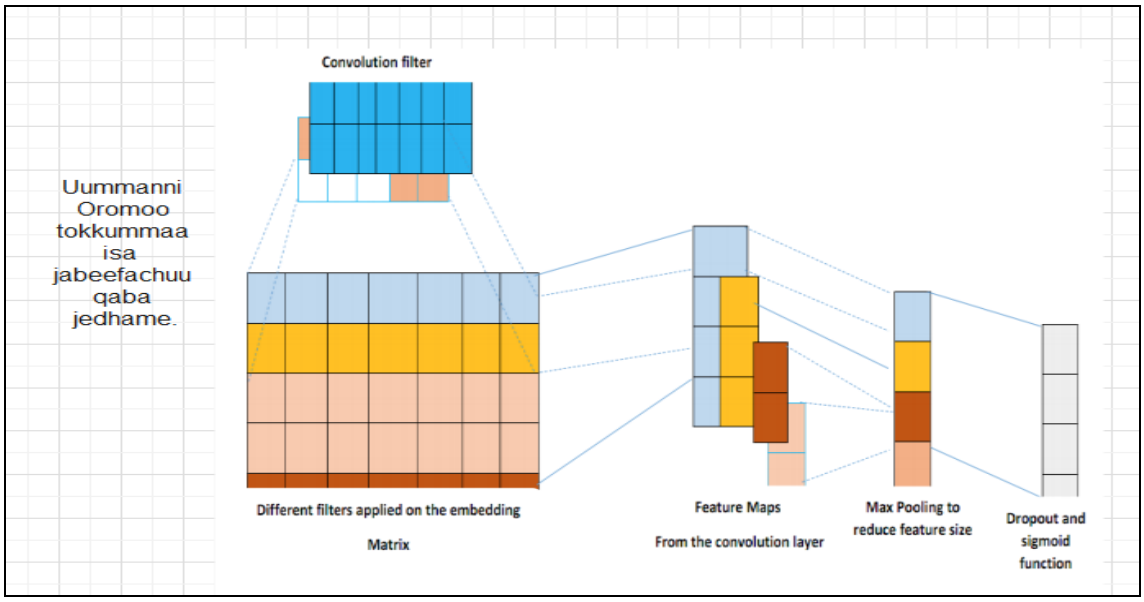

Figure 2. Operations of Data in CNN Model

To enter the input, an embedding layer converts the text inputs into a format that the CNN layer—the convolution layer—can use. Every word in a text document is transformed into a dense vector of a predetermined size in this instance. Layers inside CNN perform different functions depending on the text input. CNN uses convolution layers to extract features from the given text, while pooling layers are used from one network layer to the next to reduce the amount of output. We employed various pooling strategies while maintaining key features to minimise the size of the outputs and choose the maximum values that represent significant features [17][20].

4. Results and Discussion

In this research, many experiments are conducted utilising various Python libraries for the implementation of all the multi-label algorithms in order to evaluate our suggested model. Word2vec is trained using the Gensim Python open source toolkit. And also, the number of distinct network parameters must be chosen during the neural network design process. The following describes the model parameters needed for the suggested network: The hidden layer's neuron count: the number of nodes or processing units in the hidden layer. For instance, we used model parameters and hyperparameters for CNN training, like learning rate, Word embedding dimensions, number of epochs and batch-size 0.001,320,15,8 respectively. After the model is developed, the performance of the suggested model or solution has been assessed using a variety of performance criteria. The main metrics used to evaluate the performance of the proposed model are F1-score, precision, and recall. But we got different results based on the number of labels or classes. After an experiment was done using Word2vec and six and eight labels, we got 97.12% and 96.5 % accuracy, respectively. This experiment shows that as the number of labels increases, the accuracy of the model decreases. Therefore proposed CNN model achieves 87.52% testing accuracy and 96.5% training accuracy. And also by adjusting several parameters, including the network's layers, epoch size, batch size, and dropout, until the best fit model is discovered.

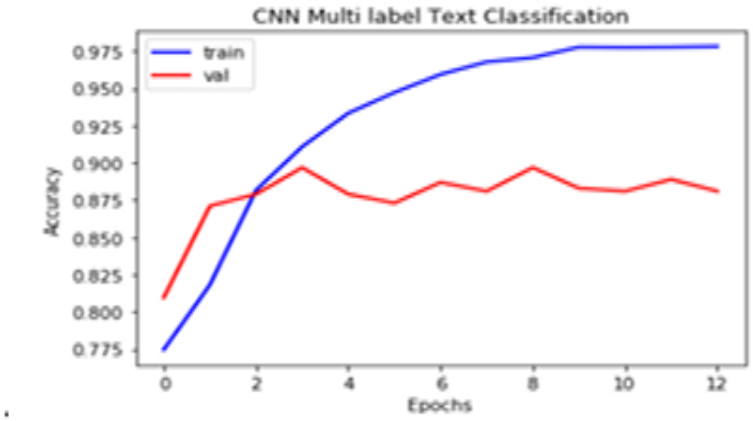

Figure 3. Shows the training accuracy of CNN

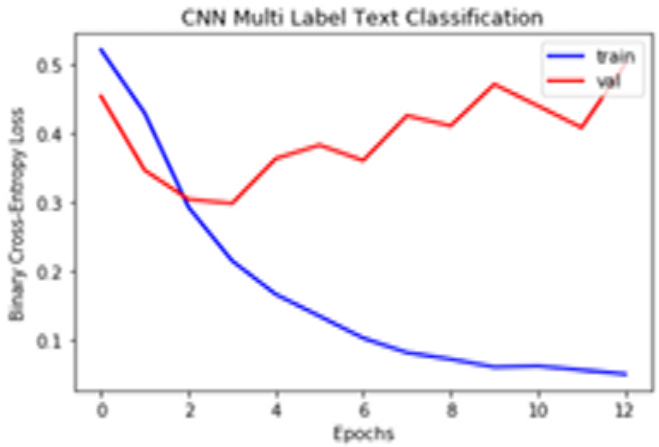

Figure 4. Shows the training loss of CNN

Training and validation accuracy rise as training and validation loss fall almost linearly, as can be shown below in Figure 3, the training loss and accuracy curve. There was no indication that over-fitting was occurring. Early Stopping, the Adam optimisation technique, and the activation function (ReLU) are to blame for this. Up until the final fourteen epochs, training and validation accuracy, as well as training loss and validation loss, are all near one another. The training and validation accuracy, training and validation loss show only slight variations over the curve. To choose the good performance, we also employed several train-test ratios. For instance, 40/60 was used and gave a very low result, 30/70 used and gave low result, 10/100 was used and gave good result, and finally, 20/80 was used and gave an intermediate result. Based on their results, we selected a 10/90 best train-test ratio proportion.

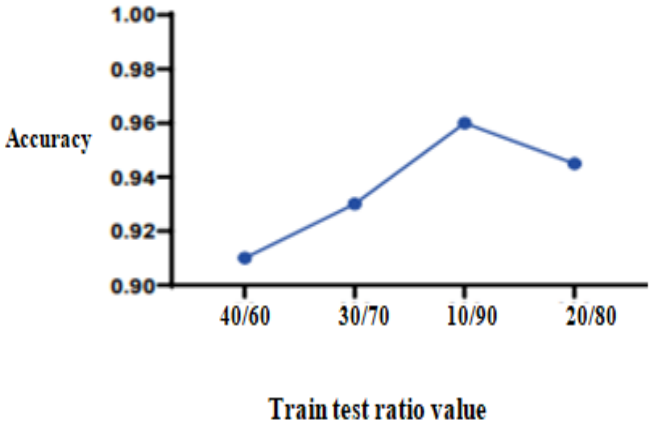

Figure 5. Example Variation of train test ratio values

5. Conclusion

In this work, we have developed CNN models for multi-label Afaan Oromo news text classification. To predict class labels, CNN uses different embeddings separately and in different combinations across multiple channels. When analysing unlabeled texts, the CNN models take into account contextual information to generate the best features to represent the contents. After comparing our suggested/developed models with one of the current models using the given parameter settings, we discovered that the suggested models outperform the original models in terms of accuracy. The suggested CNN models might be helpful for labelling news articles in Afaan Oromo. The goal of many researchers working on Afaan Oromo classifier development is to use various learning algorithms to boost the accuracy of the classification as the number of categories increases. Using various approaches, they attempted to use basic machine learning methods to address the calculation time issue. Unfortunately, all low-resource language researchers focus on flat, hierarchical, and multi-class classification types, but we created a model for multi-label text classification and attempted to apply it using a deep learning algorithm. Over 5640 Afaan Oromo news items are analysed experimentally over eight main news categories. The Python programming language served as our experimental platform for both text classification and word embedding. The problem with this study is that there isn't a published multi-label dataset; hence, the data size is smaller than in previous studies.

Acknowledgement

We acknowledge Mattu University for their initiation and support by providing different devices during the implementation, and Dr. Arulmurugan Ramu, Associate Professor, Department of Computer Science.

References

[1] C. C. Aggarwal and C. X. Zhai , “Mining text data springer Science & Business Media”, Journal of Springer Science DOI 10.1007/978-1-4614 3223-4, 2012

[2] Z. Younes, F. Abdallah, T. Denoeux and H. Snoussi “A dependent multi label classification method derived from the k- nearest neighbor rule”, EURASIP Journal on Advances in Signal Processing, 1-14 , 2011

[3] W.Tao, W., & D.Chang “ News text classification based on an improved convolutional neural network” Tehnički vjesnik, 26(5), 1400-1409 , 2019

[4] A.Tripathy, and A. Anand and S.K. Rath “Document-level sentiment classification using hybrid machine learning approach”, Knowledge and Information Systems, 53(3), 805-831, 2017

[5] A. Bakliwal et.al “sentiment analysis of political tweets” , Towards an accurate classifier. Association for Computational Linguistics, 49-58, 2013

[6] J. Fan, and Y. Keny. “High dimensional classification using features annealed independence rules”, Annals of statistics, 36(6), 2605, 2008

[7] M. Qiu et al), “Convolutional-neural-network-based Multilabel Text Classification for Automatic Discrimination of Legal Documents” Sensors and Materials, 32(8), 2659-2672, 2020

[8] W. Kelemework “Automatic Amharic text news classification”, A neural networks approach.” Ethiopian Journal of Science and Technology, 6(2), 127-137, 2013

[9] B.Jang, I. Kim, and J.W. Kim “Word2vec convolutional neural networks for classification of news articles and tweets” PloS one, 14(8), e0220976, 2019

[10] A.Diriba, “Automatic classification of Afaan Oromo News Text: The Case of Radio Fana”, etd.aau.edu.et/handle/123456789/21309, 2009

[11] N. Kannaiya, M. Raja, and N.Bakala “Multi-Label Classification for Afan Oromo Text by using Python Machine Learning Tools”, IJESC, Volume 10 Issue No.4, 2020

[12] H. Wubalem “*Multi Label Amharic Text Classification Using Convolutional Neural Network Approaches* Doctoral dissertation” IJESC, 2020

[13] C. Saunders, M. O.Stitson, J. Weston, R. Holloway, and L. Bottou “Support vector machine. Computer science” 1, 1-28, 2020

[14] J. Quinlan, “Induction of decision trees. Machine learning” ,1986
[15] P. Liu, X. Qiu, X and X. Huang, “Recurrent neural network for text classification with multi-task learning” arXiv preprint arXiv: 1605.05101, 2016
[16] A. Graves, A. Mohamed and G. Hinton, “Speech recognition with deep recurrent neural networks”, In 2013 IEEE international conference on acoustics, speech and signal processing , pp. 6645-6649, 2013
[17] D. Scherer, A. Müller and S. Behnke, “Evaluation of pooling operations in convolutional architectures for object recognition In International conference on artificial neural networks Springer, Heidelberg”, 2010
[18] R. Bulo ,“Afaan Oromo Multi-Label News Text Classification Using Deep Learning Approach”, International journal of research creativity and research ,2023
[19] K.M.Jimalo, R.B. Putchanuthala and Y.Assabie “ Afaan Oromo News Text Categorization using Decision Tree Classifier and Support Vector Machine: A Machine Learning Approach”,2017
[20] F. Van Meeuwen, “Multi-label text classification of news articles for ASDMedia,” 2013.
[21] B. Shruti, and, and G. Vishal, “Text News Classification System using Naïve Bayes Classifer,” vol. 3, pp. 209–213, 2014.
[22] N. Bakala, “A Two Steps Approach for Afan Oromo Nonfiction Text Categorization,” IJSRCSEIT , vol. 3, pp. 107–120, 2018.