Topic Classification of Tweets in the Broadcasting Domain using Machine Learning Methods

Tshephisho Joseph Sefara
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
tsefara@csir.co.za

Mapitsi Roseline Rangata
Next Generation Enterprises and Institutions
Council for Scientific and Industrial Research
Pretoria, South Africa
mrangata@csir.co.za

***Abstract*—Twitter is one of the microblogging sites with millions of daily users. Broadcast companies use Twitter to share short messages to engage or share opinions about a particular topic or product. With a large number of conversations available on Twitter, it is difficult to identify the category of topics in the broadcasting domain.**

This paper proposes the use of unsupervised learning to generate topics from unlabelled tweet data sets in the broadcasting domain using the latent Dirichlet allocation (LDA) method. Approximately six groups of topics were generated and each group was assigned a label or category. These labels were used to label the data by finding the dominating label in each tweet as the main category. Supervised learning was conducted to train six machine learning models which are multinomial logistic regression, XGBoost, decision trees, random forest, support vector machines, and multilayer perceptron (MLP). The models were able to learn from the data to predict the category of each tweet from the testing data.

The models were evaluated using accuracy and the f1 score. Linear support vector machine and MLP obtained better classification results compared to other trained models.

***Index Terms*—Topic modelling, machine learning, natural language processing, Twitter, topic classification**

I. Introduction

Twitter is a popular microblog that has become essential for both consumers and businesses to broadcast and exchange thoughts about a certain product or topic in real-time. The broadcast domain covers the conversation between users and the broadcasting radio or television channel. In most cases, radio or television channels share breaking news, sports, media personality, music, and interviews on Twitter and other local content. Users can react to these shared contents. But the reaction may diverge from the topic being shared to other new topics that users want to discuss. This creates mixed topics within a conversation and makes it difficult for Twitter to classify such conversations into relevant categories.

Recent machine learning algorithms can solve such problems by conducting both unsupervised and supervised learning in conversations to classify topics. Researchers have explored the use of machine learning algorithms in areas such as sentiment analysis and topic classification of Twitter data in different domains [1]. Supervised machine learning is an algorithm that learns from labelled training data to predict the label of the new data sample. It is commonly used to predict or discover trends and insights. There are different types of supervised algorithms, such as decision tree, support vector machine (SVM), naive Bayes, and others. Unsupervised machine learning is a method that trains models from unlabelled data (raw data) to categorise similar data into a number of groups. Topic Modelling is one of the most commonly used unsupervised machine learning techniques in natural language processing (NLP) to detect similar words in a document and cluster them into groups or topics. There are different types of topic modelling techniques known as non-negative matrix factorisation (NMF), latent semantic analysis (LSA), latent Dirichlet allocation (LDA), and others [2]. Topic modelling using LDA will be explored in this paper to enable the use of supervised learning for topic classification.

Topic classification is a challenging problem, as discussed in [3]–[5]. There are different tweet attributes that can be used to solve this problem. Location data is one of the attributes that can be used to discover popular topics at the country level that show the importance of tweets on a particular topic. Tweets are used in topic modelling [3]–[6] to create a cluster or group of topics that later need to be analysed into categories or labels. These categories enable the use of machine learning algorithms such as SVM [5], naive Bayes [3], decision trees [3] to automatically classify new tweets into correct categories.

With a large number of conversations available on Twitter, it is challenging to identify the category of topics in the broadcasting domain. This paper proposes a unique method that uses NLP tasks, such as topic modelling and machine learning models, to label data that do not have labels. In addition, the paper provides baseline methods and findings in the broadcasting domain.

Below we summarise the main contributions of this paper as follows.

- A technique to label data is proposed using both unsupervised and supervised learning.
- The machine learning methods that predict tweets topic.
- The data is made available on GitHub[1].
- We release the source code on GitHub[1] to allow for benchmarking and further development of the technique.

[1] https://github.com/JosephSefara/Topic-Classification-of-Tweets

This paper is organised as follows. The background is discussed in the next section. Section III explains the proposed methodology including data collection, data engineering, machine learning models, and model evaluation. The findings and analysis of the results are discussed in Section IV, while Section V concludes the paper with future work.

II. Background

Vadivukarassi et al. [1] propose a method to predict the category of tweets collected from the Twitter streaming API. Their aim is to study or compare the classification performance of supervised machine learning techniques in Twitter data. They first collected 1160 text documents from various categories from websites to train models and extract features to classify each category. Then, they used the models trained from a text document corpus on Twitter data to predict the category of each tweet. Their tests were conducted on a naive Bayes classifier, a linear SVM, and a multinomial naive Bayes classifier with Term Frequency-Inverse Document Frequency (TFIDF) as a feature extraction. Their findings show that the multinomial naive Bayes classifier performed better with an accuracy of 72.83%.

Lee et al. [3] developed a method to classify trending tweets into 18 various categories such as music, fashion, politics, and others. They explored supervised classification techniques such as multinomial naive Bayes classifier, K-nearest neighbour, SVM and logistic regression, where they found that multinomial naive Bayes classifier gave better results in Twitter trending topic classification. Tiwana et al. [7] also classified trending tweets into various categories such as entertainment, politics, technology, and sports; they have also incorporated context-based meaning to improve tweet classification. Furthermore, they did a comparison for the performance of classification of trending tweets by applying supervised classifiers; SVM, linear regression, logistic regression, and naive Bayes on their Twitter dataset. Dilrukshi et al. [8] classified Twitter news into 12 different predefined categories using SVM. In their analysis, they have performed dimensionality reduction to reduce the noise in the data while still preserving important features. Sriram et al. [9] classified tweets into categories such as news, opinions, deals, and private messages using eight features, one nominal (author), and seven features within the tweets such as slang, currency, percentage signs, "@username", and others. Wang et al. [10] performed a classification of private tweets that contain sensitive information into 14 categories. They used supervised machine learning, such as a naive Bayes classifier, for data classification. The data set was processed by both bag-of-words and TF-IDF. They have also used five-fold cross-validation for classification accuracy evaluation. Vashisth et al. [11] performed gender classification on Twitter data, built a labelled Twitter data set that includes gender associated with a tweet and explored NLP techniques such as Bag-of-Words, TF-IDF, and unsupervised classifiers; logistic regression, SVM, and others. Alzanin et al. [12] proposed a method to categorise tweets that are in Arabic language into five distinct categories. They tested their data using supervised classification methods such as SVM, naive Bayes, and random forest. Their findings suggest that SVM and naive Bayes were the strongest performers with an accuracy of around 98%. Labelled Latent Dirichlet Allocation (L-LDA), SVM, and L-LDA combined with SVM were proposed as approaches for identifying and categorising transportion-related tweets into incident, congestion, construction, special event, and other events. The authors used SVM to identify tweets about transportation. L-LDA incorporated with SVM was the best performer in categorising tweets with an accuracy of 98.3% [13]. Ullah et al. [14] proposed a method to identify and categorise tweets related to disasters and requesting help when disasters occur. Tweets are categorised into categories such as request for shelter, food, money, and others. They used rule-based and logistic regression, where their findings show that logistic regression is a better performer.

Topic detection on Twitter is a challenging task, as discussed by [6] who proposed a method to detect topics on Twitter using the label propagation model. Kumar et al. [4] aim to find the most frequent topics at the country level that show the importance of tweets on a particular subject; the authors used location data and tweets to detect topics at the country level using non-negative matrix factorisation topic modelling. The use of unsupervised learning to label data is a challenging task, as the data on Twitter are not labelled. Cahyani et al. [5] use machine learning methods such as SVM to classify the relevance of trending topics on Twitter, since most popular topics are not related to the content being discussed, and users are taking advantage of these hot topics to gain public attention. On average, the authors achieved an accuracy of 86%.

This paper proposes the use of LDA to generate topics using the Gensim Python library [15]. The generated topics are then used to label the data using the most dominant topic within each tweet in the data set. We then train supervised learning on the labelled data using machine learning algorithms.

III. Methodology

This section discusses the overall architectural design, data acquired, machine learning methods, and evaluation techniques.

This paper proposes a unique method shown in Fig. 1 that uses unlabelled data to create a classifier model that can label new data points in the broadcasting domain. Figure 1 shows the step-by-step architecture of the proposed method. The initial step in acquiring unlabelled data from Twitter. The data set is pre-processed to remove noise and improve the quality of the data set. Topic modelling using LDA is applied to the data set to generate latent topics. We analyse the topics to infer a better name/label for each cluster of topics. The data set is labelled using a dormant topic or label for each tweet. At this stage, the data set is labelled and we can apply supervised learning. The features are extracted by applying the TFIDF vectoriser for each tweet. The feature normalisation technique is applied to the vectorised data set. Different machine learning techniques are trained on the dataset and compared to measure the performance of the models.

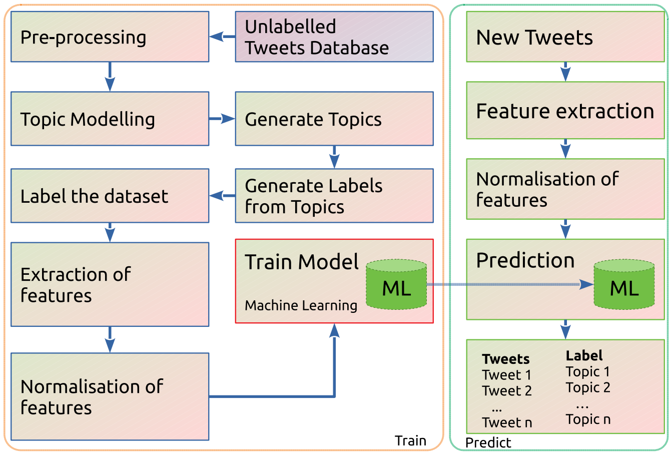

Fig. 1. Proposed architecture.

A. Data

The data was acquired from Twitter by subscribing to Twitter streaming using APIs. Data were collected from February 2023 for a period of 3 months. On average, most tweets were collected during the day. The following algorithm was used to collect tweets in real time.

Algorithm 1 Rules used to collect tweets

```
1: The tweet must be authored in South Africa.
2: The tweet must not be a retweet.
3: The must have a place.
4: The tweet must have a coarse geo-location.
5: The tweet must match one or more keywords.
```

The data structure of the tweets contains the following variables:

- User: The owner of the tweet.
- User mentions: Users that are mentioned in the tweet.
- Text: The actual tweet.
- Sensitive: The indicator that is true if a tweet is sensitive, otherwise is false.
- Place: The name of the place where the tweet was originated.
- Location: The GPS location where the tweet was originated.
- Language: The identified language of the tweet.
- Id: The unique identifier of the tweet.
- Date: The date on which the tweet was composed.

Table I shows the frequency of tweets per user. This table can also show whether a user is a robot or not. Most robots have a large number of tweets that clearly show an outlier. In this case, the authors checked the tweets and verified that the users were not robots. Exploratory data analysis was performed on the data and Fig. 2 shows the word cloud of the most mentioned hashtags. The hashtag *sabcnews* is mostly mentioned by most users because SABC is a public broadcaster of the government of South Africa with the majority of the viewers. Followed by the hashtag *DStvPrem* which is the main premier league in South Africa. The map in Fig. 3 shows the location data of where the tweets originate. Some tweets do not have a specific location but contain only the country name. Most tweets are originated from major cities such as Johannesburg which contains 125 tweets followed by Pretoria with 86 tweets. Cape Town contains 72 tweets followed by Sandton with 66 tweets. The rest of the tweets are originated from 157 places within South Africa.

TABLE I
TOP 10 POPULAR USERS.

Name	Frequency
KhayaJames	21
NjabuloMntungwa	16
TheDjManSA	15
Velile_Mnyandu	11
djvuslo	9
john_kaalie	8
Imsollyntuli	7
OmphilePhix	7
nes_kamogelo	7
DJMAOSH3	6

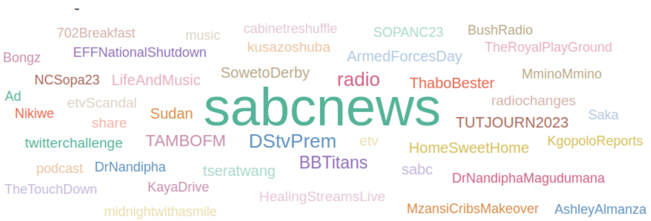

Fig. 2. Word cloud of the most mentioned hashtags.

B. Data pre-processing

The data contain noise and characters that are not important for topic modelling and machine learning modelling. The data contained 1107 tweets. The data were cleaned by removing: (i) uniform resource locators (URLs), (ii) quotes, (iii) new lines, (iv) Twitter handle, (v) hashtags, (vi) punctuation marks, (vii) white space, (viii) numbers, (ix) emails, (x) and English stopwords.

C. Topic Modelling

Latent Dirichlet allocation (LDA) is a common topic modelling algorithm that uses an unsupervised clustering method to discover topics from the data set. This method helps to label the data set that is not labelled. We performed LDA topic modelling using gensim[2] Python library on the data set to generate six clusters or groups of topics. We manually analysed the clusters to find the common label of each cluster, which are as follows: (i) Radio station segments, (ii) Network

[2] https://radimrehurek.com/gensim/

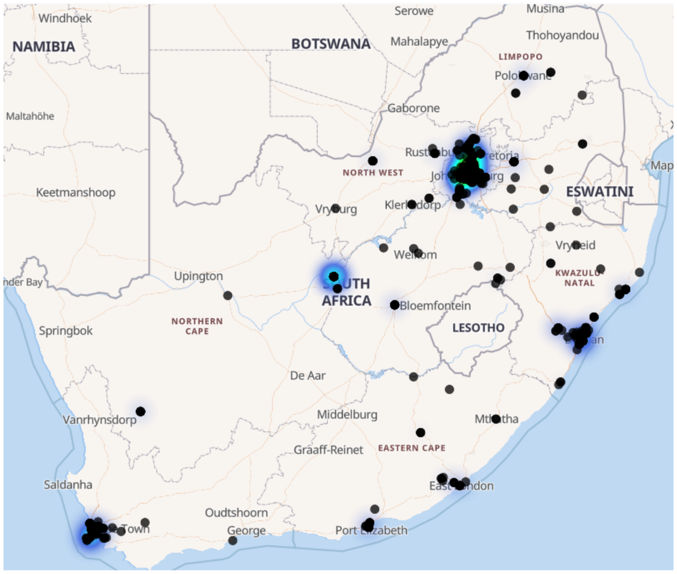

Fig. 3. Map.

signal, (iii) TV Music, News, and Sports, (iv) Radio Interviews, (v) TV Broadcasting, (vi) and General News.

The data set was labelled using new labels inferred from topic modelling. Table II shows the distribution of labels in the data set. The data set has 348 *Radio Station Segments* labels followed by *Network Signal* with 167 labels, and *TV Music, News, and Sports* has 166 labels, *Radio Interviews* has 160 labels, *TV Broadcasting* has 159 labels, and the least label is *General News* with 107.

TABLE II
LABEL DISTRIBUTION

Label	Total
Radio Station Segments	348
Network Signal	167
TV Music, News, and Sports	166
Radio Interviews	160
TV Broadcasting	159
General News	107
Total	1107

D. Machine Learning Models

This section discusses the machine learning models used for topic classification of tweets.

The data consist of 1107 tweets that were divided into 80% for training and 20% for testing. Each label was divided accordingly into the same 80% for training and 20% for testing. Data were transformed using the TFIDF vectoriser and then normalised using the standard scaler[3]. Vectorised data were used to train three machine learning models, which are defined as follows:

- Multinomial logistic regression is a supervised classification model that is modified from binary logistic regression to provide probabilities between 0 and 1 using the cross-entropy loss function. We use the logistic regression implemented in the scikit-learn [16] Python library, where we set the parameter *multi_class* to multinomial.
- XGBoost is a gradient-boosting Python library designed for lightweight and flexible distribution. It uses trees to provide parallel tree boosting to solve tasks quickly and accurately [17]. We used the XGBoost implemented in Python[4] where we set the parameter *objective* to *multi:softmax* to allow multiclass classification.
- MLP is a classifier model made up of three fully connected neural network layers, the first of which is the input layer, followed by one hidden layer, and the final of which is the output layer. We used the MLP implemented in the scikit-learn Python library [16]. Table III shows the parameters used to fit the MLP model.
- Decision tree is a supervised learning algorithm that is used mostly for regression and classification analysis. We implement the decision tree implemented in the scikit-learn Python library [16].
- Random forest is a supervised learning algorithm used mainly for regression and classification analysis. It builds trees on different samples and computes the average in regression, and selects the majority vote in the classification problem. We used the random forest implemented in the scikit-learn Python library [16].
- Linear SVM is a supervised learning algorithm that is used mostly for regression and classification analysis. We used the linear SVM implemented in the scikit-learn Python library [16].

TABLE III
MLP PARAMETERS

Parameter name	Value
Number of neurons	100
Solver	adam
Activation function	RELU
Batch size	auto
Initial learning rate	0.001

E. Evaluation

The performance of the machine learning models was evaluated using the following metrics:

1) Confusion matrix: The confusion matrix is a matrix that is applied to measure the classification performance of machine learning models. Figure 4 shows the confusion matrix, where (i) True positive (TP) is when a tweet is predicted positive and it is actually positive, (ii) False positive (FP) is when a tweet is predicted positive and it is negative, (iii) True negative (TN) is when a tweet is predicted negative while it is negative, (iv) False negative (FN) is when a tweet is predicted negative while it is actually positive.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[4] https://xgboost.readthedocs.io/en/stable/python/index.html

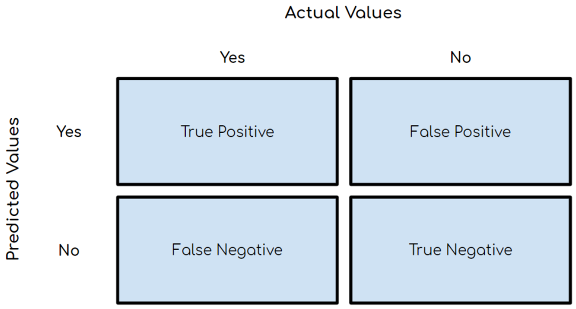

Fig. 4. Confusion matrix.

2) Accuracy: Accuracy is the number of correctly predicted tweets out of the total tweets. Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \quad (1)$$

3) F1 score: F1 score is the harmonic mean of precision and recall. F1 score is determined by this equation.

$$F1score = \frac{2*TP}{2*TP+FN+FP} \quad (2)$$

IV. Results and Discussions

This section discusses the findings generated from the experiments. Data were divided into 80% for training the models and the rest for evaluating the models. The findings in this section are based on the evaluation data that were not used during the model development. The six models discussed in Section III-D were fitted and trained on the training data. Most of the parameters were default and the parameters mentioned in Section III-D were applied. Since the data was small, the experiments were conducted on a normal computer that does not need specific resources.

We use LDA to generate categories (depicted in Table II) from the Twitter data collected; these labelled data generated by LDA were then used as training data sets to train the supervised models. Figure 5 depicts the comparison of the performance of different classifier models on testing data. All models were implemented on scikit-learn[5] except XGBoost.

In Figure 5, we observe that SVM has the best accuracy performance with 68% followed by MLP and XGBoost with 66% and 59%, respectively. The decision tree performs the least with 52% followed by logistic regression with 55% and the random forest attained 56% which is 1% higher than logistic regression. Since the data had an uneven number of labels, the classification accuracy is not enough to understand the performance of the models; therefore, we analyse the performance of the models using the f1 score.

For a better analysis of the models, we computed the f1 score from the confusion matrices in Fig. 6 and Fig. 7 for SVM and MLP, respectively. These matrices help to understand the prediction performance of each label within the data. As shown in both matrices, the label *General News* was correctly predicted for 78% f1 score out of 20 tweets by both SVM and MLP, while the label *Radio Station Segments* was correctly predicted for 76% by SVM and 71% by MLP. The label *TV Music, News, and Sports* was correctly predicted for 72% by MLP while SVM predicted for 67%. The most confusing label was *TV Broadcasting* which was predicted for 42% and 37% by SVM and MLP, respectively. This happens when most *TV Broadcasting* tweets contain features that occur in other labels. As we can observe from both matrices, *TV Broadcasting tweets* are confused with *Radio Station Segments* by both SVM and MLP. This can be an indication that the topic modelling method used can be improved by merging these two labels together. This way, the classification results may produce different results; this remains the future work.

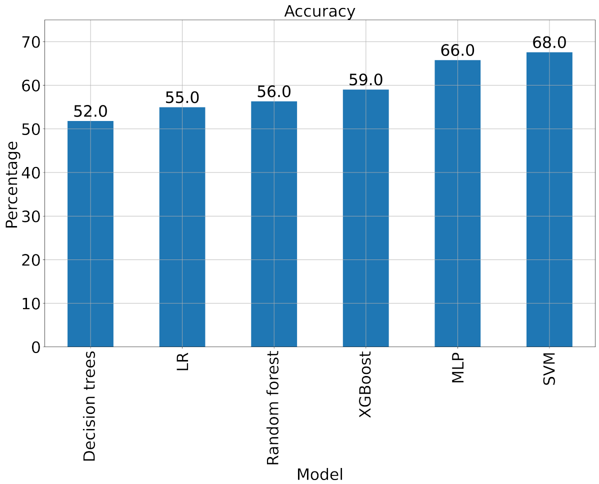

Fig. 5. Model accuracy trained on testing data.

The findings could not be compared to the literature, as we could not recreate the models found in the literature to train on the proposed data. Therefore, we are publishing the code and data on GitHub[6] to allow comparison with other future work.

This paper proposed a data set in the broadcasting domain for the South African context. The models were trained and produced the baseline results.

V. Conclusions and Future work

This paper proposed a data set containing tweets collected over a three-month period that contain tweets from the broadcasting domain. The data was geographically fixed to the South African context. The paper also proposed a unique method that uses an unsupervised technique called topic modelling using LDA to generate six groups of topics from unlabelled Twitter data. The LDA model performed well since the data was cleaned and lemmatised. The authors analysed the group of topics to come with a label or category for each group. We then

[5] https://scikit-learn.org/stable/index.html

[6] https://github.com/JosephSefara/Topic-Classification-of-Tweets

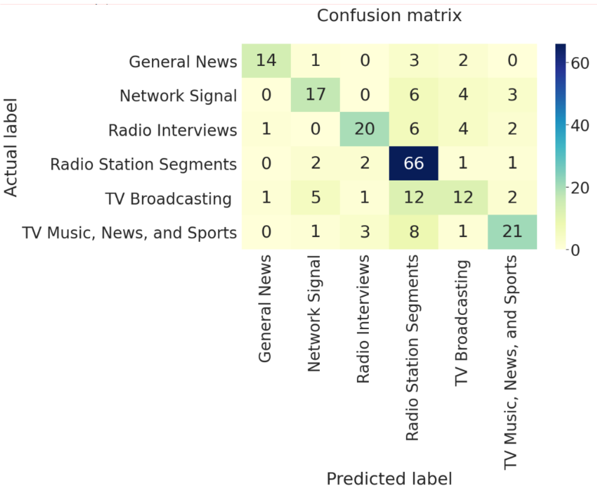

Fig. 6. SVM confusion matrix on the testing data.

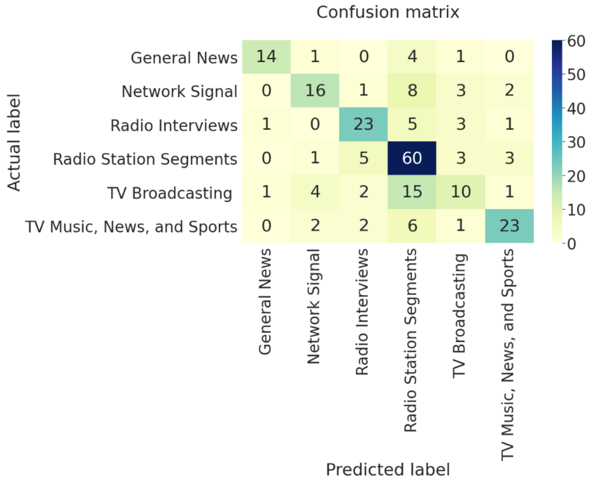

Fig. 7. MLP confusion matrix on the testing data.

calculated the most dominant label for each tweet in the data set and assigned it as the label. Once the data is labelled, we performed supervised learning. Machine learning classifiers were trained to classify tweets into different categories or labels. The best classification results were obtained from SVM with a classification accuracy of 68% and MLP with a classification accuracy of 66%. Since the data was unbalanced, the f1 score was used to analyse the performance of the machine learning models.

In summary, the literature was reviewed, the proposed data was discussed, and the performance of classification models was reviewed. From the results, SVM and MLP have indicated better accuracy performance under limited data.

Future work will improve the methods discussed by using a larger data set so that modern deep learning methods, such as transformers, recurrent neural networks, long short-term memory networks, can be used to perform text classification and generation. The data used in this paper can be found on GitHub[7].

References

[1] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "A comparison of supervised machine learning approaches for categorized tweets," in *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. Springer, 2019, pp. 422–430.

[2] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645–5657, 2015.

[3] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 251–258.

[4] J. Kumar and R. Sunitha, "Classification model for identification of country level tweet prominence from worldwide tweets using location data," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 1947–1953.

[5] D. E. Cahyani and A. W. Putra, "Relevance classification of trending topic and twitter content using support vector machine," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2021, pp. 87–90.

[6] D. Huang and D. Mu, "Topic detection in twitter based on label propagation model," in *2014 13th International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, 2014, pp. 97–101.

[7] M. S. Tiwana, F. Javeed, M. I. Lali, H. Dar, and M. Bilal, "Comparative analysis of context based classification of twitter," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, 2018, pp. 1–5.

[8] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using svm," in *2013 8th International Conference on Computer Science & Education*. IEEE, 2013, pp. 287–291.

[9] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, pp. 841–842.

[10] Q. Wang, J. Bhandal, S. Huang, and B. Luo, "Classification of private tweets using tweet content," in *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*. IEEE, 2017, pp. 65–68.

[11] P. Vashisth and K. Meehan, "Gender classification using twitter text data," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–6.

[12] S. M. Alzanin, A. M. Azmi, and H. A. Aboalsamh, "Short text classification for Arabic social media tweets," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 9, pp. 6595–6604, 2022.

[13] S. M. Khan, M. Chowdhury, L. B. Ngo, and A. Apon, "Multi-class Twitter data categorization and geocoding with a novel computing framework," *Cities*, vol. 96, p. 102410, 2020.

[14] I. Ullah, S. Khan, M. Imran, and Y.-K. Lee, "RweetMiner: Automatic identification and categorization of help requests on Twitter during disasters," *Expert Systems with Applications*, vol. 176, p. 114787, 2021.

[15] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[7] https://github.com/JosephSefara/Topic-Classification-of-Tweets