Domain-specific Sentiment Analysis of Tweets using Machine Learning Methods

Tshephisho Joseph Sefara[0000−0002−5197−7802] and Mapitsi Roseline Rangata[0000−0002−7624−2415]

Council for Scientific and Industrial Research, Pretoria, South Africa
tsefara@csir.co.za,mrangata@csir.co.za

Abstract. Most general sentiment analysers degrade quality when tested on Tweets in the broadcast domain. This domain covers both radio and television broadcast. This paper proposes domain-specific data in the broadcast domain. Furthermore, it proposes the use of machine learning methods for the sentiment analysis of tweets in this domain. Data were collected from Twitter using Twitter application programming interfaces. The data were preprocessed, and most special characters and emoticons were not removed, as sentiment analysis involves the use of opinions and emotions which are expressed using emoticons and other characters. The data were automatically labelled using a pre-trained sentiment analyser to enable the use of supervised learning on the data. Two supervised machine learning methods, such as XGBoost and multinomial logistic regression (MLR), are trained and evaluated on the data. The performance of the models was affected by two factors; limited data and the use of a general sentiment analyser to label the data in a specific domain.

Keywords: Sentiment analysis · Machine learning · XGBoost · Logistic regression · Text classification · NLP · AI.

1 Introduction

Sentiment analysis is a computational analysis in Natural Language Processing (NLP) that identifies and categorises a sentiment (people's opinion or feeling or expression) in a text about a certain event, topics or product, and others [18]. Twitter platform is one of the famous microblogging social media sites that generates a large volume of data in real time with millions of tweets posted daily, which enables researchers such as Agarwal et al. [1] to build supervised learning models to classify tweets into different categories such as positive, neutral and negative. Machine learning models such as support vector machine (SVM), naive Bayes, logistic regression, decision trees, random forest, and others are being used to create sentiment analysers. [2, 7, 9, 8] created a sentiment analyser using SVM, while [20] used naive Bayes. The use of text classification using machine learning as sentiment analysis has dominated the current research approach. Ramadhan et al. [24] used logistic regression to create a sentiment analysis model. Their model was tested on different sizes of training and testing data, also on different

numbers of K-fold cross-validation, where the results showed that the higher the training size, the better the performance, and the higher the K-fold the better the performance. While sentiment analysis is being carried out as a multiclass classification, Prabhat [23] created a sentiment analyser using logistic regression and naive Bayes as a binary classification where the labels are positive and negative.

Most sentiment analysers are trained for English text, only a few work has been done for other languages [15, 14]. Social media data is unstructured and may contain text in different languages that Twitter may not support in terms of automatic language identification during data acquisition using Twitter APIs. Mabokela et al. [14] created a multilingual annotated sentiment data set for South African languages, including Setswana, Sepedi, and English. The authors also created an application that enables automatic labelling of sentiments in a multilingual context.

Sentiment analysis is being applied in many domains, including viewership [16], political [6], food retail [5], phone retail [10], and can be applied for customer satisfaction [25]. Most sentiment analysers work with textual data, recently [28] conducted a study on the analysis of sentiments from both textual data and speech data. The authors used supervised learning methods, such as SVM for textual data and fuzzy logic for speech data. These machine learning methods recorded good results in speech and language processing [27].

In the broadcasting domain, radio and television (TV) are one of the broadcasting services that are used for communication. In this paper, the sentiments of radio and television services are analysed to obtain the opinions and feelings of people shared on Twitter about certain services such as TV shows, TV channels, news, etc., offered in South African broadcast. There are many sentiment analysers for social networks that use general data, but such analysers are trained on general domain data and consequently degrade performance when testing in the broadcasting domain. Therefore, this paper proposes a domain-specific dataset and an analysis of the sentiment analysis models trained in the broadcasting domain using limited data acquired from Twitter.

The main contributions of this paper are as follows.

- The paper proposes models for sentiment analysis in the broadcasting domain.
- Automatic labelling of data using pre-trained model to allow analytics using supervised learning.
- Data are made available on GitHub[1].
- We publish the source code on GitHub[1] to enable benchmarking of the results.

This paper is organised as follows. The next section explains the background. Section 3 explains data acquisition, data processing, machine learning models, and model evaluation. Section 4 discusses the findings and analysis, while Section 5 concludes the paper with future work.

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

2 Background

This section discusses the recent applications of sentiment analysis in different domains. In the viewership domain, Malik et al. [16] performed sentiment analysis to predict viewers on streaming TV shows using Twitter data. For each TV show, the authors collected tweets using a certain keyword for a specific period. They also explored the performance of various supervised machine learning methods for the classification of sentiments from tweets about TV programmes.

In the food retail domain, El Rahman et al. [5] conducted a sentiment analysis on Twitter data to show which restaurant is popular between McDonald's and KFC. They used different supervised classification algorithms. Maxtent, as one of the explored, was a top performer for McDonald's and KFC with a four-fold cross-validation of 74% (McDonald's) and 78% of a four-fold cross-validation (KFC).

For the food supply domain, based on the public sentiments expressed on Twitter, Neogi et al. [21] conducted a sentiment analysis on India's farmer protests. Their study used a lexicon-based method; TextBlob, together with an exploration of machine learning classification methods, to evaluate public sentiment in India about farmer protests. TF-IDF was used as a feature extractor in their modelling simulation. Random Forest was one of the classifiers investigated in their study and was found to be the best performer in terms of accuracy.

In the phone retail domain, Hasan et al. [10] analysed a sentiment analysis on Twitter data related to particular phone brands to obtain sentiment to show which phone brand is popular between the iPhone and Samsung, where the authors have collected 1000 tweets. The authors explored the use of both term frequency-inverse document frequency (TF-IDF) and bag-of-words (BoW) to generate features for modelling. Their results showed that both TF-IDF and BoW improve the performance of the results, showing that the iPhone was more popular compared to Samsung devices.

In the health domain, Vijay et al. [29] analysed Twitter data related to covid-19 for sentiment classification in certain parts of India from November 2019 to May 2020 to evaluate the progression of virus spread and people's opinions based on covid-19 and the measures implemented by the government to curb virus spread. TextBlob was used in their study to analyse Twitter data obtained during that time period.

In the political domain, Elbagir et al. [6] analysed tweets from the 2016 US elections for sentiment classification to categorise tweets into 4 categories; positive, negative, neutral, highly negative, highly positive using Valence Aware Dictionary and sEntiment Reasoner (VADER) [11] and Natural Language Toolkit (NLTK). On the other hand, Joyce et al. [12] also performed sentiment analysis in the 2016 US elections for comparison of sentiment analysis approaches known as lexicon-based sentiment analysis and machine learning classifiers for sentiment. They used manual and automated labelled tweets in both approaches. Their results show that the lexicon-based sentiment analysis performed better. In their research, the author [30] analysed the 2020 US presidential election for sentiment analysis to obtain opinions on Twitter between the two candidates running for

president, and also to predict the outcome of the election, five machine learning classifications were explored in the data for performance evaluation. Among the five classifiers, multilayer perceptron (MLP) was the top performer. The sentiment analysis method to extract the individual's point of view on Twitter about electricity hikes in a developed and developing country was developed by Kaur et al. [13]. In their study, they proposed a lexicon-based method known as VADER to infer sentiment polarity together with classification algorithms. TF-IDF was used for feature extraction. Four algorithms were explored for performance evaluation. Random forest and decision tree were the most successful models.

3 Methodology

In this section, we discuss the proposed architectural design, the acquired data, the exploratory analysis of the acquired data, the proposed supervised machine learning methods, and the evaluation techniques of the models.

3.1 Architectural Design

This paper proposes a method illustrated in Fig. 1 that uses a pre-trained sentiment analysis model to label the new data that are collected. Figure 1 shows the architecture of the proposed method. The first step is to acquire data from Twitter that are not labelled. We pre-process the data to remove unnecessary characters and to enhance the quality of the acquired data. We extracted the features using the TF-IDF vectoriser for each tweet. The pre-trained sentiment analyser is applied to the data to create labels. We labelled the data using the labels generated by the pre-trained sentiment analyser for each tweet. At this stage, the data are labelled and supervised learning can be applied to the labelled data. We trained machine learning methods on the labelled data and compared the performance of the trained models. We further analysed the predictions of the models using confusion matrix.

3.2 Data

The data was collected from Twitter using Twitter streaming APIs in real time. Data are collected from February 2023 for a period of 3 months. The following rules were sent to Twitter to match the tweets in real time. We did not retrieve historical Twitter data.

3.3 Exploratory Data Analysis

This section analyse the acquired data to determine the quality of the data, the relationship between the entities, the patterns within the data, and errors and anomalies. We use the following entities from the data:

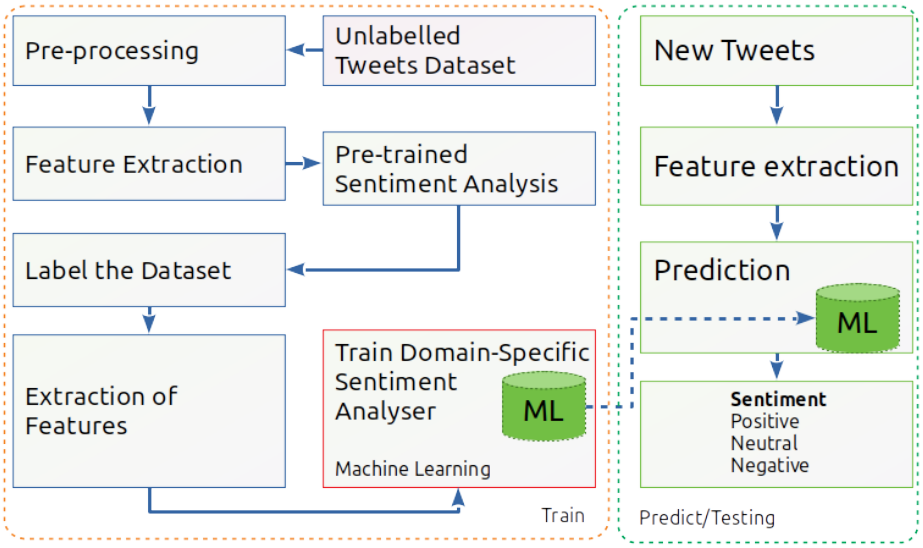

Fig. 1. Proposed architecture.

Algorithm 1 Rules used to collect tweets on Twitter APIs

1: The tweet must be original or quoted tweets not retweet.
2: The tweet must be created by a user in South Africa.
3: The tweet must be tagged with a place.
4: The geo-location of the tweet should be near a tagged place.
5: The tweet must contain one or more keywords provided to Twitter.

– User: The account authored a tweet.
– Hashtags: The hashtags mentioned in a tweet.
– User mentions: Accounts that are mentioned in a tweet.
– Text: The actual tweet.
– Place: The geographical place where a tweet was authored.
– Date: The date and time on which a tweet was composed.

The data consisted of 1125 number of tweets where some tweets are authored by same user. Figure 2 shows the top 10 most frequent social media influencers or users. This figure validates whether some users are robots or not. We validate the users with the most frequent tweets and ascertain that the users are not robots. About 173 tweets are authored by the same influencers, and 572 tweets are authored by unique influencers.

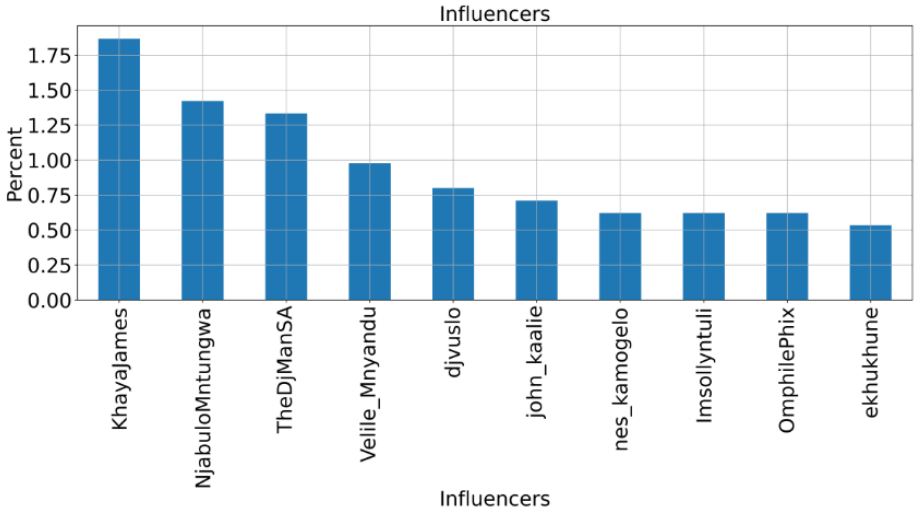

Fig. 2. Top 10 social media influencers.

The data contained tweets generated by users in South Africa. The country name will be shared when a user does not share the actual location. Figure 3 shows the top 10 most frequent places where a tweet was authored. The first place with the most frequent tweets is South Africa followed by the following cities: Johannesburg, Pretoria, Cape Town, Sandton, Randburg, Midrand, Durban, Roodepoort, and Centurion. This figure shows that most of the tweets are authored from populated and developed cities in South Africa. The most frequent tweets are authored from 83 places, whereas other tweets are authored from 78 unique places.

The tweets may mention other users. Figure 4 shows the top 10 most frequent user mentions with *etv* being the highest followed by *DStv*, *METROFMSA*, *Official_SABC1*, and others.

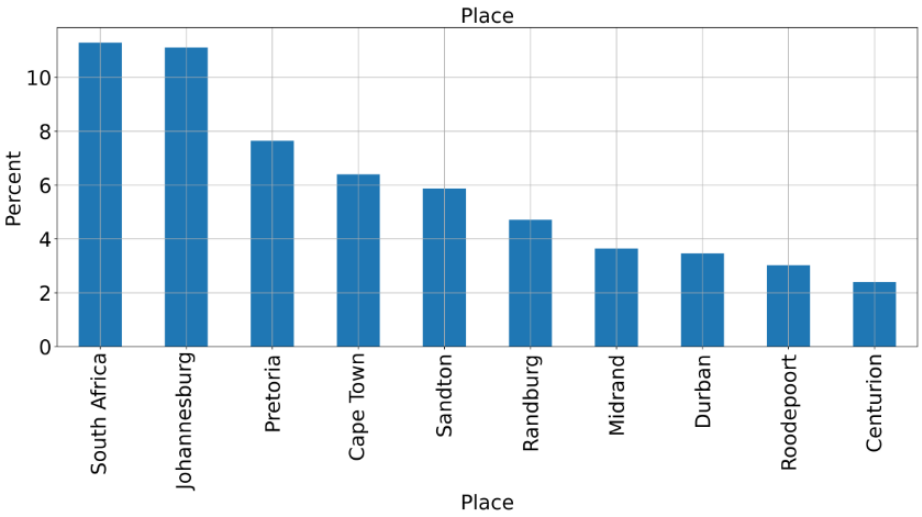

Fig. 3. Top 10 places.

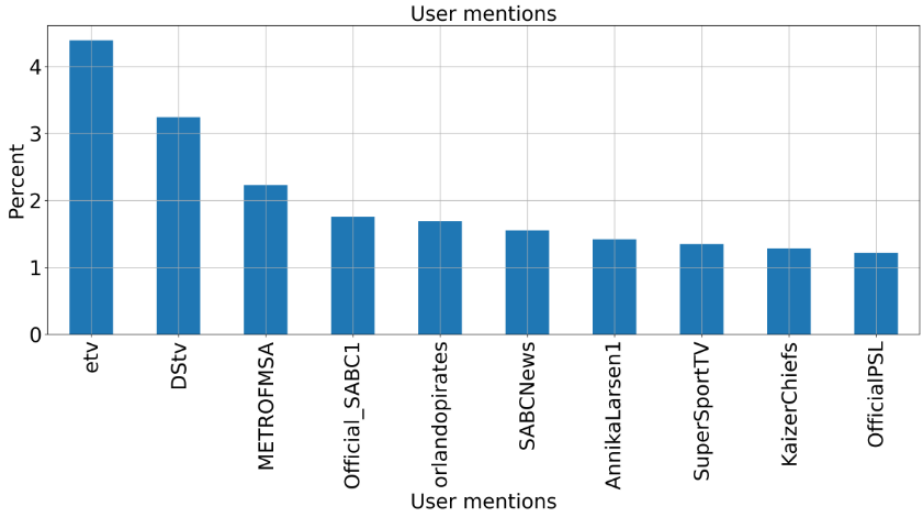

Fig. 4. Top 10 user mentions.

Tweets may contain hashtags. Figure 5 shows the top 10 most frequent hashtags with *sabcnews* having the highest mentions followed by *DStvPrem*, *radio*, and others.

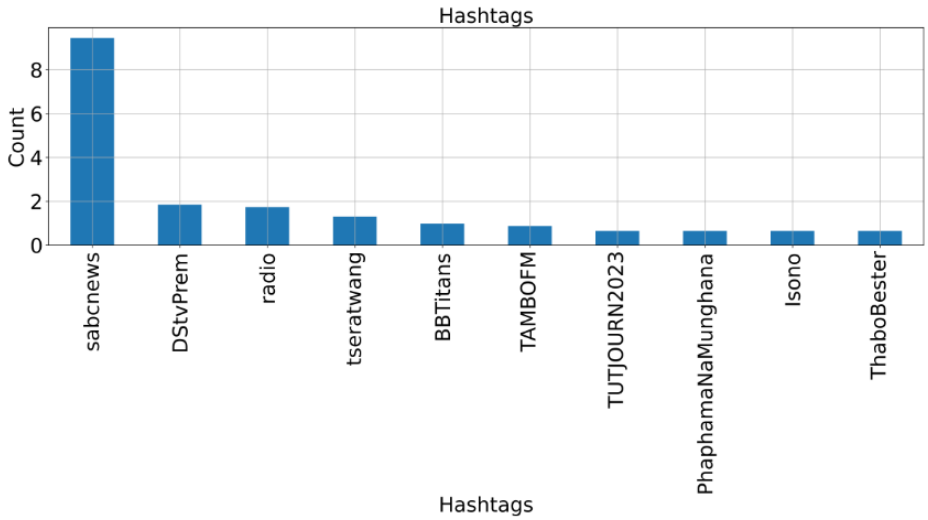

Fig. 5. Top 10 hashtags.

The tweets contain a date entity that tells when a tweet was authored. Figure 6 shows the time distribution of tweets for a 24-hour period in a day. It shows that most tweets are authored from 5:00 in the morning until 19:00 in the afternoon. Since Twitter uses GMT timezone, the local timezone for South Africa is GMT+2 which implies that most tweets are authored from 7:00 until 21:00, but around 17:00 fewer tweets are authored, this happens when most users travel from work to their homes.

3.4 Data preprocessing

The data was pre-processed to validate that the tweet exists. In sentiment analysis, special characters and emoticons express emotions and meaning, which is an important feature during model building. We preserve the original tweet without removing any characters.

3.5 Pre-trained Sentiment Analysis

The data was analysed using VADER sentiment analysis method to generate labels. Figure 7 depicts the findings of the VADER sentiment analysis, with more than 47% of the tweets being positive and around 30% neutral and around

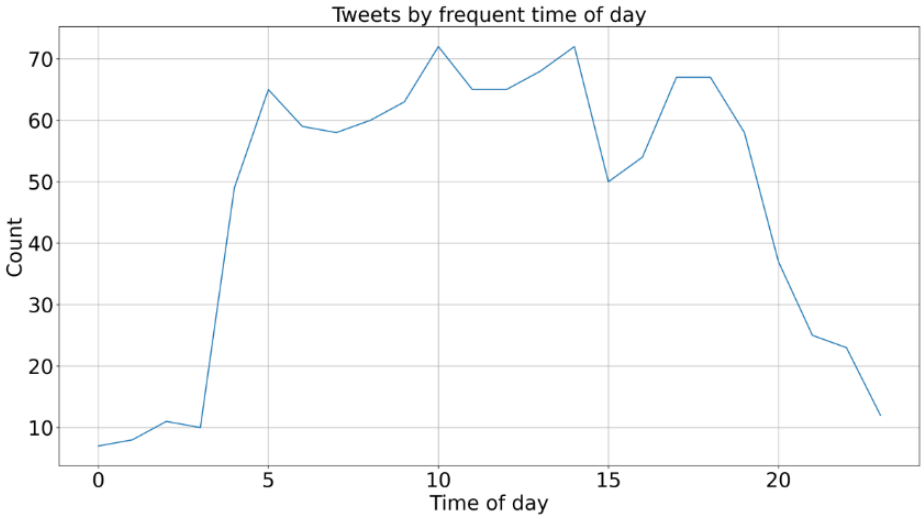

Fig. 6. Frequent tweets during 24 hour time in a day.

23% negative. These labels are used to create domain-specific sentiment analyser models that aid in analysing why and when a tweet is being classified as negative/neutral/positive.

3.6 Domain-specific Sentiment Analysis Models

The data was first transformed using the TF-IDF vectoriser[2] with up to 3 trigrams.

This step is the most important part of feature engineering. The following machine learning models are trained on the data.

- **Multinomial logistic regression (MLR)**: is a supervised machine learning model that is trained on the data. MLR model can output the predictions and the importance of features.
- **XGBoost**: is an implementation of gradient-boosting decision trees that is trained on the data. XGBoost model can output the predictions and importance of features.

3.7 Evaluation

The quality performance of the machine learning models was tested using the confusion matrix, the F measure, and the accuracy metrics:

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

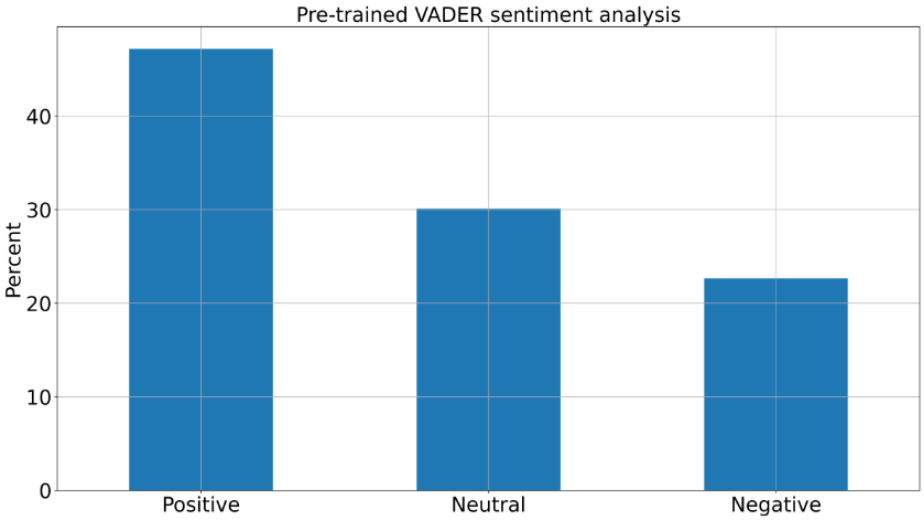

Fig. 7. Pre-trained sentiment analysis using VADER.

Confusion matrix The confusion matrix is defined as a matrix that is applied to measure the classification performance of machine learning models. This matrix is used to calculate the precision, recall, F measure, and accuracy of machine learning models. The matrix also helps to understand the accuracy of predicting each label or class. The confusion matrix is depicted in Fig. 8, where:

1. True positive (TP) is when a tweet is correctly predicted as positive,
2. False positive (FP) is when a tweet is incorrectly predicted as positive,
3. True negative (TN) is when a tweet is correctly predicted negative,
4. False negative (FN) is when a tweet is incorrectly predicted negative.

Accuracy Accuracy is defined as the total number of correctly predicted tweets divided by the total number of tweets. Accuracy is determined as:

$$Accuracy = \frac{TN + TP}{FP + FN + TN + TP} \tag{1}$$

F Measure F Measure is the harmonic mean of precision and recall. F Measure is defined by the following equation.

$$FMeasure = \frac{2 * TP}{2 * TP + FP + FN} \tag{2}$$

4 Results and Discussions

This section explains the findings of the domain-specific sentiment analysis models. The acquired data was splitted into 80% for building the model and the rest

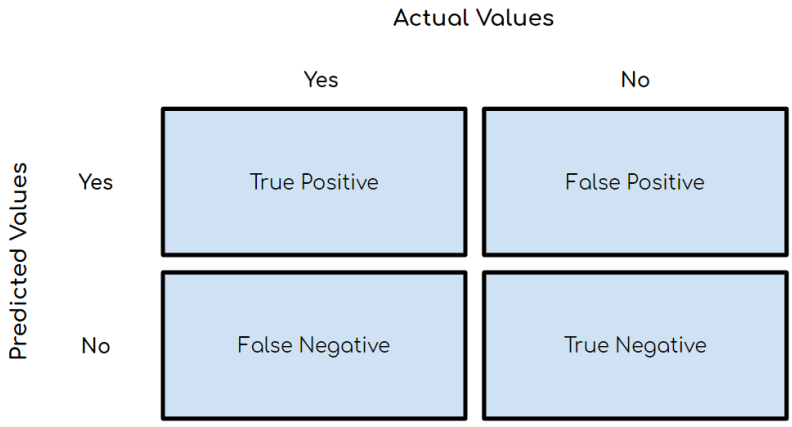

Fig. 8. Confusion matrix.

for testing the model. The test data were not used to build the model to allow for a proper evaluation of the models on unseen data. MLR and XGBoost were fitted on the training dataset, and most of the parameters are default. The logistic regression implemented in scikit-learn [22] was used as the first model in which the parameter *multi_ class* was set to *multinomial* to make it suitable for multiclass classification, since this type of model is well-known for binary classification. The XGBoost model implemented in [3] was used as the second model in which the parameter *objective* was set to *multi:softmax* to make it suitable for multiclass classification, as this type of model is well-known for binary classification. MLR and XGBoost obtained an accuracy of 56% and 57%, respectively. The quality of the performance of the models depends on the size of the data [19], and since our data were limited, this affected the performance of the results in this domain. Another factor that affects the accuracy is the use of a general model (VADER) to infer labels in the broadcasting domain. Furthermore, the F measure was calculated for MLR and XGBoost, which obtained 45% and 56% respectively, as shown in Table 1.

Table 1. Model prediction results in percentage based on the proposed data.

Authors	**Model**	**Accuracy**	**F measure**
Baseline	MLR	56%	45%
Baseline	XGBoost	57%	56%

For a better analysis of the models, we computed the confusion matrices in Fig. 9 and Fig. 10 for MLR and XGBoost, respectively. These matrices aid

to understand the prediction performance for each label within the data. The models were better in predicting the positive sentiments with 49% for MLR and 34% for XGBoost. The XGBoost predicted neutral sentiment better than MLR, while MLR incorrectly predicted most of neutral and negative sentiments as positive sentiments.

This paper proposed a limited Twitter data in the specific domain (broadcasting domain) for the South African context. Models were built and generated the baseline results in this domain.

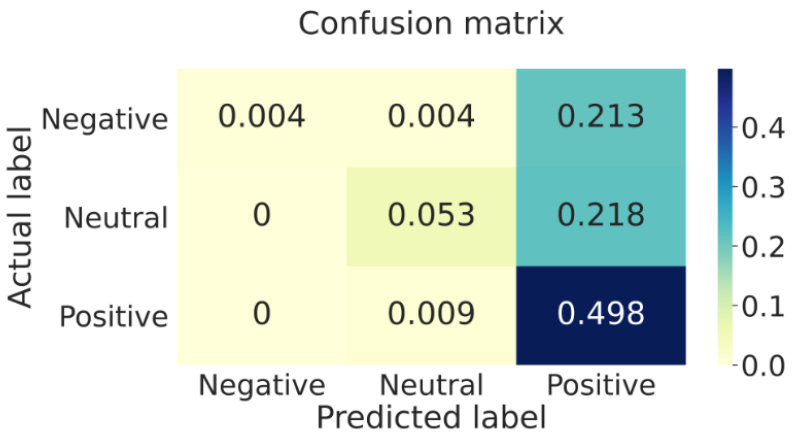

Fig. 9. MLR confusion matrix.

5 Conclusions and Future work

This paper proposed the use of unlabelled data to create sentiment analysis models for a broadcasting domain. The data were labelled using a pre-trained VADER sentiment analyser to enable us to conduct supervised learning on the data. The two machine learning methods; MLR and XGBoost are trained and tested on the data. The models are evaluated using accuracy and the F measure. The models were evaluated using the confusion matrix to show the performance of each label. The results revealed that a variety of factors affected the models' performance; limited data, and using a general model to label the data that are specific to broadcasting domain.

In conclusion, the paper addresses an essential issue of domain-specific sentiment analysis, which is often challenging due to the unique language usage in different fields. The methods used, namely XGBoost and MLR, are widely recognised in machine learning and have demonstrated effectiveness in sentiment analysis. The approach of keeping emoticons and special characters can

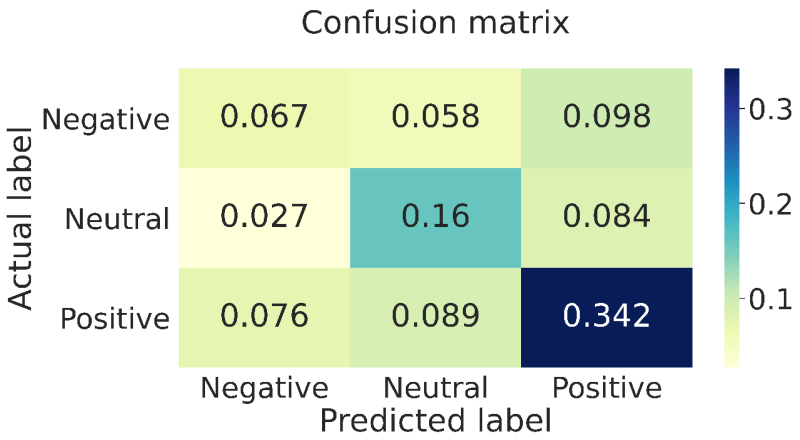

Fig. 10. XGBoost confusion matrix.

potentially improve the models' performance, since they often carry significant sentiment information in tweets. The use of limited data is a known problem in this field, and there are existing methods to mitigate this problem [17, 4].

In Future, this work will focus on:

- Acquiring more data in this domain and updating it on Github.
- Manually labelling the data to enable the use of both unsupervised and supervised machine learning on the data.
- Training deep learning models [26].

References

1. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.J.: Sentiment analysis of Twitter data. In: Proceedings of the workshop on language in social media (LSM 2011). pp. 30–38 (2011)
2. Anjaria, M., Guddeti, R.M.R.: Influence factor based opinion mining of Twitter data using supervised learning. In: 2014 sixth international conference on communication systems and networks (COMSNETS). pp. 1–8. IEEE (2014)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
4. Dhole, K., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahadiran, A., Mille, S., Shrivastava, A., Tan, S., et al.: NL-Augmenter: A framework for task-sensitive natural language augmentation. Northern European Journal of Language Technology **9**(1) (2023)
5. El Rahman, S.A., AlOtaibi, F.A., AlShehri, W.A.: Sentiment analysis of Twitter data. In: 2019 international conference on computer and information sciences (ICCIS). pp. 1–4. IEEE (2019)

6. Elbagir, S., Yang, J.: Twitter sentiment analysis using natural language toolkit and VADER sentiment. In: Proceedings of the international multiconference of engineers and computer scientists. vol. 122, p. 16 (2019)
7. Fouad, M.M., Gharib, T.F., Mashat, A.S.: Efficient Twitter sentiment analysis system with feature selection and classifier ensemble. In: The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018). pp. 516–527. Springer (2018)
8. Gautam, G., Yadav, D.: Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. In: 2014 Seventh international conference on contemporary computing (IC3). pp. 437–442. IEEE (2014)
9. Ghiassi, M., Skinner, J., Zimbra, D.: Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with applications **40**(16), 6266–6282 (2013)
10. Hasan, M.R., Maliha, M., Arifuzzaman, M.: Sentiment analysis with NLP on Twitter data. In: 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2). pp. 1–4. IEEE (2019)
11. Hutto, C., Gilbert, E.: Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media. pp. 216–225 (2014)
12. Joyce, B., Deng, J.: Sentiment analysis of tweets for the 2016 US presidential election. In: 2017 ieee mit undergraduate research technology conference (urtc). pp. 1–4. IEEE (2017)
13. Kaur, P., Edalati, M.: Sentiment analysis on electricity Twitter posts. arXiv preprint arXiv:2206.05042 (2022)
14. Mabokela, K.R., Celik, T., Raborife, M.: Multilingual sentiment analysis for under-resourced languages: A systematic review of the landscape. IEEE Access (2022)
15. Mabokela, R., Schlippe, T.: A sentiment corpus for South African under-resourced languages in a multilingual context. In: Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages. pp. 70–77 (2022)
16. Malik, H., Shakshuki, E.M., et al.: Approximating viewership of streaming TV programs using social media sentiment analysis. Procedia Computer Science **198**, 94–101 (2022)
17. Marivate, V., Sefara, T.: Improving short text classification through global augmentation methods. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) Machine Learning and Knowledge Extraction. pp. 385–399. Springer International Publishing, Cham (2020)
18. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal **5**(4), 1093–1113 (2014)
19. Mokgonyane, T.B., Sefara, T.J., Manamela, M.J., Modipa, T.I.: The effects of data size on text-independent automatic speaker identification system. In: 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD). pp. 1–6. IEEE (2019)
20. Neethu, M., Rajasree, R.: Sentiment analysis in Twitter using machine learning techniques. In: 2013 fourth international conference on computing, communications and networking technologies (ICCCNT). pp. 1–5. IEEE (2013)
21. Neogi, A.S., Garg, K.A., Mishra, R.K., Dwivedi, Y.K.: Sentiment analysis and classification of Indian farmers' protest using Twitter data. International Journal of Information Management Data Insights **1**(2), 100019 (2021)

22. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. the Journal of machine Learning research **12**, 2825–2830 (2011)
23. Prabhat, A., Khullar, V.: Sentiment classification on big data using naive Bayes and logistic regression. In: 2017 International Conference on Computer Communication and Informatics (ICCCI). pp. 1–5 (2017). https://doi.org/10.1109/ICCCI.2017.8117734
24. Ramadhan, W., Astri Novianty, S., Casi Setianingsih, S.: Sentiment analysis using multinomial logistic regression. In: 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC). pp. 46–49 (2017). https://doi.org/10.1109/ICCEREC.2017.8226700
25. Ramirez, L.A.P., Marquez, B.Y., Magdaleno-Palencia, J.S.: Neuromarketing to discover customer satisfaction. In: Guarda, T., Portela, F., Augusto, M.F. (eds.) Advanced Research in Technologies, Information, Innovation and Sustainability. pp. 191–204. Springer Nature Switzerland, Cham (2022)
26. Sefara, T.J., Zwane, S.G., Gama, N., Sibisi, H., Senoamadi, P.N., Marivate, V.: Transformer-based machine translation for low-resourced languages embedded with language identification. In: 2021 conference on information communications technology and society (ICTAS). pp. 127–132. IEEE (2021)
27. Sefara, T.J., Mokgonyane, T.B.: Emotional speaker recognition based on machine and deep learning. In: 2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC). pp. 1–8. IEEE (2020)
28. Vashishtha, S., Susan, S.: Inferring sentiments from supervised classification of text and speech cues using fuzzy rules. Procedia Computer Science **167**, 1370–1379 (2020)
29. Vijay, T., Chawla, A., Dhanka, B., Karmakar, P.: Sentiment analysis on covid-19 Twitter data. In: 2020 5th IEEE international conference on recent advances and innovations in engineering (ICRAIE). pp. 1–7. IEEE (2020)
30. Xia, E., Yue, H., Liu, H.: Tweet sentiment analysis of the 2020 US presidential election. In: Companion proceedings of the web conference 2021. pp. 367–371 (2021)