EVALUATING THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS AND LEXICAL ANALYSERS FOR THE SENTIMENT ANALYSIS OF TAMIL TWEETS

A.R.F. Shafana
Department of Information and Communication Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka
arfshafana@seu.ac.lk

M.I.F. Nihla
Department of Management and Information Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka
nihla@seu.ac.lk

A.F. Musfira
Department of Information and Communication Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka
ameermusfi@seu.ac.lk

M.M.F. Naja
Department of Software Engineering
Universiti Malaya
Kuala Lumpur, Malaysia
mmfnaja@gmail.com

***Abstract*— The proliferation of social media and microblogging platforms allows the public to readily publish their thoughts and opinions. Especially, Twitter generates a huge amount of textual data that could be used to perform fruitful analyses. Sentiment Analysis is one such analysis that has gained wider research attention that helps researchers gain insights into public opinion on a specific topic. Despite the fact that many efforts have been pledged in this regard, limitations still exist for non-English languages. This study is an attempt to compare the performance of the popular lexical-based approach and classical machine learning-based approach for classifying the sentiments of a low-resource language like Tamil. The study extracted the Tamil tweets using the TwitterAPI for this purpose which resulted in 45852 tweets in total. Two subject experts in the field randomly selected a set of 300 tweets and then classified to their respective sentiments. This annotated data was then used as the ground truth, and six separate evaluations were performed on the pre-processed and cleaned data. Two lexical-based analysers (VADER and TextBlob) and four machine learning algorithms (Random Forest, eXtreme Gradient Boosting, Support Vector Machine, and Gaussian Naïve Bayes) were used in this analysis. The outcome of the results indicated that the machine learning algorithms were still effective over the lexical-based analysers for Tamil sentiment analysis. Specifically, the Support Vector Machine achieved the highest performance score of all. This study serves as empirical evidence for the interested society in performing sentiment analysis on Tamil language tweets.**

Keywords— Tamil Language, Sentiment Analysis, Lexicon, Twitter API, Supervised Learning

I. INTRODUCTION

Twitter is an excellent microblogging platform where millions of users directly communicate information [1]. Recent statistics report that the total number of registered Twitter users rounds up to 152 million [2], and the monthly visit count is over 500 million people [1]. The limited character count feature of Twitter has made it emerge as an appropriate tool for text analysis [3]. This creates a vast opportunity for scholars to conduct many studies using the ample data available.

Opinion mining on Twitter data is one such study which enables data scientists to categorize the opinions from people as either being positive or negative or neutral, based on the polarity of the tweet. This is generally conducted either using a machine-learning approach or a lexical-based approach. The machine learning approach typically follows the methodology presented by Pang et al. [4] that utilises the supervised learning approach, where the manually annotated data is used to train the classifier. This classification is often binary (positive or negative). However, the limitation in obtaining the annotated data makes this approach less feasible to apply to a new set of data [5]. In addition, the same set of annotated data produces a less accurate result when applied to another context [6]. Previous studies have employed various machine learning algorithms such as Gaussian Naïve Bayes, Support Vector Machine, eXtreme Gradient Boosting, and Random Forest for sentiment classification of textual data.

On the other hand, the lexical-based approach is gaining popularity as this approach does not necessarily depend on pre-defined annotations and supports a ternary classification. Of the lexical-based approaches, the analysing tools have shown more efficient classification accuracy than its traditional benchmarks such as Affective Norms for English Words (ANEW), Linguistic Inquiry and Word Count (LIWC), the General Inquirer, Senti WordNet, and certain other [6]. VADER and TextBlob are two most prevalently used tools for performing sentiment analysis in recent studies with proven higher accuracy.

A vast of literature analysing the applicability of the above approaches is especially for the English Language. Tamil Language, a Dravidian Language, has no specific standard annotated corpora for sentiment analysis besides SAIL data, which makes the number of studies much limited [7]. Thus, this study aims to apply existing tools and algorithms on the Tamil Twitter data and compare their performance in analysing sentiments. Accuracy, precision, recall, and F1-measure were used as the standard metrics for the evaluation. To the best of our knowledge, the sentiment analysis on Tamil Twitter data is minimal, and this comparative approach would be beneficial for this language. Thus, the outcome of this research can be used by scholars who are interested in

performing sentiment analysis on Tamil Twitter data to wisely choose the efficient approach based on their context of the application.

II. Related works

Sentiment analysis is a machine learning method wherein machines assess and classify sentiments, emotions, and opinions conveyed in the form of text or voice regarding any specific topic or object [8]. As the amount of textual data on the internet grows, most current research is focused on sentiment analysis. Researchers are particularly interested in developing and designing a system that recognises and categorises feelings expressed in textual form. As a result, the two most extensively utilised methodologies in this direction are the machine-learning approach and the lexical-based approach.

By comparing text terms with pre-prepared lexicons, the lexicon-based technique assigns precise weights to each word based on the polarity of the word to which it belongs and identifies the attitudes. The study by Al-Shabi, [9] takes a lexicon-based approach, focusing on five of the most well-known lexicons used in the field of sentiment analysis on Twitter data that, includes VADER, SentiWordNet, SentiStrength, Liu and Hu opinion lexicon, and AFINN-111. Overall classification accuracy and the F1-measure were compared, and the results demonstrated that the classification accuracy with the Vader lexicon is higher for both positive and negative sentiments. Another study uses a lexicon-based technique for sentiment analysis. The authors use NLTK, Text blob, and VADER Sentiment analysis tools to categorise movie reviews and compare them to identify the optimum tool for sentiment classification. The results of this study's experiments show that VADER excels in the Text blob [10]. Nur Syahirah et al. [5] compared the performance of two lexicons, VADER and TextBlob, in performing sentiment analysis on 7,997 tweets. According to the findings, both lexicons have an adequate accuracy rate, with VADER outperforming TextBlob for English tweets.

The other extensively used approach for sentiment analysis is machine-learning-based algorithms. Tamil movie reviews were categorised into positive and negative categories where the study used machine learning methods such as SVM, Decision tree, Maxent classifier, and Naive Bayes [7]. TamilSentiwordnet has been used to extract features. The study found that SVM performed well for the classification of reviews. Shihab & Jing [6] suggested using a combination of character-based Deep Bidirectional long short-term memory neural networks (DBLSTM) for Tamil tweets analysis. According to the findings of another study, SVM and RNN classifiers that utilise the TF-IDF and Word2vec features of Tamil text scored higher than grammar rules-based categorisation and certain other classifications which employ the existence of words, Term Frequency (TF), and Bag of Words (BoW) since features work better [11].

III. Materials and Methods

A. Data Collection and Preprocessing

Twitter data has been acquired using the public streaming Twitter API to retrieve tweets in the Tamil language alone. Scraping was performed on a random date, and the result consisted of 45852 tweets in total. The query did not specify a keyword in order to retrieve the maximum number of tweets available in the Tamil Language.

The next process involves data preprocessing, where the tweets are cleaned by first removing the null and duplicate values. Irrelevant information to this specific study, such as URLs, images, usernames, and emoticons, was also removed. This is rather an important step since the accuracy of the final classifier relies on this. The Natural Language Toolkit (NLTK) was used to obtain the processed data that is comprised of the main Tweet message. Special characters, retweets, URLs, user mentions, and unnecessary punctuations were also removed using regular expressions in Python. The consecutive steps were done to prepare the dataset suitable for sentiment analysis.

B. Development of Ground Truth

The preprocessed tweets were provided to the subject experts in the field to develop a ground truth for the comparison. Tweets were provided to two subject experts, and the tweets with the same class of sentiments from both experts were filtered for further analysis. This count was also limited to 300 tweets which could be clearly segregated as neutral, negative, and positive, where the count is similar to the study of Nur Syahirah et al. [5].

C. Analysis using lexical analysers

The sentiment analysis using lexical analysers was performed simultaneously using VADER and TextBlob tools. VADER is a sentiment analysing tool that has been widely used in previous studies. This is a parsimonious rule-based sentiment analysing tool [12]. TextBlob is another tool based on NLTK corpora for sentiment classification [13]. Both tools have been excellently performing ternary classification whereby the tweets can be classified as either positive, negative or neutral. The threshold values for the polarity used in the study were in line with the previous studies employing these tools for sentiment analysis. Table I provides a detailed view of the threshold values.

TABLE I. Threshold values of the lexical-based analysers

Sentiments	Analysers	
	***VADER* (Compound Score from Analyser)**	***TextBlob* (Polarity Score from Analyser)**
Positive	>= 0.05	> 0
Neutral	< 0.05 AND > -0.05	= = 0
Negative	<= -0.05	< 0

D. Analysis using machine learning algorithms

The machine learning approach typically uses a supervised learning approach where a set of data is first used to train the dataset, and another set is used to validate the performance of the trained model. There have been many different algorithms in this approach. Support Vector Machine is one such algorithm that has been consistently involved in classifying sentiments on textual data [14,15,16]. XGBoost is gaining popularity as an efficient ensemble algorithm in the field of sentiment analysis [17]. Similarly, the use of the Naïve Bayes classifier [18, 19] and Random Forest Classifier [20, 21] can be widely seen in studies involving the classification of sentiments using tweets with higher accuracies.

IV. RESULTS

The performance of the lexicons and algorithms was calculated by comparing them against the ground truth. The metrics used for the evaluation are accuracy, precision, recall and the F1 measure, as represented by the following formulae.

$$\text{Accuracy} = \frac{\text{number of correctly classified positive,negative and neutral tweets}}{\text{number of positive,negative neutral tweets in ground truth}} \quad (1)$$

$$\text{Precision} = \frac{\text{number of correctly classified positive,negative,neutral tweets}}{\text{number of positive,negative,neutral tweets classified by lexicons}} \quad (2)$$

$$\text{Recall} = \frac{\text{number of correctly classified positive,negative and neutral tweets}}{\text{number of positive,negative neutral tweets in ground truth}} \quad (3)$$

$$\text{F1-Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The measures were calculated for each of the approaches used in the study and are presented in the following tables. Table II and Table III present the confusion matrices of the VADER analyser and TextBlob analyser on Tamil tweets.

TABLE II. CONFUSION MATRIX OF VADER ANALYSER

	POSITIVE (Predicted)	***NEUTRAL (Predicted)***	***NEGATIVE (Predicted)***
POSITIVE (Actual)	4	45	2
NEUTRAL (Actual)	3	73	3
NEGATIVE (Actual)	12	156	2

TABLE III. CONFUSION MATRIX OF TEXTBLOB ANALYSER

	POSITIVE (Predicted)	***NEUTRAL (Predicted)***	***NEGATIVE (Predicted)***
POSITIVE (Actual)	1	49	1
NEUTRAL (Actual)	0	79	0
NEGATIVE (Actual)	0	170	0

The performance measures obtained for both lexical-based approaches are given in Table IV.

TABLE IV. PERFORMANCE MEASURE OF SENTIMENT ANALYSIS (FOR VADER AND TEXTBLOB ON TAMIL TWEETS).

ANALYSER	***Polarity***	***Precision***	***Recall***	***F1-Measure***	***Accuracy***
VADER	Positive	0.21	0.08	0.11	0.26
	Neutral	0.27	0.92	0.41	
	Negative	0.29	0.01	0.02	
TextBlob	Positive	1.00	0.02	0.04	0.27
	Neutral	0.27	1.00	0.42	
	Negative	0.00	0.00	0.00	

The following results were obtained when machine learning-based algorithms were used for the sentiment analysis of the Tamil tweets. The results are tabulated in Table V.

TABLE V. PERFORMANCE MEASURE OF SENTIMENT ANALYSIS (FOR MACHINE LEARNING-BASED APPROACHES ON TAMIL TWEETS).

ANALYSER	***Polarity***	***Precision***	***Recall***	***F1-Measure***	***Accuracy***
Support Vector Machine	Positive	0.78	0.83	0.81	0.75
	Neutral	0.61	0.85	0.71	
	Negative	1.00	0.42	0.59	
Random Forest	Positive	0.69	0.94	0.80	0.72
	Neutral	0.75	0.46	0.57	
	Negative	1.00	0.33	0.50	
eXtreme Gradient Boost	Positive	0.71	0.97	0.82	0.72
	Neutral	0.67	0.31	0.42	
	Negative	0.83	0.42	0.56	
Gaussian Naïve Bayes	Positive	0.81	0.71	0.76	0.68
	Neutral	0.62	0.77	0.69	
	Negative	0.46	0.50	0.48	

The results are presented diagrammatically for a comparative analysis in Fig. 1 below. Since the classification is multinomial, the weighted averages of the measures were used for this purpose.

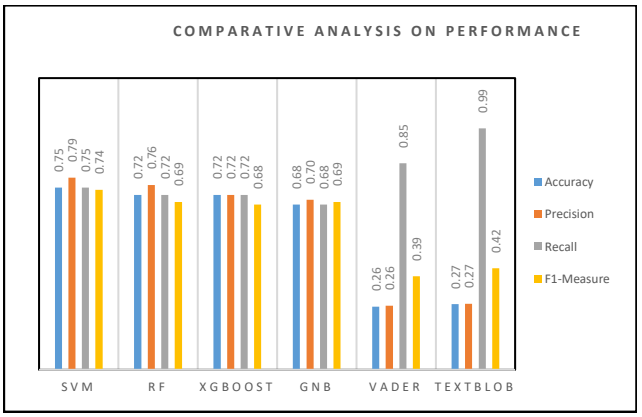

Fig. 1. Comparative Analysis on performance

V. DISCUSSION

The outcome of the study indicates that the machine-learning-based approach is still a good choice for performing sentiment analysis of Tamil tweets. Support Vector Machine performs outstandingly well when compared with eXtreme Gradient Boosting, Random Forest and Gaussian Naïve Bayes method with an accuracy of about 79%. The precision measure (79%), recall measure (75%) and F-measure (74%) of the Support Vector Machine are also comparatively better than the other algorithms. Although the performance of the lexical-based approach is inferior in this context, the performance of TextBlob seems to be comparatively better than its other lexical analyser, the VADER. Apparently, all the machine-learning approaches have obtained significantly greater accuracy, precision and F1-measure when compared with the lexical-based approach. However, the recall measures of the lexical analysers are still better than its counterpart in the study.

Despite the fact that the performances of these lexical analysers are low in this context, the validity of the tools

cannot be degraded since the tools have had proven results in many of the studies in the past [5,10]. Thus, this study reveals that the performance of the tools is much of context-specific. Furthermore, we believe this result might have been influenced by the fact that these tools rely on the corpora, and the corpora for the Tamil language are relatively limited [7].

However, we also agree that a machine-learning-based approach cannot always be considered an optimal algorithm for sentiment analysis owing to the difficulty in manually annotating the texts to their respective sentiments. This also consumes much time. Thus, we propose that consistent research must be undertaken to widen the corpus of the Tamil language as carried out by Thavareesan & Mahesan [11] as a future work of this study. The expansion of the corpus will have profound benefits in applying lexical analysing tools directly for movie reviews, customer reviews, and certain other aspects of sentiment analysis pertaining to the Tamil language.

VI. Conclusion

Sentiment Analysis in the Tamil language has many potential applications, including product marketing, product reviews, and movie reviews. The proliferation of social media has provided an efficient platform for the public to share their views publicly online. This enables the data scientists to dig deeper into their opinions and classify them. This study has utilised the Twitter platform to scrape the Tamil tweets and compared the performance of the sentiment classification using two approaches, namely machine learning-based and lexical-based. The Support Vector Machine algorithm of the machine learning approach has obtained a better performance score than other approaches. The performance of other machine learning-based approaches also has obtained a relatively better performance score. However, based on the challenges in both approaches, we conclude that this analysis is context-specific, and the expansion of the Tamil corpus would improve the efficiency of the tools as well.

References

[1] Rufai, S. R., & Bunce, C. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. Public Health |, 42(3), 510–516. https://doi.org/10.1093/pubmed/fdaa049

[2] Twitter, Inc. - Financial information - Quarterly results. (n.d.). Retrieved Feb. 24, 2022, from https://investor.twitterinc.com/financial-information/quarterly-results/default.aspx

[3] Alharbi, A. S. M., & de Doncker, E. (2019). Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research, 54, 50–61. https://doi.org/10.1016/J.COGSYS.2018.10.001

[4] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? 79–86. https://doi.org/10.3115/1118693.1118704

[5] W. Nur Syahirah Wan Min, N. Zareen Zulkarnain, and F. Teknologi Maklumat Dan Komunikasi, "Comparative Evaluation of Lexicons in Performing Sentiment Analysis," 2020. [Online]. Available: https://www.researchgate.net/publication/342451911

[6] E. Shihab and Y. Jing, "Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment," in Proceedings of the International MultiConference of Engineers and Computer Scientists 2019, Mar. 2019.

[7] S. Shriya, R. Vinayakumar, M. Anand Kumar, and K. P. Soman, "Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms," Indian J Sci Technol, vol. 9, no. 45, Dec. 2016, doi: 10.17485/ijst/2016/v9i45/106482.

[8] S. Anbukkarasi and S. Varadhaganapathy, "Analyzing Sentiment in Tamil Tweets using Deep Neural Network," in Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020, Mar. 2020, pp. 449–453. doi: 10.1109/ICCMC48092.2020.ICCMC-00084.

[9] M. A. Al-Shabi and M. A. Al-Shabi, "Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining Fuzzy Environment View project Accident Problem View project Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining." [Online]. Available: https://www.researchgate.net/publication/343473213

[10] V. Bonta, N. Kumaresh, and N. Janardhan, "A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," Asian Journal of Computer Science and Technology, vol. 8, no. S2, pp. 1–6, Mar. 2019, doi: 10.51983/ajcst-2019.8.s2.2037.

[11] S. Thavareesan and S. Mahesan, "Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts," in Moratuwa Engineering Research Conference (MERCon) 2020, 2020.

[12] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, no. 1, pp. 216–225, May 2014, Accessed: Jan. 08, 2022. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14550

[13] R. A. Laksono, K. R. Sungkono, R. Sarno, and C. S. Wahyuni, "Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes," Proceedings of 2019 International Conference on Information and Communication Technology and Systems, ICTS 2019, pp. 49–54, Jul. 2019, doi: 10.1109/ICTS.2019.8850982.

[14] M. R. Huq, A. Ali, and A. Rahman, "Sentiment Analysis on Twitter Data using KNN and SVM," 2017. [Online]. Available: www.ijacsa.thesai.org

[15] M. Ahmad, S. Aftab, and I. Ali, "Sentiment Analysis of Tweets using SVM," 2017.

[16] M. Ahmad, S. Aftab, M. Salman Bashir, N. Hameed, I. Ali, and Z. Nawaz, "SVM Optimization for Sentiment Analysis," 2018. [Online]. Available: www.ijacsa.thesai.org

[17] R. H. Hama Aziz and N. Dimililer, "SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier," https://doi.org/10.1080/02533839.2021.1933598, vol. 44, no. 6, pp. 562–572, 2021, doi: 10.1080/02533839.2021.1933598.

[18] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naive Bayes and K-NN Classifier," International Journal of Information Engineering and Electronic Business, vol. 8, no. 4, pp. 54–62, Oct. 2016, doi: 10.5815/ijieeb.2016.04.07.

[19] A. Goel, J. Gautam, and S. Kumar, "Real time sentiment analysis of tweets using Naive Bayes," Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016, pp. 257–261, Mar. 2017, doi: 10.1109/NGCT.2016.7877424

[20] M. A. Fauzi, "Random forest approach for sentiment analysis in Indonesian language," Indonesian Journal of Electrical Engineering and Computer Science, vol. 12, no. 1, pp. 46–50, Oct. 2018, doi: 10.11591/ijeecs.v12.i1.pp46-50.

[21] Y. Hegde and S. K. Padma, "Sentiment analysis using random forest ensemble for mobile product reviews in Kannada," Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017, pp. 777–782, Jul. 2017, doi: 10.1109/IACC.2017.0160.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.