Paper No: SC-21 **Smart Computing**

Sentiment classification of Sinhala content in social media

Pradeep Jayasuriya*
Faculty of Computing
Sri Lanka Institute of Information Technology, Sri Lanka
pradeep.jayasuriya@my.sliit.lk

Sarith Ekanayake
Faculty of Computing
Sri Lanka Institute of Information Technology, Sri Lanka
sarith.ekanayake@my.sliit.lk

Ranjiva Munasinghe
Faculty of Business
Sri Lanka Institute of Information Technology, Sri Lanka
ranjiva.m@sliit.lk

Bihara Kumarasinghe
Faculty of Computing
Sri Lanka Institute of Information Technology, *Sri Lanka*
bihara.kumarasinghe@my.sliit.lk

Isuru Weerasinghe
Faculty of Computing
Sri Lanka Institute of Information Technology, Sri Lanka
don.isuru@my.sliit.lk

Samantha Thelijjagoda
Faculty of Business
Sri Lanka Institute of Information Technology, Sri Lanka
samantha.t@sliit.lk

Abstract: **In this study, we focus on the classification of Sinhala social media sentiments into positive and negative classes for a particular domain (sports). We have employed machine learning algorithms and lexicon-based sentiment classification methods. We also consider a hybrid approach by constructing an ensemble classifier in which we combine Machine Learning and Lexicon based methods. For individual methods, machine learning algorithms performed best in terms of accuracy. The ensemble classifier was able to improve performance further.**

Keywords: *Machine Learning, Natural Language Processing, Sentiment Analysis, Social Media, Supervised Learning*

I. INTRODUCTION

Social Media has a major impact on the world today with global usage in 2018 estimated to be 2.65 billion. Social media has become the major platform where people share their opinions on various topics such as products, services, people, places, organizations, events, news, ideas etc. Many insights can be gained from understanding what is being said on social media – e.g. from a business perspective, social media is a great source for understanding where their products or services are positioned among the customers. Accordingly, social media sentiment analyzing researches have been conducted [1], [2], [3], [4] and tools have been developed for popular languages such as English (e.g. Social Studio, Hootsuite etc.) which can provide insights for businesses to improve their products and business processes. Social media monitoring is also important for monitoring social unrest [5].

In the local context, Sri Lanka has over 6 million social media users which represent close to 30% penetration. In particular, social media users expressing their opinions in the Sinhala language has also increased significantly. At present, there are some research efforts at the state universities to use Natural Language Processing (NLP) on the Sinhala language, in particular focusing on morphological analysis on traditional texts and machine translations e.g.: Multilingual document generation. To the best of our knowledge, the work done on analyzing Sinhala content in social media is limited[1], particularly in polarity classification of sentiments i.e. classifying sentiments into positive negative and neutral classes. This study has classified Sinhala sentiments using a dictionary-based lexicon method [6].

Sentiment analysis is an area of study within NLP for extracting sentiments from text via automated techniques. Opinion mining and sentiment analysis are well-established in linguistic resource-rich languages (e.g. English). The success of an opinion mining approach depends on the availability of resources, such as special lexicons, coding libraries, and WordNet type tools for a particular language. Due to the lack of such resources, it is more difficult to analyze the sentiments of languages that are less commonly used like Sinhala [7]. Another challenge for NLP analysis for Sinhala (and other Indic languages) is due to their diglossic nature - whereby there are both formal and informal dialects of the same language which are very different. It is the informal language that is more frequently used in Sinhala content on social media. The domain is also important, as algorithms that are trained for one particular domain provide poor results in a different domain. Other challenges include the use of English loan words (written using the letters from the Modern Latin alphabet) and the use of 'Singlish' – where Sinhala words are spelled out phonetically in English. A more complete list of challenges in Indic languages can be found in [5].

There is two main sentiment analysis approaches Machine Learning (ML) based [8] and lexicon-based

[1] There are a considerable number of studies on Hate Speech and Racism detection for Sinhala on Social media

sentiment analysis. In this study, we have implemented both these methods. Lexicon-based sentiment analysis is generally not used for social media sentiment analysis due to the use of informal language in social media. We have proposed a new approach to lexicon library construction as a solution to this matter. We also employ an ensemble approach [9] to improve the predictive performance of the sentiment analysis, in which we combine both machine learning and lexicon-based methods as a hybrid sentiment analysis approach. Sentiment classification in all the methods we have employed in this paper are supervised approaches [10]. In particular, we work with a binary classification of sentiments into positive and negative (polarity) classes. YouTube is selected as the social media platform and 'sports' is the selected domain of this study. We have focused on comment-level sentiment classification where a comment contains one or several sentences that are considered a single entity by the sentiment analysis process.

The paper is structured in the following manner – we begin with the methodology section where we discuss the construction of the sentiment analysis model. In particular, we describe the dataset, data pre-processing, feature extraction and different analysis approaches taken. The next section discusses our results and findings. The paper ends with a summary and discussion of the current study and our planned future work.

II. METHODOLOGY

This section describes the construction of the sentiment analysis model for analyzing Sinhala social media content. It involves data tokenization, pre-processing, feature extraction and sentiment analysis. Python is used as the language for the development of this model.

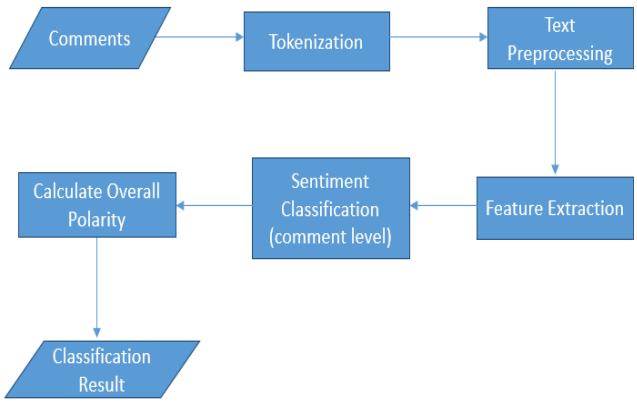

Fig. 1. Sentiment analysis flow chart.

A. *Data-set description*

Comments were obtained from sports-related videos (involving cricket, rugby, and athletics) from YouTube. The next step was to label these comments into classes (positive or negative) by sentiment, resulting in a dataset suitable for supervised learning. When creating the dataset, longer comments (comments with more than five sentences) were manually split in a way that a split contains a complete and independent sentiment. We also ensured the dataset allowed for stratified sampling. A total of 2210 comments were grouped as follows for training and testing purposes.

TABLE I. DATASET DESCRIPTION

	Train set	Test set	Total
Positive comments	830	275	1105
Negative comments	830	275	1105
Total	1660	550	2210

Source: [10]

The dataset consists of 2810 total sentences and 1346 of them are distributed in the 1105 positive comments and the remaining 1464 sentences are distributed in the 1105 negative comments. There is a total of 21,573 words in the dataset. They are distributed as 8389 words in positive comments and 13,184 words in negative comments.

B. *Data pre-processing*

The first step in this stage is tokenization. Text is first tokenized as comments and then into words for carrying out further analysis. Text cleaning follows tokenization, where only the main Sinhala characters were considered and all non-Sinhala characters, punctuation and numerical text were removed from comments.

After the initial cleaning, stop words were removed from the text. Stop word removal is an important task in sentiment analysis and was first introduced by Hans Luhn [11]. Stop words are common words with a high term frequency in a document that does not have any sentimental value. There are different methods available for stop-word removal [12], and the stop word removal greatly enhances the performance of the feature extraction algorithm [1, 13]. Removing stop words also reduces the dimensionality of the data sets. It will leave key opinion words which will make the sentiment analyzing process more accurate. Stop words are taken from a customized list of stop words for the particular domain. At the simplest level stop, words are iterated in a word list and removed from the text.

C. *Feature extraction*

Feature extraction is a very important task in sentiment analysis because the accuracy of the analysis depends on it. We have used the word N-gram method for feature exaction - in particular unigram, bigram and trigram features [14]. N-gram feature vectors are created as a bag of n-grams representations from tokenized and pre-processed text.

D. *Machine learning-based sentiment analysis*

In our approach we start by using different ML algorithms for our classification: Naïve Bayes classifier (NB), Logistic Regression Classifier (LR) and support vector machine classifier (SVM) [13]. Each algorithm is trained using unigram, bigram and trigram feature extraction methods to train 3 classifiers for each algorithm – resulting in 9 classification methods. In a subsequent section, we describe creating an ensemble classifier based on combinations of the above base learners and lexicon-based classifier.

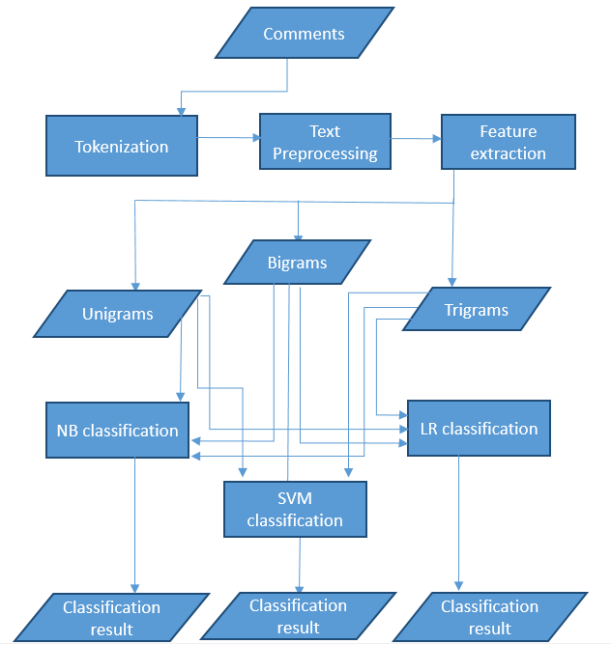

Fig.2..ML-based sentiment analysis: Base learner training

E. Lexicon-based sentiment analysis

We describe the lexicon-based method in a little more detail as we have combined existing approaches in a novel way. This section starts with 1) lexicon library construction which consists of two subsections: 1.1) subjective lexicon construction and 1.2) expanding subjective lexicon. Expanding subjective lexicon involves 1.2.1) string similarity analysis and 1.2.2) mismatches removal strategy. Finlay the 2) lexicon-based sentiment classification is discussed.

For the lexicon-based analysis, we employ a Bayesian analysis method [15] and improve upon this by using the Levenshtein ratio analysis for string similarity analysis [16].

1) *Lexicon library construction*

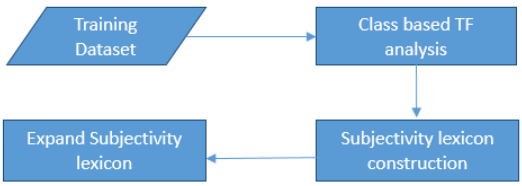

Fig.3. Lexicon library construction

In our approach of lexicon library construction, we have constructed a subjectivity lexicon [17] using the training data set and expanded it by using a Sinhala-English dictionary.

a) Subjective lexicon construction

In the literature, a list of words with the polarities assigned to each word in the list is called a subjective lexicon [18]. We calculate a positive and negative probability score for each word i.e. unigram feature in the training data set based on its term frequency (TF) in each class. The difference in each class for each word is also calculated to construct a difference score. These scores are then assigned as sentiment scores for a particular word. From the training dataset, we obtained a subjective lexicon consist of 4389 words at the end of this analysis.

$$\text{Positive polarity score} = \frac{\text{Frequency of positive}}{\text{Total positive comments}} \quad (1)$$

$$\text{Negative polarity score} = \frac{\text{Frequency of negative}}{\text{Total negative comments}} \quad (2)$$

$$Difference = Positive\ polarity\ score - Negative\ Polarity\ Score \quad (3)$$

As the Difference score of a word increases, it has a higher sentiment value. The difference score is used for sorting the subjective lexicon so that the sentiment analyzing process can identify its significance.
Bigram and trigram features will not be considered for lexicon-based analysis.

b) Expanding subjective lexicon using a Sinhala-English dictionary

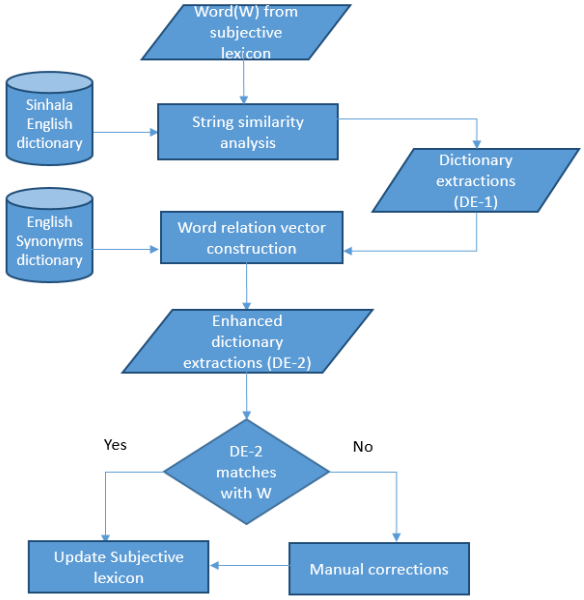

Fig. 4. Expanding subjective lexicon

Subjective lexicon is expanded using string similarity analysis upon a Sinhala English dictionary and by employing an additional mismatches removal strategy.

I. String similarity analysis

Each word in the constructed subjective lexicon is checked against a Sinhala-English dictionary for String similarity analysis. The Sinhala-English dictionary has a structure as each English word has one or more Sinhala meanings. The purpose of string similarity analysis is to extract synonyms and similar words (extracting different words of the same lemma) for words in the subjective lexicon from the dictionary. We have employed Levenshtein ratio analysis [16] for the purpose of string similarity analysis. Levenshtein ratio analysis provides matched entries from the

dictionary for a particular word in the subjective lexicon as a result.

e.g.: A word from the subjective lexicon: ‘ආදරෙයි‘

Dictionary extractions from the Levenshtein ratio analysis for ‘ආදරෙයි‘:

- 'affection': 'ආවේදනාව', 'චින්දනය', 'අනුරාගය', 'ආදරය', 'ආබාධය', 'දයාව', 'රෝගය', 'ස්නේහය'
- 'love': 'ආදරය', 'ආදරය කරනවා', 'ආලය', 'ආලය කරන වස්තුව','පෙම්වතිය', 'මෛත්රීය', 'ස්නේහය', 'සෙනෙහස'
- 'respect': 'ගෞරවය', 'ආකාරය', 'ආදරය', 'උපහාරය', 'ගරු කරනවා', 'විෂයය', 'සම්මානය', 'සැලකිලි දක්වනවා'

Underlined words are the detected words from Levenshtein ratio analysis. Sinhala synonyms from each dictionary extraction are collected to update the subjective lexicon. In this step, it is important to exclude duplicates that are already present in the subjective lexicon.

II. Mismatches removal strategy

Results provided from the Levenshtein ratio analysis sometimes can be inaccurate because of the following reasons.

a. Collected synonyms from dictionary extractions may include irrelevant words due to language differences [6].

e.g.: Word from the subjective lexicon: ඉක්මන් (Means ‘Quick’ in Sinhala)

Dictionary extractions from String similarity analysis:

- කඩිසර, ශීඝ්‍ර(Means ‘Quick’ in Sinhala)
- උපවාසය, උපවාසයෙහියෙදෙනවා (Means a Hunger strike in Sinhala)

These errors will be corrected manually at the end of this analysis.

b. String similarity analysis may detect irrelevant Sinhala words. Even if the two words are unrelated, Similarity in the structure (Letters and their order which constructs the word) of the two irrelevant words is the reason for this issue.

In order to overcome b, English words of the dictionary word extractions ('affection', 'love', 'respect' as for the above example) are leveraged. For a particular Levenshtein ratio analysis result of a word in the subjective lexicon, resulting dictionary extractions’ English words are checked for English synonyms from an online English synonyms dictionary using python web automation. Each resulting English synonym set of each English word from the dictionary extractions is then analyzed by checking for common synonyms between them to construct word relation vectors. In this regard 1) If word A and B are having a common synonym and 2) B and C is having a common synonym A and C is also considered as having a common synonym for indicating their relation. As for this example, words A, B and C will be included in the same word relation vector. These word relation vectors indicate the relations between the dictionary extractions. From the constructed word relation vectors, we considered the vector with the highest number of words as the mismatches removed vector. (Sometimes this hypothesis may cause inaccurate dictionary extractions. They will be manually corrected at the end of this analysis)

Finally, the Sinhala words are picked from the selected vector’s dictionary extractions and they are assigned the same polarity scores of the Analyzed Sinhala word in the subjective lexicon.

e.g.: As for the above example since ‘affection’, ‘love’ and ‘respect’ are words which have English synonyms in common, all corresponding Sinhala words are selected and subjective lexicon is updated as follows.

```
ආදරෙයි   0.308   5.218   0.12   |ආවේදනාව|චින්දනය|අනුරාගය|ආදරය|
```

Fig. 5. Updating subjective lexicon

This analysis provided Sinhala synonym sets for 789 words in the subjective lexicon and 138 manual corrections were applied for them.

2) *Sentiment classification using lexicon library*

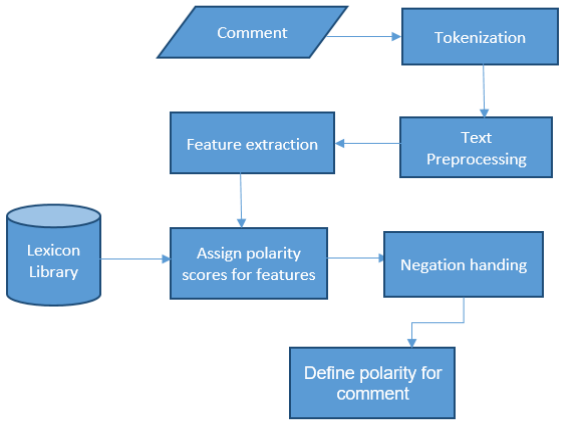

Fig. 6. Lexicon based sentiment analysis

Words in a comment are traversed through the lexicon library and if a particular word is present in the library it’s positive and negative sentiment scores will be collected to construct a positive and a total negative score for the comment. These total positive and negative scores are then compared and the score with the higher absolute value is selected as the final sentiment score. If the higher value is positive, the comment is classified as positive and if the higher value is negative, the comment is classified as negative.

Negation handling is another important task which increases the accuracy of lexicon-based sentiment analysis. Negation shifters (Fig.7) change the polarity of the features around them. A negation shifter is associated with a feature word within its context distance and the polarity of the feature word is shifted by the negation shifter. We have considered context distance as the first 2 words before the negation shifter. A negation shifter will multiply the value of any feature within the context distance of the negation shifter by a factor of -2 in our analysis process as negation rules used for the Sinhala language in [18].

නෑ නෑ, නැහැ, නැත, නැති (all means no or not)
බෑ, බෑ, බැහැ, බැරි, බැරිය (all means can't)
එපා (don't)

Fig. 7. Selected negation shifters used for lexicon-based sentiment analysis

F. Ensemble classifier

We combine ML and lexicon-based methods using a majority voting ensemble classifier. A comment is classified by each base learner. The list of base learners are as follows:

- 4 Unigram classifiers :

  1) Lexicon-based classifier, 2) NB unigram classifier, 3) LR unigram classifier, 4) SVM unigram classifier
- 3 Bigram classifiers :

  5) NB bigram classifier, 6) LR bigram classifier, 7) SVM bigram classifier
- 3 Trigram classifiers :

  8) NB trigram classifier, 9) LR trigram classifier, 10) SVM trigram classifier

The final classification is determined by the majority vote of the base learners incorporated in the ensemble. We compare different ensembles by considering different combinations of the base learners.

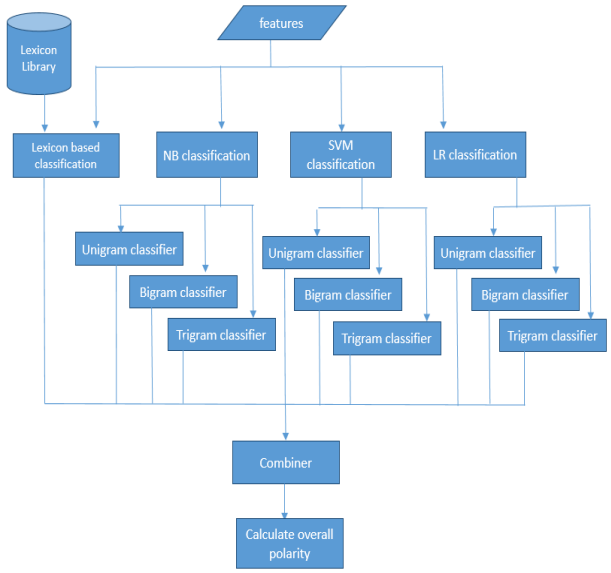

Fig. 8. Combining base learners using an Ensemble classifier.

III. RESULTS

We present our results which are the percentages of sentiments in the testing set that were predicted correctly by the model. F1-score and Accuracy are used as evaluation metrics. Both of these metrics range between 0 and 1, with higher values indicating better classification/prediction.

TABLE IIII. INDIVIDUAL ANALYSIS METHODS COMPARISON

Classifier		With stop word removal		Without stop word removal	
		F1 Score	Accuracy-Score	F1 Score	Accuracy-Score
Naïve Bayes	Unigram	0.73	0.78	0.71	0.77
	Bigram	0.66	0.61	0.64	0.54
	Trigram	0.61	0.53	0.60	0.50
	Unigram	0.74	0.73	0.74	0.73
SVM	Bigram	0.66	0.56	0.43	0.55
	Trigram	0.57	0.50	0.56	0.47
Logistic Regression	Unigram	0.77	0.78	0.76	0.76
	Bigram	0.67	0.60	0.43	0.59
	Trigram	0.62	0.54	0.61	0.51
Lexicon based analysis	Unigram	0.70	0.72	0.64	0.68

The results indicate that removing stop words improves prediction metrics. Bigram classifiers are the most positively affected classifiers from stop word removal.

For the majority voting ensemble classifier we found that even though the stop word removal improved the base learner prediction metrics, it reduced the corresponding metrics of the ensemble classifiers by 1-2%. The majority voting ensemble classifier performed the best (including improving upon the base learner performance) without stop word removal.

TABLE III. MAJORITY VOTING METHODS COMPARISON

Majority voting Ensemble classifier (without stop words removal)	NO. of base learners	F1 Score	Accuracy-Score
1. Unigram classifiers only	3	0.762	0.775
2. Unigram and bigram classifiers only. (LR bigram classifier excluded)	5	0.781	0.797
3. Unigram and bigram classifiers with Lexicon based classifier	7	0.786	0.800
4. Unigram, Bigram and Trigram classifiers only	9	0.768	0.756
5. one of three trigram classifiers excluded from 10 base learners	9	0.787	0.790

The results indicate the best f1 score and accuracy-score combinations were given by the following two combinations of base learners: 1) Unigram and bigram classifiers with lexicon-based classifier 2) Unigram, bigram and trigram classifiers with lexicon-based classifier.

IV. CONCLUSIONS & FUTURE DIRECTIONS

In this research, we focused on analyzing Sinhala sentiment by considering comments on sports-related videos on the YouTube social media platform. We employed both ML algorithms and lexicon-based methods for sentiment analysis. In addition, we considered a hybrid approach by combining the above methods using a majority vote ensemble classifier.

In terms of feature extraction, the unigram method was the most effective. N-grams (N>1) tend to improve language coverage [19] and performance when the corpus is larger [19]. Based on our results Logistic Regression unigram classifier with stop-word removal was the most accurate among individual ML algorithms according to the F1-score and accuracy metrics.

For the lexicon-based methods, the best performance was obtained by the use of Levenshtein ratio analysis with the

Sinhala – English dictionary to expand the subjective lexicon. This method also incorporated stop-word removal.

When considering individual methods, ML methods were more accurate than the lexicon-based methods.

For the majority vote ensemble classifier, we found that not removing stop words provides better performance. A possible explanation for this is that we are creating a strong classifier from an ensemble of weak base learners, whereas an ensemble of strong base learners may not show improvement over the strongest base learner. We also found that hybrid sentiment classification (the combination of ML and lexicon) was more accurate than ML or lexicon-based methods alone. The lexicon-based approach helps to improve the diversity of the base learners used for the ensemble classifier which helps for employing further ensemble techniques.

For future work, we hope to explore the following paths of research:

- Using other base learner algorithms (e.g. kNN Clustering, Random forest).
- Using further ensemble learning methods such as boosting (AdaBoost, Gradient boosting, extreme gradient boosting and light gradient boosting) and stacking methods.
- Employing character N-grams as features.
- Developing a combination of rule-based methods (e.g. POS tagging) and ML algorithms using linguistic features of the Sinhala language.
- Use of Deep Learning and other Neural Network-based methods.

APPENDIX

In each step of the lexicon construction, performance improvements were recorded as follows (with stop word removal).

TABLE IV. LEXICON BASED ANALYSIS

Steps of lexicon construction	F1 Score	Accuracy
1. Initial subjective lexicon (without expanding using Sinhala English dictionary)	0.60	0.62
2. Expanding subjective lexicon: without manual corrections	0.67	0.71
3. Expanding subjective lexicon : with manual corrections	0.70	0.72

ACKNOWLEDGMENT

We thank the Faculty of Graduate Research Studies - Sri Lankan Institute of Information Technology for their partial support for this research

REFERENCES

[1] M. Hu and B. Liu, "Mining Opinion Features in Customer Reviews," in *Proceedings of the 19th National Conference on Artifical Intelligence*, San Jose, California, 2004.

[2] F. Neri, C. Aliprandi and M. Cuadros, "Sentiment Analysis on Social Media," in *International Conference on Advances in Social Networks Analysis and Mining*, 2012/08/28.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonnea, "Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Language in Social Media ({LSM} 2011)*, Portland, Oregon, Association for Computational Linguistics, 2011-June, pp. 30-38.

[4] S. Jayasanka, T. Madhushani, E. Marcus, . I. Aberathne and S. Premaratne, "Sentiment Analysis for Social Media," in *Information Technology Research Symposium*, 2013/11/22.

[5] P. Bhattacharyya, H. Murthy, R. Munasinghe and S. Ranathunga, "Indic language computing," in *Communications of the ACM*, vol. Vol. 62 no. 11, November 2019, pp. 70-75.

[6] N. Medagoda, S. Shanmuganathan and J. Whalley, "Sentiment Lexicon Construction Using SentiWordNet 3.0," in *2015 11th International Conference on Natural Computation (ICNC)*, Auckland, 2015.

[7] N. de Silva, "Survey on Publicly Available Sinhala Natural Language Processing Tools and Research," Cornell University, 05 06 2019. [Online]. Available: https://arxiv.org/abs/1906.02358. [Accessed 03 07 2019].

[8] K. Al-Barznji and A. Atanassov, "BIG DATA SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHMS," in *26 th International Symposium "Control of Energy, Industrial and Ecological Systems"*, Bankia , Bulgaria, 2018.

[9] Y. Yang, "Chapter 4 - Ensemble Learning," in *Temporal Data Mining Via Unsupervised Ensemble Learning*, Elsevier, 2017, pp. 35-56.

[10] J. E. T. Akinsola, "Supervised Machine Learning Algorithms: Classification and Comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, pp. 128 -138, 08 06 2017.

[11] H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159-169, 1958.

[12] J. Kaur and P. K. Buttar, "STOPWORDS REMOVAL AND ITS ALGORITHMS BASED ON DIFFERENT METHODS," *International Journal of Advanced Research in Computer Science*, Vols. Vol 9, No 5, pp. 81-88, 20 10 2018.

[13] D. Ly, K. Sugiyama, Z. Lin and M.-Y. Kan, "Product Review Summarization from a Deeper Perspective," in *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL*, Ottawa, ON, Canada, 2011/01/01, pp. 311-314.

[14] A. Esuli and . F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining," in *5th Conference on Language Resources and Evaluation*, Genoa, Italy, 2006 - may.

[15] T. Hastie, R. Tibshirani and J. Friedman, "8.3 Bayesian Methods," in *The Elements of Statistical learning*, 2 ed.

[16] S. Zhang, Y. Hu and G. Bian, "Research on string similarity algorithm based on Levenshtein Distance," in *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 25-26 March 2017.

[17] S. Park and . K. Yanggon , "Building thesaurus lexicon using dictionary-based approach for sentiment classification," in *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, 2016.

[18] N. Medagoda, "Framework for Sentiment Classification for Morphologically Rich Languages: A Case Study for Sinhala," Auckland University of Technology, Auckland, 2017.

[19] A. Vlasblom, "Coursera Data Science Capstone Milestone Report," July 2015.