SENTIMENT CLASSIFICATION OF SINHALA CONTENT IN SOCIAL MEDIA: A COMPARISON BETWEEN WORD N-GRAMS AND CHARACTER N-GRAMS

Pradeep Jayasuriya
SLIIT Business School
Sri Lanka Institute of Information Technology
Malabe, Sri Lanka
pradeep.jayasuriya@my.sliit.lk

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

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

Abstract: **In this study, we focus on the classification of Sinhala posts on social media into positive and negative class sentiments. We focus on the domain of sports. We employ machine learning algorithms for sentiment classification where we compare feature extraction methods using Character N-grams (for N ranging from 3 to 7) and Word N-grams (for N ranging from 1 to 3). We find that Character N-grams outperform Word N-grams in sentiment classification. Further, we find that a) lower level character N-grams (N = 3 or 4) outperform higher level character N-grams (N ranging from 5 to 7) and b) the combinations of N-grams of different orders outperforms individual N-gram results (N: 1, 2 for words and N: 3, 5 for characters). In addition, Character N-grams enable the sentiment classifier to a) detect spelling mistakes and b) function as a stemmer which results in higher sentiment analysis accuracy.**

Keywords: Sentiment Analysis, Natural Language Processing, Sinhala, Social Media, N-grams, Machine Learning

I. INTROODUCTION

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 Sri Lanka, there are over 6 million social media users – i.e. a penetration of approximately 30%. In particular, social media users expressing their opinions in the Sinhala language has also increased significantly.

There is a considerable amount of research effort on Sinhala Natural Language Processing (NLP), however, to the best of our knowledge, the work done on analyzing Sinhala content in social media is limited [1]. In particular, polarity classification[2] of sentiments in Sinhala social media content is not well-researched.

Sentiment analysis is an area of study within NLP for extracting sentiments from text via automated techniques. Opinion mining and sentiment analysis is well-established in linguistic resource-rich languages such as 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 the 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 [6]. Other challenges for Sinhala NLP analysis include a) the fact that Sinhala is a morphologically rich language and b) Sinhala is diglossic, whereby the formal and informal dialects are very different. It is the informal

[1] There are a considerable number of studies on Hate Speech
[2] Classifying sentiments into positive, negative and possibly neutral classes.

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 code-mixed text (use of English words in Sinhala sentences) 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].

In this study we use machine learning algorithms for sentiment classification of social media comments using character level N-grams (char N-grams) and word level N-grams feature extraction. In particular, we work with a binary classification of sentiments into positive and negative (polarity) classes and assess the performance of the respective methods. We have employed supervised classification [7] for this study. 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 which contains one or several sentences and is then considered a single entity by the sentiment analysis process.

The rest of our paper is structured in the following manner – we begin with a brief introduction to N-grams, followed by a short discussion on the use of N-grams. The next section is the methodology where we discuss the sentiment analysis model. In particular, we describe the dataset, data pre-processing, feature extraction and different approaches taken in the sentiment analysis. 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. N-GRAMS & RELATED WORK

Given a sentence S, Word N-grams of S are a sequence of N word combinations, made out of all possible combinations of adjacent words of length N.

Ex:

'He is the best player of our generation'

- Unigrams (N=1):
  [He, is, the, best, player, of, our, generation]
- Bigram (N=2):
  [He is, is the, the best, best player, player of, of our, our generation]

Similarly, given a sentence S, Character N-grams of S are a sequence of N character combinations, made out of all possible combinations of adjacent characters of length N.

Ex:

Ex: 'best match ever'

- Character Trigrams (N=3):
  ['bes', 'est', 'st ', 't m', ' ma', 'mat', 'atc', 't m', 'ch ', 'h e', ' ma', 'est', 'ver', 'est']

The use of character N-grams in place of words has been used for various NLP tasks – for example:

- Text categorization [8]
- Numerical classification of multilingual documents and information retrieval [9]
- Author identification [10]
- Language detection [11].

In the study [8] of text categorization, newspaper articles from English, Japanese and Chinese newspapers are classified using the 'FRAM' (Frequency Ratio Accumulation Method). It is a new proposed classification technique that adds up the ratios of term frequency among categories. Adopting character N-grams as feature terms has improved the accuracy of these experiments.

In the study [12], it has been demonstrated that Character N-grams perform better than Word N-grams for text classification. They have used the IMDB movie review data set (English) [13] in this study. Using Character N-grams as feature terms improves the FRAM.

III. METHODOLOGY

This section describes 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 development of this model.

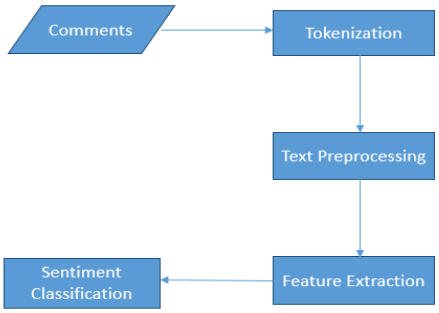

Fig. 5. Sentiment analysis flow chart.

A. *Data-set description*

Sinhala comments were obtained from sports-related videos (cricket, rugby and athletics) from YouTube. The next step was to label these comments into sentiment classes (positive or negative) to create 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 an 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.

1) DATASET DESCRIPTION

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

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 text cleaning where only the main Sinhala characters are considered. All non-Sinhala

characters, numerical text, and punctuation (except the full stop) were removed from the comments.
After the initial cleaning, comments are tokenized by separating strings by white spaces. These tokens are further processed using two steps: a) Sentence separation correction and b) stop word removal.
Sentence separation is important for the tokenization accuracy because comment level classification is employed. Social media comments may include the full stop at the end of a sentence, but as for the following example the second comment may not be separated properly because of the missing white space:

1). තරඟ 3 දින්නා. සුබ පතනවා (Properly separated)
2). තරඟ 3 දින්නා.සුබ පතනවා (Improperly separated)

This creates a single token as 'දින්නා.සුබ' which includes 2 different words in Sinhala Language. These tokens will be corrected by removing the full stop and dividing in to 2 new tokens. As for the above example the token 'දින්නා.සුබ' will be separated into two tokens as 'දින්නා' and 'සුබ'.
Stop words were removed from the text by removing corresponding tokens. Stop word removal is an important task in sentiment analysis and was first introduced by Hans Luhn [14]. Stop words are common words with a high term frequency in a document that does not have any sentiment value. There are different methods available for stop word removal [15], and in doing so greatly enhances the performance of the feature extraction algorithm [1, 16]. 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

In the feature extraction, comments are tokenized into N-grams for carrying out further analysis where the bag of word words representation is used to represent features in a comment. N-grams tend to improve both language coverage and classification performance when the corpus is larger [17]. Character N-gram features are less sparse than word N-grams features, and are expected to have a performance overhead compared to the processing time of the word N-grams.
Character N-grams are used in tools for spelling mistakes [18] and stemmers [19]; thus its use in Character N-gram feature extraction allows the corresponding classifier to function as both a stemmer and tool for correcting spelling mistakes. Mis-spellings and noise (caused by wordplay and creative spelling) tends to have a minimum impact on substring patterns (substrings of words) than word patterns when analyzed by machine learning algorithms.
In Sinhala script, characters can be consonants, vowels or diacritics. Sinhala diacritics are called 'Pilli' (vowel strokes). A Sinhala letter in Sinhala script can be a consonant, vowel or a compound form of a consonant and a vowel stroke. In contrast, a Sinhala letter is formed by character unigram or a character bigram in Sinhala script.

FORMATION OF SINHALA LETTERS

Formation (Consonant + Vowel)	Pilla (Vowel Stroke)	Compound Form
ක් + ඈ	◌ෑ	කෑ
ක් + ඒ	◌ේ	කේ
ක් + උ + ර්	◌්‍ර	ක්‍ර

The following N-grams / N-gram combinations were considered in this study.

1. Word N-grams
   - Unigrams
   - Bigrams
   - Trigrams
   - Unigrams + Bigrams
   - Unigrams + Bigrams + Trigrams

2. Character N-grams:
   - Individual char N-grams
     - 2/3/4/5/6/7 characters
   - Char N-gram combinations
     - (2,3),(2,3,4),(2,3,4,5),(2,3,4,5,6), (2,3,4,5,6,7)
     - (3,4) ,(3,4,5) ,(3,4,5,6), (3,4,5,6,7)
     - (4,5) ,(4,5,6), (4,5,6,7)

Space character is an important aspect of character N-gram tokenization. It gives awareness about word boundaries. The N-grams described above were further tested in 2 different tokenizing methods as follows:
1) **With adjacent word awareness in N-gram tokens:** In this method, a complete sentence is considered as one string for generating N-grams. N-grams are generated from inside and outside of word boundaries (beginning and the end of a word which are marked with an underscore). This method provides awareness about the adjacent words by considering N-grams shared by two adjacent words.
2) **Without adjacent word awareness in N-gram tokens:** Words of a sentence are considered as separate entities for generating N-grams in this method. N-grams are generated only inside word boundaries. N-grams do not include any information about adjacent words in the tokens.
E.g.: Character N-grams (N=4) of phrase'අපේම්_කට්ටිය' (Space character is replaced with an underscore)

- Without considering space:
  ['අ ප ේම්' , 'ක ට ්ට' , 'ට ්ට ි' , '්ට ි ය ']

- Considering space:
  ['අ ප ේම්' , 'ප ේම් _' , '**ේම් _ ක**' , '**ම් _ ක ට**' , '_ ක ට ්' , 'ක ට ්ට' , 'ට ්ට ි' , '්ට ි ය']

The highlighted N-grams include a space in the middle of the N-gram indicating the end of one word and beginning of the adjacent word.

D. Machine Learning-Based Sentiment Analysis

We have employed several machine learning algorithms from the Python Scikit-learn library to test the performance of the classification model:

A) Naïve Bayes Classifiers:
   1. Bernoulli Naïve Bayes
   2. Complement Naïve Bayes
   3. Multinomial Naïve Bayes

B) Support vector machine Classifiers:
   4. SVC
   5. Linear SVC
   6. NuSVC

C) Boosting Classifiers:
   7. Ada-Boost Classifier
   8. Xg-Boost classifier(XGB)
   9. Gradient Boost Classifier(GBM)

D) Other Classifiers:
   10. Logistic Regression Classifier
   11. Decision Tree Classifier
   12. Random Forest Classifier(RF)
   13. K-Nearest Neighbors classifier(KNN)

IV. RESULTS

WE use the F1-score, Accuracy and Kappa as the metrics to evaluate the classifier performance. The F1-score and Accuracy metrics range between 0 and 1, with higher values indicating better classification/prediction. Kappa measures the improvement above a random classifier and is theoretically bound above by 1 with higher scores indicating better classification/prediction. A kappa of zero would indicate the classifier is as good as random guessing. It can take negative values as well. We use 6-fold cross-validation to evaluate the classifier performance.

Table III and IV presents a comparison of N-values for word N-grams and char N-grams respectively with logistic regression. Character N-grams were more accurate than word N-grams but processing times were much lower for the word N-grams.

WORD N-GRAMS COMPARISON

N (N-gram/ N-gram combination)	Processing Time(ms)	F1 Score	Accuracy	Kappa
N=1	36	0.74	74.35	0.487
N=2	80	0.69	63.62	0.272
N=3	105	0.67	53.40	0.068
N:1,2	141	0.75	75.02	0.500
N:1,2,3	131	0.74	74.92	0.498

CHAR N-GRAMS COMPARISON

Character N-gram	Processing Time(ms)	F1 Score	Accuracy	Kappa
N=2	150	0.77	77.08	0.543
N=3	180	0.79	79.77	0.595
N=4	198	0.80	80.65	0.613
N=5	210	0.79	79.10	0.582
N=6	146	0.77	77.86	0.557
N=7	122	0.78	78.27	0.565

We also present the results of comparison between 1) generating N-grams only inside word boundary and 2) generating N-grams both inside and outside of word boundary in the following table. It demonstrates the effect of awareness of adjacent words in char N-gram tokens. Logistic Regression is the classification algorithm used in this comparison. Best results were obtained by method 1)

EFFECT OF ADJACENT WORD AWARENESSS IN CHAR N-GRAM TOKENS

Char N-gram	Without Adjacent Word Awareness in Tokens			With Adjacent Word Awareness in Tokens		
	F1 Score	*Accuracy*	*Kappa*	*F1 Score*	*Accuracy*	*Kappa*
N=2	0.77	77.08	0.54	0.76	76.31	0.52
N=3	0.79	79.77	0.59	0.79	79.27	0.59
N=4	0.80	80.65	0.61	0.80	80.05	0.60
N=5	0.79	79.10	0.58	0.79	77.86	0.55
N=6	0.77	77.86	0.55	0.77	75.23	0.50

Combinations of character N-grams produced the best results of this study. Multinomial Naïve Bayes, Complement Naïve Bayes and Logistic Regression provided the best results (above 80%) among the 13 algorithms tested in this experiment. The following graphs of N-gram combinations starts with a particular value of N and next value of N is added to feature extraction to measure the change of the accuracy-score and compare the N-gram combinations.

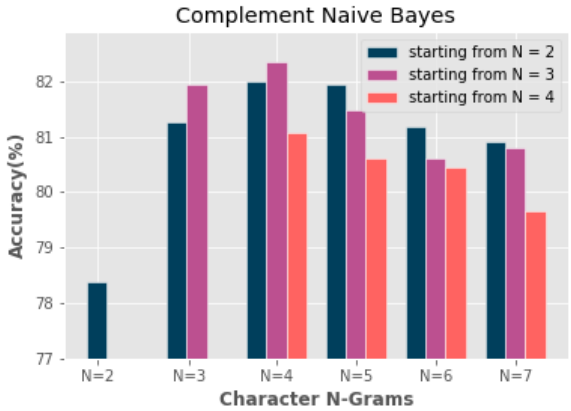

Fig. 6. Complement Naïve Bayes char N-gram results

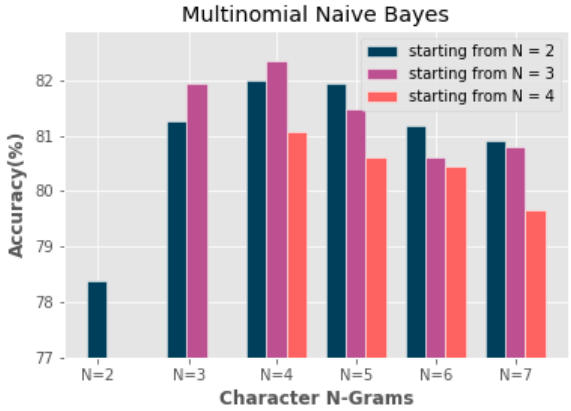

Fig. 7. Multinomial Naïve Bayes char N-gram results

Fig. 8. Logistic Regression char N-gram results

COMPARISON OF MOST ACCURATE CLASSIFIERS

Algorithm	Char N-gram combination	Processing Time (ms)	F1	Acc	Kappa
Multinomial Naïve Bayes	N: 2,3,4	470	0.80	81.99	0.63
	N: 3,4	303	0.81	82.35	0.64
Complement Naïve Bayes	N: 2,3,4	437	0.80	81.99	0.63
	N: 3,4	293	0.81	82.35	0.64
Logistic Regression	N=4	437	0.80	80.60	0.61
	N: 3,4	293	0.80	80.39	0.60

V. 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 ML algorithms for the sentiment analysis.

In terms of feature extraction, we investigated the performance of word N-grams (N ranging from 1 to 3) and character N-grams (N ranging from 2 to 7). Character N-grams outperform word N-grams in terms of accuracy. This indicates that character N-grams require less prior knowledge compared to word N-grams and word N-gram features are too strict in collecting evidence from the underlying sentiments. Given the size of the data set we assume that word N-grams needs more labeled data to achieve higher accuracies compared to character N-grams. Character N-grams results in a performance overhead.

Combinations of N-grams produced higher accuracies for both word and character N-grams, but their processing durations were longer than that of any of the corresponding individual N-grams. N: 1, 2 was the best combination for word N-grams and N: 3, 4 was the best combination for character N-grams. In Sinhala script, in order to construct a string of 2 Sinhala letters a minimum of 2 characters and a maximum of 4 characters are combined. Combinations of n values: 2, 3, 4 and n values: 3, 4 has provided the bet results indicating that 2 letterer character N-gram features are the most informative features.

Complement Naïve Bayes and Multinomial Naïve Bayes algorithms produced the best results for character N-grams analysis.

Character level N-grams provide the ability to detect the similarity of words which allows machine leaning algorithms employed in sentiment analysis to detect spelling mistakes. This is because mis-spelling and noise tends to have a minimum impact on substring patterns than word patterns. In addition it also provides the ability to the machine leaning algorithms to function as a stemmer.

Adjacent word awareness in char N-gram tokens was expected to improve the results. However it has reduced a very small percentage of the classification accuracy. We assume that with a larger corpus adjacent word awareness will become more effective.

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

- Exploring the use of Sinhala syllables as features
- Reducing Performance overhead by constructing a character N-gram based Sinhala stemmer
- Using ensemble learning methods to improve accuracy.
- Using Deep Learning and other Neural Network-based methods.

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. 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].

[7] 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.

[8] M. S. N. Y. Y. T. and S. H. , "Multilingual text categorization using Character N-gram," in *IEEE Conference on Soft Computing in Industrial Applications*, 2008.

[9] P. Majumder and M. Mitra, "N-gram: a language independent approach to IR and NLP," in *International Conference on Universal Knowledge and Language*, Goa, India, November 2002.

[10] . J. Houvardas and E. Stamatatos, "N-Gram Feature Selection for Authorship Identification," in *Artificial Intelligence: Methodology, Systems, and Applications, 12th International Conference*, Varna, Bulgaria, 09-2006.

[11] . B. Gyawali, G. Ramirez-de-la-Rosa and T. Solorio, "Native Language Identification: a Simple n-gram Based Approach§," in *Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 2013/06/01.

[12] T. Hartmann, S. Klenk, A. Burkovski and G. Heidemann, "Sentiment Detection with Character n-Grams," 2011.

[13] "movie-review-data," [Online]. Available: http://www.cs.cornell.edu/people/pabo/movie-review-data/. [Accessed 15 03 2020].

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

[15] 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.

[16] 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.

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

[18] S. Roy and F. Ali, "Unsupervised Context-Sensitive Bangla Spelling Correction with Character N-gram," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 12-2009.

[19] J. Xu and W. B. Croft, "Corpus-Based Stemming Using Cooccurrence of Word Variants," New York, NY, USA, January 1998.