ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 1-5
doi: 10.18178/wcse.2020.02.001

Sentiment Analysis System for Myanmar News using K Nearest Neighbor and Na ïve Bayes

Thein Yu [1+]and Khin Thandar Nwet[2]

[1] University of Computer Studies, Yangon, Myanmar

[2] University of Information Technology, Myanmar

Abstract. With the explosive growth of internet technology, there are very large amount of information on the web for the internet users. Users not only use that information but also provide opinions for decision making process. Sentiment analysis or opinion mining is one of text classification techniques that identify and extract opinion described in a piece of text. Our aims in this paper are to develop automatic sentiment analysis system for Myanmar news and to annotate sentiment news. Therefore, this system creates sentiment annotated corpus for Myanmar news. Feature extraction and selection are very important for sentiment analysis to get higher performance. N-grams, Countvectorizer, and TF-IDF are used for feature selection and feature extraction. In this system, Myanmar news sentiment analysis system is implemented by using K Nearest Neighbor (KNN) and Na ïve Bayes machine learning algorithms.

Keywords: Sentiment analysis, Na ïve Bayes, K Nearest Neighbor, N-gram, TF-IDF.

1. Introduction

Opinion mining or sentiment analysis is a popular research field in the combination of information retrieval (IR) and natural language processing (NLP) and have common characteristics with other disciplines such as text mining and information extraction. Sentiment analysis or opinion mining is also the task that intends to refer the sentiment orientation in a document. Opinion mining is a technique of text mining that provides a way for individual and corporation to exploit the large amount of information available on the internet and to detect and extract polarity orientation of subjective information in text documents. There are three levels of sentiment: (i) document-based level; (ii) sentence-based level; and (iii) aspect-based level. In document-based and sentence-based sentiment analysis, it is implicitly assumed that the analysed document or sentence only discusses a single object. In general, sentiment analysis determines the sentiment orientation of a writer about some aspect or the overall contextual polarity of a document. Sentiment classification is a recent sub division of text classification which is concerned not only with the topic of document, but also with the expressed opinion. Sentiment classification also has different names, among which opinion mining, sentiment analysis, sentiment extraction, or affective rating. News can be good or bad, or seldom neutral. The statistical analysis of sentiment cues can give a powerful sense of how the latest news impacts important entities [1].

Sentiment analysis for Myanmar language has many challenges due to scarcity of resources such as automatic feature extraction tools, stemming, anaphora resolution, and name entity recognition etc. In this paper, sentiment analysis system of Myanmar news is proposed. Feature extraction and transformation are used in sentiment analysis to get better performance. N-Gram and TF-IDF are used in this system for feature extraction and transformation. K nearest neighbour and na ïve bayes algorithms is applied in implementing sentiment analysis system.

[+] Corresponding author. Tel.: +959898681179
E-mail address: theinyu@ucsy.edu.mm.

The remaining parts of this paper are organized as follows; the related works are explained in section 2. The methodology is described in section 3. In section 4, experiment showed. Conclusion and future work is presented in section 5.

2. Related Work

Many sentiment analysis systems are existed for English language and other languages. Different methods are used with different resources at different levels for different systems.

Joseph Lilleberg, Yun Zxu, and Yanqing Zhang proposed a text classification system with semantics features. They used TF-IDF model and combined with word2vec model. They show performance result of TF-IDF, word2vec combined with wighted TF-IDF with stop words and without stop words[2]. In paper [3], authors implemented Chinese text classification system using N-gram based feature selection and text representation methods. They used four feature selection methods such as absolute text frequency, relative text frequency, absolute n-gram frequency and relative n-gram frequency and three text representation methods such as 0/1 logical value, n-gram frequency numeric value (TF) and TF-IDF value with support vector machine and naive bayes machine learning algorithm.

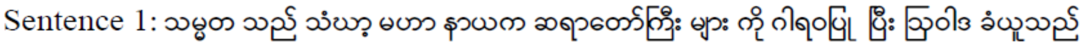
In paper [4], authors developed twitter sentiment system that use N-gram feature selection and combination. They use Sanders product review dataset. They model system using kern lab classifier, decision tree classifier, and naive bayes classifier. In paper [5], authors implemented improved text sentiment classification model that use TF-IDF and next word negation for feature. They used movie review dataset, product review dataset, and SMS spam dataset that are trained with linear support vector machine, maximum entropy, random forest, and multinomial naive bayes.

3. Proposed System

3.1. Preprocessing

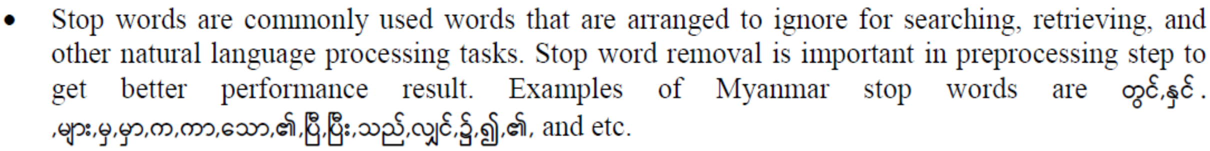
Preprocessing is the important step in natural language processing. There are three preprocessing steps in the proposed system.

- Word Segmentation is the fundamental task in natural language processing that identify boundaries of word. Myanmar word segmentation is the process of placing spaces into textual data without other replacing or rewriting operations. This system used word segmentation tool from NLP Lab, University of Computer Studies,Yangon, Myanmar examples of segmented result are as follow:



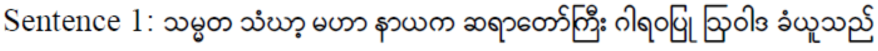
Sentence 1: သမ္မတ သည် သံဃာ့ မဟာ နာယက ဆရာတော်ကြီး များ ကို ဂါရဝပြု ပြီး ဩဝါဒ ခံယူသည်

- Tokenization is the process of separating up a sequence of strings into words, phrases, keywords and other elements. Tokens or words are separated and identified by white space, punctuation marks or line breaks.



- Stop words are commonly used words that are arranged to ignore for searching, retrieving, and other natural language processing tasks. Stop word removal is important in preprocessing step to get better performance result. Examples of Myanmar stop words are တွင်,နှင့် . ,များ,မှ,မှာ,က,ကာ,သော,၏,ပြီ,ပြီး,သည်,လျှင်,၌,၍,၏, and etc.

After removing stop words, sentence 1 may be as follow:

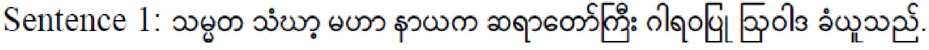


Sentence 1: သမ္မတ သံဃာ့ မဟာ နာယက ဆရာတော်ကြီး ဂါရဝပြု ဩဝါဒ ခံယူသည်

3.2. Feature Extraction and Transformation

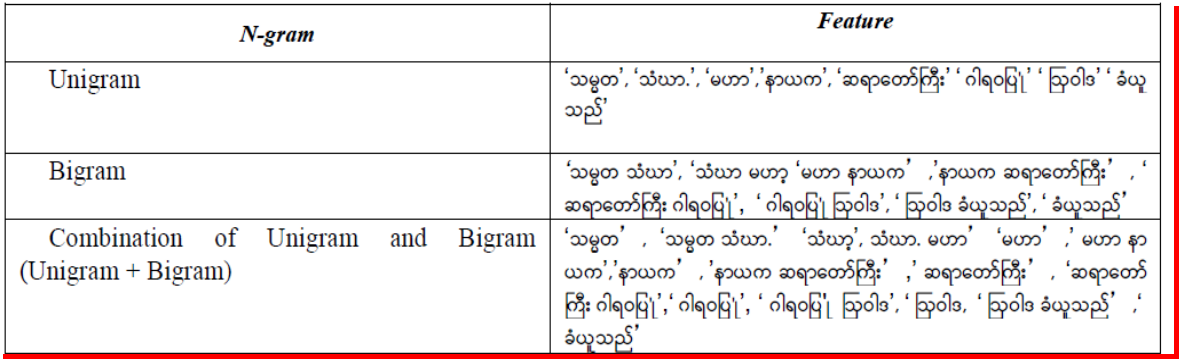
- **N-gram :** N-gram is a language models that assign probabilities to the sequences of words. N-gram is based on bag of word model and has a sequence of word with n length. N-gram with length (n=1) is called unigram and length (n = 2) is also called bigram and then length (n = 3) is also called trigram. Text classification also depends on text representation to get higher accuracy [6].

Examples of n-gram words for sentence 1 after removing stop words is shown in table 1.



Sentence 1: သမ္မတ သံဃာ့ မဟာ နာယက ဆရာတော်ကြီး ဂါရဝပြု ဩဝါဒ ခံယူသည်.

Table 1: Examples of N-gram



N-gram	*Feature*
Unigram	'သမ္မတ', 'သံဃာ.', 'မဟာ','နာယက', 'ဆရာတော်ကြီး' ' ဂါရဝပြု' ' ဩဝါဒ' ' ခံယူသည်'
Bigram	'သမ္မတ သံဃာ', 'သံဃာ မဟာ့ 'မဟာ နာယက' ,'နာယက ဆရာတော်ကြီး' , ' ဆရာတော်ကြီး ဂါရဝပြု', ' ဂါရဝပြု ဩဝါဒ', ' ဩဝါဒ ခံယူသည်', ' ခံယူသည်'
Combination of Unigram and Bigram (Unigram + Bigram)	'သမ္မတ' , 'သမ္မတ သံဃာ.' 'သံဃာ့', သံဃာ. မဟာ' 'မဟာ' ,' မဟာ နာယက','နာယက' ,'နာယက ဆရာတော်ကြီး' ,' ဆရာတော်ကြီး' , 'ဆရာတော်ကြီး ဂါရဝပြု',' ဂါရဝပြု', ' ဂါရဝပြု ဩဝါဒ', ' ဩဝါဒ, ' ဩဝါဒ ခံယူသည်' ,' ခံယူသည်'

- **CountVectorizer :** The Countvectorizer is one of the bag of word model to tokenize text document and build a vocabulary of known words. And then, it also encodes document into document term matrix vector with that vocabulary. The encoded term matrix vector contains length of entire vocabulary and integer count number of each word presented in the document. The encoded vector is transformed to array version of vector that can be directly used by machine learning algorithm.

- **TF-IDF Vectorizer:** TF-IDF is a term weighting method which shows the importance of a term in a document to present textual data. It compares the frequency of word described in an individual document as opposed to the entire document. TF-IDF is based on the bag-of-words (BoW) model and does not necessary to have position in text, semantics, co-occurrences in different documents, etc[7].

3.3. Machine Learning Algorithm.

- **Na ïve Bayes:** The bayesian classification is one of probabilistic learning method for text classification. Naive bayes classifier is an independent features model that the inclusion (or exclusion) of a particular feature of a class is unrelated to the inclusion (or exclusion) of any other feature[8]. The probability function of sentiment class given document is calculated using equation 1and 2.

$$P(c/d) = P(c)P(d/c)/P(d) \tag{1}$$

Where, c=sentiment class, p(c/d) = probability of sentiment class given document, p(d/c)= likelihood of document, p(d)= probability of document, p(c) =prior probability d=document
P(d) has the same value, so p(d)can be drop. Document can be presented as many features such as f_1, f_2, f_3,…f_n, function can be as follows:

$$P(c/d) = argmaxP(c)\prod_{f\epsilon F} P\,(f/c) \tag{2}$$

Where, f= features vector, c= sentiment class, d=document, p(f/c)= likelihood of features given to sentiment class

- **K Nearest Neighbor (KNN) Algorithm :**Nearest neighbor algorithm is one of the simplest machine learning algorithms. The KNN algorithm is also a lazy learning because the computation for the generation of the predictions is postponed until classification. The idea is to acquire the training set and then to predict the label of any new instance on the closest labeled neighbors in the training set. The algorithm works based on the rule that chooses minimum distance from the test data to the training samples to determine the K nearest neighbor. After defining K nearest neighbor, a simple majority of them is used to predict test data. The KNN works as follows: The distance between the test data and all the training samples are calculated. The distance may be calculated by any standard means. Euclidean distance is usually used. The K nearest neighbor may be considered if the distance of the training samples to the test samples is less than or equal to Kth smallest distance. The quality of the predictions relies on the distance measure. The KNN algorithm is suitable for applications for sufficient domain knowledge [9, 10].

4. Experiment

4.1. Experimental Apparatus

Na ïve bayes and K nearest neighbor algorithms were experiment on the Myanmar News dataset. The dataset was split in the ratio of 80 % for training and 20 % for testing purposes. These experiments were carried out using an open source ‘scikit-learn’ Python library and NLTK Library on Jupyter Notebook.

4.2. Training Dataset

Myanmar news data are collected from web sites and ALT Tree bank for training and testing data. Sentiment corpus contains 2000 news for positive and 1000 news for negative. The task is to identify if positive or negative sentiment is expressed at document and sentence level. There has been found that many researchers used unbalanced dataset for sentiment analysis. Therefore, recently dataset are used in the proposed system.

4.3. Experimental Result

The hold out evaluation method is used in the experiment. Na ïve bayes with countvectorizer (unigram) features gets greatest accuracy score values. The evaluation metrics such as accuracy, precision, recall and F1 measure were calculated using equation 3-6. Table 2 and 3 show performance results.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{4}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{5}$$

$$\text{F-Measure}=\frac{2*Precision*Recall}{Precision+Recall} \tag{6}$$

TP defines the number of positive news that are correctly classified, as positive,
FP is the number of negative news that are incorrectly classified as positive.
TN is the number of negative instances that are correctly classified as negative.
FN is the number of positive tuples that are incorrectly classified as negative news.

Table 2: Performance Results 1

Feature with TFIDF	**Na ïve Bayes**	**KNN**
	Accuracy %	**Accuracy %**
Unigram	75.67	80.83
Bigram	69.33	74
Trigram	67.83	70
Unigram + Bigram	71.83	79.67
Bigram +Trigram	68.50	73.50
Unigram + Bigram+ Trigram	69.50	79.83

Table 3: Performance Results 2

Features with CountVectorizer	**Na ïve Bayes**	**KNN**
	Accuracy %	**Accuracy %**
Unigram	82.33	68.67
Bigram	77.00	68.33
Trigram	69.50	68.33
Unigram + Bigram	81.17	68.17
Bigram+Trigram	74.17	68.33
Unigram + Bigram+ Trigram	79.83	68.00

5. Conclusion and Future Work

The proposed system is implemented to classify sentiment of news in Myanmar language. By using system, user can easily feel emotion of news. K nearest neighbor and na ïve bayes machine learning algorithms are used in this system. This system compares accuracies of those algorithms with unigram, bigram, trigram, combination of unigram and bigram, combination of bigram and trigram (bigram+trigram), and combination of unigram, bigram, and trigram(unigram+ bigram + trigram) features. Na ïve bayes with Countvectorizer(unigram) feature has highest accuracy value in the proposed system. In future, we intend to classify with many algorithms such as CNN, ANN, and RNN deep learning algorithms.

6. Acknowledgements

I specially thank to my supervisor, Dr Khin Thandar Nwet, Lecturer, University of Information Technology, Yangon, Myanmar. I would like to thank Dr Win Pa Pa, Professor, Natural Language Processing Lab, University of Computer Studies, Yangon, for word segmentation tool. I deeply thanks to anonymous reviewers for my paper.

7. References

[1] B. Liu. Sentiment Analysis and Opinion Mining", Synthesis Lecturer on Human Language Technologies, Morgan & Claypool Press, 2012

[2] J. Lilleberg. Y. Zhu, and Y. Zxang. Support Vector Machines and Word2vec for Text Classification with Semantic Features, Proc. 20151IEE 14th Inl'l Coni. on Cognitive Inlormatics & Cognitive Computing IIccrcn51, 2015

[3] Z. Wei, D. Maio, J. H. Chauchai, and R. Z. Wenli. N-gram based feature selection and text representation for Chinese text classification, International Journal of computational Intelligence Systems, Vol.2, No.4, pp- 365-374December, 2009

[4] P.B. Awachate and V.P. Jsgursagar. Improved Twitter Sentiment Analysis using N Gram feature selection and Combination, International Journal of Adavanced Research in computer and communication engineering, Vol.5Issue 9, September 2016.

[5] B. Das and S. Chakraborty. An Improved Text Sentiment Classification Model using TF-IDF and Next word Negation,

[6] Jurafsky ,D., & Martin, J.,H.(2018) .N-gram Language Models. Speech and Language Processing, September 23, 2018.

[7] A. Aizawa.(2003) .An information-theoretic perspective of tf–idf measures. Information Processing and Management 39 (2003) 45–65.

[8] How to Prepare Text Data for Machine Learning with scikit-learn (2003). https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn.

[9] http://nlp.stanford.edu/IR-book/html/html edition / naive-bayes-text-classification-1.html.

[10] S.Shwartz, S.and S. Ben-David. Understanding Machine Learning. Cambridge University Press, 2014.