Annotation and Sentiment Analysis System for Myanmar News Using Naïve Bayes Algorithm

[1] Thein Yu, [2] Khin Thandar Nwet, [1] May Kyi Nyein
[1] Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
[2] Faculty of Computer Software, University of Information Technology, Yangon, Myanmar
{theinyu, maykyinyein}@ucsy.edu.mm; khinthandarnwet@uit.edu.mm

***Abstract*— Sentiment analysis or opinion mining is a combination technique of computational linguistics, natural language processing, and text analytics. Sentiment analysis provides important pieces of subjective information during decision making process. News provides very important information for people. There are many sentiment analysis researches for English and other language, but there is a little sentiment analysis research for Myanmar language. In Myanmar language, there are challenges in scare of resources. We construct Myanmar news sentiment tagged corpus and news data are collected from many websites such as duwun.com, news-eleven.com, 7daydaily.com and popularmyanmar.com, ayeyarwady.com and so on. In developing sentiment analysis, identification and extraction of feature information in source domain are also needed to perform. We use N-gram feature extraction process in this research to get more performance. Naïve Bayes is very similar and intuitive classification algorithm. In this paper, statistical approach such as Naïve Bayes algorithm is applied in document level sentiment analysis. We implement this system using Python Jupyter Notebook. We use 600 news data set including 300 positive news and 300 news positive news.**

Keywords: Sentiment Analysis, Naïve Bayes, N gram

I. INTRODUCTION

The success of Web 2.0 technologies along with the growth of social content available online have stimulated and provided many opportunities for understanding the opinions and trends, not only of the general public and consumers, but also of companies, banks, and politics[1]. Sentiment analysis has been more known in recent years for automatic customer satisfactions analysis of online services such as blogging and social network as it can provide business insights by classifying public opinion on social data [2]. It is also widely applied in data mining, web mining, and information retrieval [4]. Sentiment analysis is one of the text mining tasks. Forums, blogs, review sites etc. are providing the liberty to individuals to speak their mind with no constraints. Much of that text data is derived by the sentiments of the person which provides a new way to categorize those chunks of data. Tagging those text data based on sentiments provides a better, more clearer and a strong insight to the reader. Sentiment classification also plays a significant role in classifying and tagging important news and information from the more casual content [3]. There are three computational sentiment analysis approaches such as linguistic, statistical and hybrid. Different approaches use difference domain resources. And there are three levels for sentiment analysis such as word or aspect, sentence, and paragraph.

Sentiment analysis applications have applied to almost every possible domain, from consumer products, services, healthcare, and financial services to social events and political elections. There are many sentiment analysis systems for English and other languages. But, recently Myanmar automatic sentiment analysis system is not widely applied. Therefore, developing sentiment analysis systems for Myanmar documents is a challenging task due to scarcity of resources of the language like automatic tools for tokenization, feature selection and stemming etc.

In this paper, an automatic sentiment analysis system for Myanmar news is proposed. The proposed system is implemented by using supervised learning approach which assigns the pre-defined category labels to the text documents based on the likelihood suggested by the training set of labeled documents. For the training data set, news from Myanmar media websites are manually collected, labeled and stored in the training data set.

N-gram is a simply all combinations of N items from a given text with *n* represent starting from 1.We use N-gram feature extraction process in this research to get more performance. Naïve Bayes is very similar and intuitive classification algorithm. In this paper, statistical approach such as Naïve Bayes algorithm is applied in document level sentiment analysis. We implement this system using Python Jupyter Notebook.

The remaining parts of this paper are organized as follows. In section II, related works is discussed. Methodology and system design are explained in section III. Conclusion and future work is presented in section IV.

II. RELATED WORKS

There are many researches for sentiment analysis system. Different Systems use different resources and different algorithms. In the paper written by Y Al-Amrani, M Lazaar, K E El Kadiri, system is designed for sentiment analysis of messages (SMS, Facebook, Twitter). They use bag of word model for vector transformation. They use "SMS Spam Collection Data Set . They implement hybrid method of decision tree and support vector machine and compare six different learning methods such as (Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), PART and Logistic Regression (LR)) and hybrid approach Decision Tree Support Vector Machine (DTSVM). They show the proposed approach has high accuracy and low CPU time than the other algorithms [1].

N Godbole, M Srinivasaiah, and S Skiena implement system that assigns score value to news and blogs. They create sentiment lexicon and corpus[2]. P Waykar, K

Wadhwani, P More from Department of Computer Engineering, DYPIET, Pimprihas been implement sentiment classification system using supervised approach. The unigram algorithm was used to derive feature set from twitter dataset and Naïve Bayes classifier was then applied on derived features for final categorization [4]. S Tripathi presented Naïve Bayes classifier to classify movie data from the Internet Movie Database(IMDb)[5]. M A Sghaier and M Zrigui implement a tool for sentiment analysis able to find the polarity of opinions in reviews extracted from e-commerce magazines and blogs in Arabic language. They implement a small (symbol to Word) converter for the use of emoticons and a checker for the use of the elongated words. They present the reviews in six different models (unigrams, bigrams, tri-grams, unigrams + bigrams, bigrams + trigrams, unigrams + bigrams + trigrams) for feature extraction. They get the best results by the combining Unigrams, Unigrams + Bigrams, Unigrams + Bigrams + Trigrams with a standard corpus with the application of the Arabic light stemming classified by Naive Bayes. They also obtain the best results by the combinations: Unigrams + Bigrams, Unigrams + Bigrams + Trigrams with a preprocessed corpus, stemmed and classified by Support Vectors machine (SVM) [6]. In the paper written by T. S.N Ayutthaya and K. Pasupa, authors implement Thai sentiment analysis for Thai children stories. They work incorporation of part-of-speech and sentic features. They classify sentiment by using combination of Bidirectional Long Short-term Memory and Convolutional Neural Networks models. Their best result of f1 score is 78.79. They use 40 Thai stories dataset [11]. In paper written M.Okada , H. Yanagimoto and K. H. ESSAND , authors developed sentiment analysis system for customer review. They use Amazon Product Review dataset and Japanese review dataset of TripAdvisor. They also used gated convolutional neural networks model and then compare the best results of gCNN faster than RNN with different language datasets [12].

A. Myanmar Language and Corpus Creation

Myanmar language is an official language of the Republic of the Union of Myanmar. The Myanmar alphabet consists of 33 letters and 12 vowels, and is written from left to right. It doesn't need spaces between words, although modern writing usually consists of spaces after each clause to enhance readability. Myanmar words are sequences of syllables and written in circular shape. Its sentences structure is the subject-object-verb (SOV) order. In Myanmar language, words such as ကောင်းသော, သာယာသော,လှပသော, နားလည်မှု , ဝမ်းမြောက်ကြောင်း are tagged as positive word. And then, words such as ဆိုးသော, ရှုပ်ဆိုးသော, မကောင်းသော, ရေနစ်သည်, ပယ်ဖျက်သည်,သဘာဝဘေးအန္တရာယ်, ငလျင်လှုပ် are tagged as negative words.

Examples of positive sentences are as follows:

- နိုင်ငံတော်၏အတိုင်ပင်ခံပုဂ္ဂိုလ်နှင့်.အိန္ဒိယနိုင်ငံပြည်ပရေးရာဝန်ကြီးဌာနForeignSecretaryတို့.တွေ့.ဆုံပီးနှစ်နိုင်ငံအစိုးရအကြားရခိုင်ပြည်နယ်ဖွံ့.ဖြိုးရေးဆိုင်ရာနားလည်မှုစာချွန်လွှာလက်မှတ်ရေးထိုးခဲ့သည်
- မြန်မာနိုင်ငံ၏နှစ်၇၀ပြည်.လွတ်လပ်ရေးနေ.အထိမ်းအမှတ်အဖြစ်သမ္မတဦးထင်ကျော်ထံကမ္ဘာ.ခေါင်းဆောင်များကဝမ်းမြောက်ကြောင်းသဝဏ်လွှာပေးပို့.သည်။

Examples of negative sentences are as follows:

- ငလျင် လှုပ်ခက်မှု ကြောင်. အီရန်မြို့.တော် တီဟီရန်တွင် ပြည်သူများ လမ်းမများပေါ် ပြေးထွက်ခဲ့ရသည်။
- လတ်တလောရင်ဆိုင်ကြုံတွေ့.နေရသည့်သဘာဝဘေးအန္တရာယ်ကျရောက်မှုများအပေါ် သက်ဆိုင်ရာဝန်ကြီးဌာနများအလိုက်နိုင်ငံတကာအဖွဲ့.အစည်းများနှင့်.ဆက်သွယ်ပီးကယ်ဆယ်ရေး၊ထောက်ပံ့ရေးနှင့်.ပြန်လည်ထူထောင်ရေးလုပ်ငန်းများကိုပေါင်းဆောင်ရွက်လျက်ရှိသည်။

III. METHODOLOGY AND SYSTEM DESIGN

A. Word Segmentation

Word Segmentation is a basic task for Myanmar text as it doesn't consist of white space to delimit the words like other Asian languages. Segmentation is merely a preprocessing task for a wide variety of further tasks such as parsing, information extraction, and information retrieval. We use word segmentation tool to segment word in the proposed system. [7][8].

C. N-Grams

N-grams are items (words) next to each other [9]. N-gram is composed of adjacent words or letters of length n. When length n is 1, it defines to a unigram; when n is 2, it refers to a bigram; when n is 3, it refers to a trigram. In this system, N-gram are used to select feature. In nature of Myanmar language, sometime single word shch as စွန့်ပစ်မှု,လှုပ်ခတ်မှု etc. can even lead to misclassification. Sequences of word (eg.Bigram)စည်းကမ်းမဲ့|စွန့်ပစ်မှု and ငလျင်|လှုပ်ခတ်မှု etc. may overcome this problem. Therefore, N gram is used in this system than other methods. By using n gram, the system accuracy is higher. Examples of Myanmar negative word are shown in table1 and examples of positive Myanmar words are shown in table2.

TABLE 1. N-GRAM NEGATIVE MYANMAR WORDS

Negative Myanmar Words	
Unigram	**Bigram**
ဝင်တိုက်	စည်းကမ်းမဲ့|စွန့်ပစ်မှု
ဖောမလင်	ဖမ်းဆီး၊စစ်ဆေး
စိုးရိမ်	ငလျင်|လှုပ်ခတ်မှု

TABLE 2. N-GRAM POSITIVE MYANMAR WORDS

Positive Myanmar Words	
Unigram	**Bigram**
ဆောင်ရွက်နိုင်	ထိန်းသိမ်း၊ စောင့်ရှောက်
နားလည်မှု	ထုတ်လုပ်မှု၊ မြင့်တက်
ဝမ်းမြောက်	အနည်းငယ်၊ ကျဆင်း
အကျိုးပြု	ပြန်လည်၊ လွတ်မြောက်လာခဲ့ ပြီ
ထိမ်းသိမ်း	ဆုတောင်း ၊မေတ္တာ

D. Naïve Bayes Myanmar Text Classifier

Naïve Bayes method which is based on Bayesian theorem is used for classifying sentiment. It is also an approach to text classification. A naive Bayes classifier is a simple probabilistic classifier and is particularly suited when the dimensionality of the inputs are high. Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling [10]. Naive Bayes is also a classification algorithm to solve binary (two-class) and multi-class classification problems. The basic idea in Naïve Bayes approach is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. In Naïve Bayes classifier, each document is viewed as a collection of words and the order of words is considered independent. Given a documents d for classification, the probability of each sentiment c is calculated as follows:

$$P(c/d) = P(P(c)P(d/c))/P(d) \quad (1)$$

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

Where, F is the feature vector. Naïve Bayesian classifier simplifies (naive) to assume about how the features interact. Naïve Bayes method is suitable for sentiment classification with independent features. The proposed system is implemented by using Python NLTK Toolkit on Jupyter Notebook.

E. *System Design Architecture*

In our system, we need to preprocess news data using segmentation and then generate features with n-gram. And those features are classified using Naïve bayes to deliver sentiments of news. Validation and Evaluation processes are done to test performance.

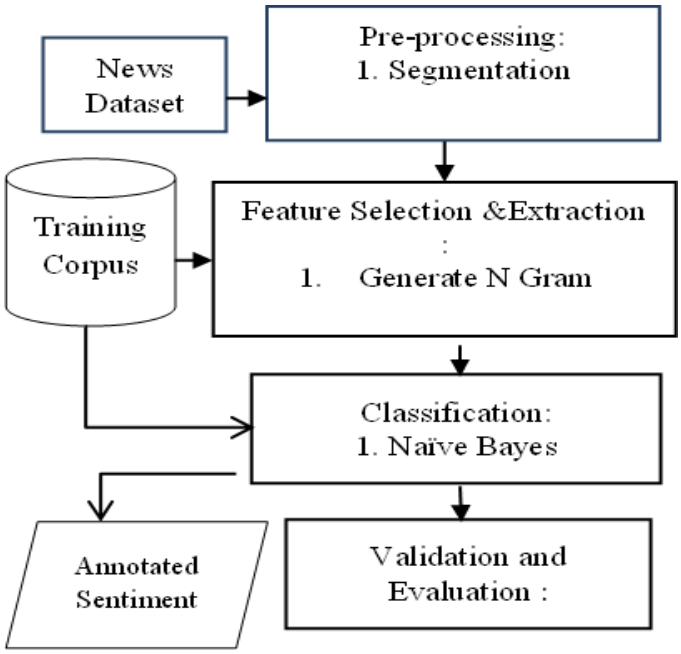

Figure1. System design

F. Dataset

Dataset are manually collected from Myanmar news websites. This system contains 600 news which consist of 300 positive tagged words and 300 negative tagged words. Each news contains average 5 sentences and totally 3000 sentences. In my research, structure news data are used

IV. Conclusion

Opinions are central to almost all human activities and key influencers of behaviors. Sentiment analysis is commonly used in several areas that include tracking sentiment towards products, movies, politicians, and companies. In this system, Myanmar news data are classified as positive and negative sentiments. Data sets are collected from various Myanmar web sites. Sentiment annotated corpus is created. Myanmar news sentiment analysis can be useful for natural language processing research environment. This system is developed using supervise machine learning method, Naïve Bayes. More training data will give more performance. In future, we intend to train and test more dataset. We aim to test other classifiers such as SVM, neural network, Randon Forest etc.

Acknowledgment

Word Segmentation process is supported by NLP Lab, UCSY.

References

[1] Y. Al-Amrani, M. Lazaar, and K. E. E Kadiri, “Sentimrnt Analysis Using Hybrid Method Of Support Vector Machine And Decision Tree”, Journal of Theoretical and Applied Information Technology, Vol.96. No 7, 15th April 2018.

[2] N. Godbole ,*M. Srinivasaiah.and* S. Skiena , “Large-Scale Sentiment Analysis for News and Blogs” , ♦Dept. of Computer

Science, Stony Brook University, Stony Brook, NY 11794-4400, USA, ICWSM'2007 Boulder

[3] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[4] P. Waykar, K. Wadhwani, and P.More, "Sentiment analysis in twitter using Natural Language Processing (NLP) and classification algorithm ", International Journal of Advanced Research in Computer Engineering & Technology(IJARCET), Volume 5 Issue 1, January 2016, ISSN:2278-1323

[5] S.Tripathi, "Bigram Extraction and Sentiment Classification on Unstructured Movie Data, International Journal of Electrical and Electronics Research , Volume.3, Issue 3, July-September 2015

[6] M.A Sghaier and M. Zrigui"Sentiment Analysis for Arabic e-commerce websites", June 2015.

[7] W. P. Pa, N. Thein," Myanmar Word Segmentation Using a Combined Model", University of Computer Studies, Yangon, Myanmar

[8] C. DING, Y. K .THU, M .UTIYAMA, and E .SUMITA, "Word Segmentation for Burmese (Myanmar)", National Institute of Information and Communications Technology, ACM Trans. Asian Low-Resour. Lang. Inf. Process., Vol. 15, No. 4, Article 22, Publication date: May 2016.

[9] https://www.sketchengine.eu/user-guide/user-manual/n-grams/

[10] https://machinelearningmastery.com/naive-bayes-for-machine-learning/

[11] T. S.N Ayutthaya and K. Pasupa, "Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features", Faculty of Information Technology King Mongkut's Institute of Technology Ladkrabang Bangkok 10520, Thailand,

[12] M.Okada , H. Yanagimoto and K. H. ESSAND "Sentiment Classification with Gated CNN for Customer Reviews",