Automatic Sentiment Analysis System for Myanmar News

Thein Yu, KhinThandarNwet

University of Computer Studies, Yangon

{theinyu, khinthandarnwet}@ucsy.edu.mm

Abstract

Sentiment analysis and opinion mining mainly emphasize on opinions which express or describe positive or negative sentiments. Sentiment analysis is mostly concerned with NLP. The inception and rapid development of the field concerned with the social media on the Web, e.g., post, reviews, forum discussions, blogs, microblogs, Twitter, and social networks. There is a little research in Myanmar language. This system implements sentiment analysis system for Myanmar news using well-known machine learning approach such as Support vector machine. Feature extraction and selection are also need for sentiment analysis to improve accuracy. This system uses N gram and TF-IDF method for feature selection and extraction. Sentiment Lexicon is constructed and news corpus is used for training and testing. This system uses News data set from different web page in Facebook social media

Keywords: **Sentiment Analysis, Opinion Mining, Natural Language Processing, N-gram, TF_IDF, SVM**

1. Introduction

Sentiment is basically also termed as the expression of sensitive feeling in art and literature. Sentiment Analysis also defined as Opinion Mining is a Natural Language Processing and Information Extraction task that intend to get writer's feelings expressed in positive or negative comments. Sentiment analysis is the computational technique for extracting, classifying, understanding and determining the opinions describes in various subject. It concerned with natural language processing (NLP) and computational techniques to automatically extracted or classified of sentiment from typically unstructured text. There are two approach for sentiment analysis: linguistic approach and statistical approach.

Research in sentiment analysis not only has an effect on NLP, but may also have a profound effect on management sciences, political science, economics, and social sciences as they are all affected by people's opinions. Sentiment analysis is now important role of the social media research.

The social media has now placed the major space on the Web. In Myanmar, social media, especially Facebook are used by most people to express their opinions about specific topics in the Myanmar language. The content that read is on the negative aspect is spreading negativity in environment. The news that provide positive aspect is dominating positivity in environment. Many sentiment analysis systems for English and other language are exist, but recently Myanmar sentiment analysis system are not widely used. There are challenging tasks to develop sentiment analysis systems are scarcity of resources of the language like automatic tools for tokenization, feature selection and stemming etc.

This research intends to analyze sentiment of Myanmar news as in positive and negative. Supervise machine learning algorithm are used for sentiment classification. Support vector machine (SVM) is used in text classifier. Feature selection and extraction are used to get more accuracy in

classification. Tf-Idf and N-gram are used for this system.

The remaining parts of this paper are organized as follows. The related works are explained in section 2. In section 3, the nature of Myanmar language and sense annotated corpus. System design are shown in section 4. Conclusion and future work are presented in section 5.

2. Related Work

There are many sentiments analyses research for various domain and language. But, there are a little research for Myanmar language. In the paper written by Y Al-Amrani, M Lazaar, K E El Kadiri, 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 developed hybrid method of decision tree and support vector machine and compare six different learning method 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 describe the proposed approach has high accuracy and low CPU time than the other algorithms [1]. Y M Aye and S S Aung create the Myanmar sentiment lexicon for food and restaurant domain and analyze the sentiment of Myanmar text customer reviews for recommendation. They also generate the context-independent rules for Myanmar language[3].

A Gupta, J Pruthi, N Sahu show that hybrid manner of various machine learning approaches can gives better result than using these approaches in isolation. They use two machine learning K-Nearest Neighbours (KNN) and Support Vector Machine in hybrid manner. They use more tweet data and use adjective as feature [5]. In paper [6], authors explore the use of virtual examples to improve performance of text classification with support vectors machine. They use Reuters-21758 data set collection.

There are many approaches that have been used in sentiment classification, using different resources at different levels. In machine learning approach, there are many classification algorithms for sentiment analysis such as SVM, Neural Network, Naïve Bayes, Bayesian Network and Maximum Entropy

3. Myanmar Language and Corpus

Myanmar language is an official language of Myanmar country. Myanmar alphabet composed of 33 letters and 12 vowels, and is written from left to right. The order of subject-object-verb (SOV) is Myanmar sentence structure. Its word structure are sequences of syllables and written structure is circular shape. In corpus, news that feel happy, delithed, wonderful are annotated as positive news. In Myanmar language, နားလည်မှု , ပမ်းမြောက် etc are used positive words as follow.

- နိုင်ငံ တော်၏အတိုင်ပင်ခံပုဂ္ဂိုလ်နှင်. အိန္ဒိယနိုင်ငံ ပြည်ပရေးရာပန်ကြီး ဌာန Foreign Secretaryတို.တွေ.ဆုံပီးနှစ်နိုင်ငံ အစိုးရအကြား ရရိုင်ပြည်နယ် ဖွံ.ဖြိုးရေး ဆိုင်ရာ နားလည်မှုစာချွန်လွှာ လက်မှက်ရေးထိုး
- မြန်မာနိုင်ငံ၏ နှစ် ဂုပပြည်. လွတ်လပ်ရေးနေ. အထိမ်းအမှတ်အဖြစ် သမ္မတ ဦးထင်ကျော်ထံ ကမ္ဘာ.ခေါင်းဆောင်များက ပမ်းမြောက်ကြောင်း သဝဏ်လွှာပေးပို.

News that feel sad, nervous are annotated as negative news. In Myanmar language, ဘေးအန္တရာယ်, ငလျင်လုပ် , word etc are used as negative words as follow

- ငလျင်လှုပ်ခက်မှုကြောင့်. အီရန်မြို့.တော်တီဟီရန်တွင်ပြည်သူများလမ်းမများပေါ် ၇ ပြေးထွက်ခဲ့ရ
- လတ်တလောရင်ဆိုင်ကြုံတွေ့.နေရသည့် သဘာဝဘေးအန္တရာယ်ကျရောက်မှုများအပေါ် သက်ဆိုင်ရာပန်းကြီးဌာနများအလိုက်နိုင်ငံတကာအဖွဲ့.အစည်းများနှင့်. ဆက်သွယ်ပီးကယ်ဆယ်ရေး၊ ထောက်ပံ့ရေးနှင့်.ပြန်လည်ထူထောင်ရေး လုပ်ငန်းများကို ပေါင်းဆောင်ရွက်လျက်ရှိသည်။

4. . System Design

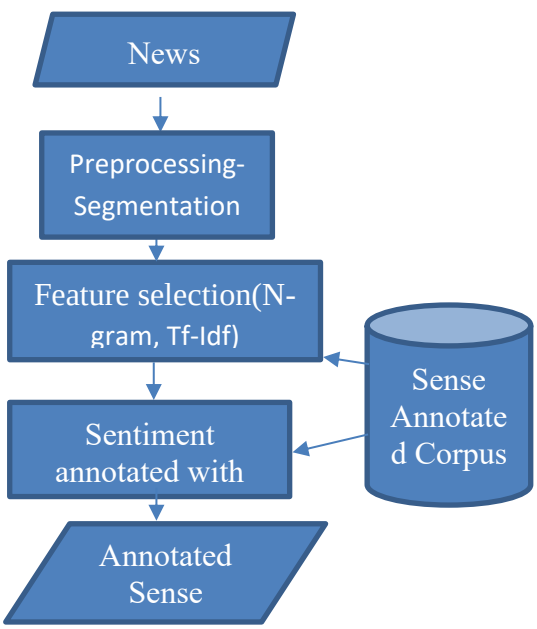


Figure 1. Proposed System Architecture

As shown in figure 1, news are need to preprocessing and segmentation process is done. And then, feature extraction process will proceed. N-gram feature selection is used. These features are then giving weighting value by using feature weighting algorithm, TF-IDF. These features are trained by support vector machine to get sentiment of news.

4.1 Preprocessing

Preprocessing analyzes the opinions from syntactical point of view and original syntax of sentence is not disturbed [10]. Preprocessing is important role in text classification to get more accuracy. Preprocessing is used to remove the noise of data to facilitate feature extraction. Tokenization and segmentation are first part of text classification.

4.2 Feature selection

Feature selection is the finding and selecting processes of more useful features in a dataset. Feature selection do the training process of the machine learning algorithm more faster and accuracy by eliminating vocabulary size and noise feature.

4.2.1 N-Gram

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. By using n gram, the system accuracy is higher. Examples of Myanmar negative word are shown in table1and positive word are shown in table 2:

Unigram	Bigram	Trigram
မီးလောင်	သား၊မယား၊ပြု	ဂင်၊တိုက်၊မှု
ငလျင်လုပ်	ဖမ်းဆီး၊စစ်ဆေး	ပိတ်၊မိ၊နေ
ရုပ်သိမ်း	အခက်၊အခဲ	စိတ်၊ကြွ၊ရွံးသွပ်
ပျက်ဆီး	ထပ်မံ၊ဖမ်းဆီး	ထပ်မံ၊ဖမ်းဆီး
စျေးကျ	အမှု၊ဖွင့.	သတင်း၊မှား၊ပေး

Table 1: N gram Negative words

Unigram	Bigram	Trigram
ဆောင်ရွက်နိုင်	ပို၊ကောင်းအောင်	ထိန်းသိမ်း၊စောင့်၊ရှောက်
နားလည်မှု	ပြန်လည်၊ရရှိ	ထုတ်လုပ်မှု၊မြင့်၊တက်
ဂမ်းမြောက်	အနည်းငယ်၊ကျဆင်း	ဆုတောင်း၊မေတ္တာ၊ပို.သ
အကျိုးပြု	ဆက်လက်၊ကျဆင်း	ပြန်၊လွတ်၊ပေးရေး
ထိမ်းသိမ်း	ဆုတောင်း၊မေတ္တာ	ပေးအပ်၊ချီးမြှင့.

Table 1: N gram positive words

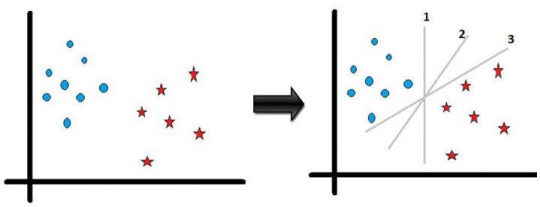
4.2.2 Tern Frequency-Inverse Document Frequency

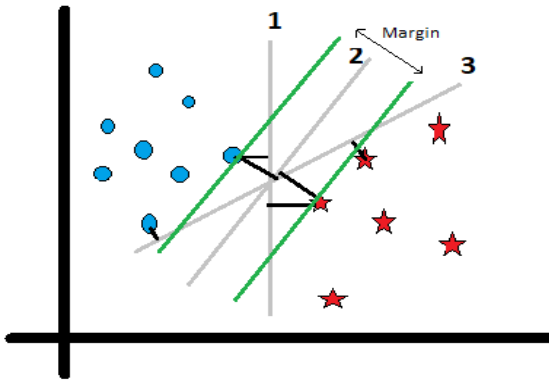
TF-IDF calculates a weight which states the importance of a term inside a document. It compares the frequency of usage inside an individual document as opposed to the entire data set. TF-IDF is based on the bag-of-words (BoW) model, therefore it does not get position in text, semantics, co-occurrences in different documents, etc.

- TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document)
- IDF(t) = log_e(Total number of documents / Number of documents with term t in it).
- Value = TF * IDF

4.3 Support Vector Machine

Support Vector Machine (SVM) is one of most useful supervised machine learning algorithms which can be applied for both classification or regression challenges. SVM are often defined as the classifier that makes the greatest accuracy outcomes in text classification issues. Each data item is placed as a point in n-dimensional space (n is number of features) with the value of each feature being the value of a particular coordinate. Classification is done by searching the hyper-plane that differentiate the two classes very well. SVM define the hyperplane that has the maximum distance (margin)between the nearest data points of either class.





SVM has four basic methods:

- The separating hyperplane
- The maximum-margin hyperplane
- The soft margin
- The kernel function

A linear classifier is based on a linear *discriminant function* of the form

$f(x) = w x + b.$

The hyperplane:

$\{x : f(x) = w x + b = 0\}$

$wx + b > 0$ for positive case

$wx + b < 0$ for negative case

X mean input feature vector, w means weight vector, b means bias.

5. Data used for Experiment

The experiment is implement using data collected from Myanmar news websites using Facebook 4G crawler. Currently, this contain 700 news. Each news contain average 5 sentences and totally 3500 sentence

6. Conclusion and Future Work

Opinions are central to almost all human activities and are key influencers of behaviors. Sentiment analysis is commonly used in several areas that include tracking sentiment towards products, movies, politicians, and companies. In Myanmar language, there is a little sentiment analysis system for many domains. In this system, Myanmar news are classified as positive and negative sentiments. Data sets are collected from various Myanmar web sites. Sense 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, support vector machine. More training data will give more performance, training data must be carried out. In future, we intend to test other classifier such as decision tree, decision list, maximum entropy, ensemble.

References

[1] Y AL-AMRANI, M LAZAAR, K E EL 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] V A. Kharde and S.S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[3] "Sentiment Analysis of Review of Restaurant in Myanmar Text", June 26-28, 2017, Kanazawa, Japan, 2017 IEEE SNPD 2017

[4] S M. Mohammad National Research Council Canada 1200 Montreal Rd., Ottawa, ON, Canada, “Challenges in Sentiment Analysis”, A Practical Guide to Sentiment Analysis 2015.

[5] A Gupta1, J Pruthi2, N Sahu3. ““Sentiment Analysis of Tweets using Machine Learning Approach”, International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 6, Issue. 4, April 2017, pg.444 – 458

[6] P Waykar, K Wadhwani, and P More, “Virtual Examples for Text Classification with Support Vector Machines “, Manabu Sassano Fujitsu Laboratories Ltd. 4-1-1, Kamikodanaka, Nakahara-ku, Kawasaki 211-8588, Japan

[7] M Z Asghar1, A Khan2, S Ahmad1, F Masud Kundi1” A Review of Feature Extraction in Sentiment Analysis”, Journal of Basic and Applied Scientific Research , ISSN 2090-4304

[8] M A Sghaier and M Zrigui“Sentiment Analysis for Arabic e-commerce websites”, June 2015.

[9] https://www.sciencedirect.com/science/article/pii/S2090447914000550#.

[10]. C. C. Aggarwal and C.-X. Zhai, Mining Text Data, Springer, 2012..