Annotation and Sentiment Analysis for Myanmar News

Thein Yu, Khin Thandar Nwet

University of Computer Studies, Yangon

{Thein Yu, khinthandarnwet}@ucsy.edu.mm

Abstract

This paper presents a corpus of Myanmar News and its metadata, annotated with sentiment polarity. It contains 100 news from Myanmar media websites such as news-eleven.com, 7daydaily.com and popularmyanmar.com. This news dataset is of at most value as all the annotation is done manually and this makes it a very rich dataset for training purposes. In this work, we describe the creation and annotation process, content, and the possible uses of the dataset. As an experiment, we have built a basic classification system to identify the emotion polarity of the news. N-gram feature selection algorithm is used in this system to select the most relevant features from training dataset. In Experiment, we classify the News by using well-known Naïve Bayes classifier.

Keywords: Sentiment Analysis, Opinion Mining, Natural Language Processing, Naïve Bayes, N-gram

1. Introduction

The development of Web 2.0 technologies 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].

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. For the training data set, news from Myanmar media websites are manually collected, labeled and stored in the training data set. N-Gram is used as a feature selection method and Naïve Bayes theory is applied in implementing the text classifier. In this Naïve Byes text classifier, word frequencies are used as features.

The remaining parts of this paper are organized as follows; the related works are explained in section 2. In section 3, labeling of training data is discussed. Then, the overview of the proposed system is shown in section 4 and experiment work is illustrated in section 5. Conclusion and future work is presented in section 6.

2. Related Work

Many algorithms have been applied for sentiment analysis system in English and different languages. However, very few researches have been done on Myanmar text.

In the paper written by BAgarwal, V K Sharma, and N Mittal, Point-wise Mutual Information (PMI) based method is used with extracted sentiment-rich phrases by using Partof-speech (POS) based rules and dependency relation in the document [5].

PranavWaykar, KailashWadhwani, Pooja More from Department of Computer Engineering, DYPIET, Pimprihas developed sentiment classification system using supervised approach. The unigram algorithm was applied to derive feature set from twitter dataset and Naïve Bayes classifier was then used on derived features for final categorization [6]. ShubhamTripathi presented Naïve Bayes classifier to classify movie data from the Internet Movie Database(IMDb)[3]. The algorithm is used for extracting bi gram feature. He applied methods using R Studio software. In paper [7], the authors implemented Ensemble classifier on twitter data sets. Bag of words approach is used by adding semantics on feature. To increase the accuracy of prediction,Word Sense Disambiguation and WordNet synsets are also added to the feature vector.

In paper [8], authors proposed the implementation of 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 developed a small (symbol to Word) converter for the use of emoticons and a checker for the use of the elongated words. They represented the reviews in six different models (unigrams, bigrams, tri-grams, unigrams + bigrams, bigrams + trigrams, unigrams + bigrams + trigrams) for feature extraction. They obtained 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 obtained the best results by the combinations: Unigrams + Bigrams, Unigrams + Bigrams + Trigrams with a preprocessed corpus, stemmed and classified by Support Vectors machine (SVM).

There are many approaches that have been applied 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[9].

3. Creation Training Dataset

Myanmar News data are extracted from news-eleven.com, the voice, 7daydaily.com and popularmyanmar.com (*weather, political, news*)…… for training and testing data. The collected corpora need to be segmented to mark the word boundary. Although the sentences of Myanmar are clearly delimited by a unique marker, but words are not always bordered by the spaces. Therefore, we have manually word-segmented the text from the web sites and change the font into Myanmar3 and character encoding into standard utf-8 format respectively. On the basis of the level defined, the task is to identify if positive or negative sentiment is expressed at that level. Following are the features of the dataset:

- originally procured 100 Myanmar News.
- 50 news annotated as positive in the dataset.
- 50 news annotated as negative in the dataset.
- total number of words in the dataset are 3000.
- average number of words in a news are rounded off to 30.
- total number of words in positive news are 1540.
- average number of words in a positive news can be rounded off to 34.
- total number of words in negative news are 1460.
- average number of words in a negative news can be rounded off to 26

3.1 Annotation

We aim to identify whether the news evokes a positive or a negative emotion for training data. Hence, it is best for us to look at document level classification.

Each Myanmar news was annotated as positive or negative. News evoke a certain emotion or mood, and these can be classified as those with positive or negative valence according to Russell's Circumplex Model, as shown in Figure 1. The news that evoke emotions ranging from 'aroused' to 'sleepy' including 'calm', 'satisfied', 'delighted', 'excited', happy, pleasant etc. are to be tagged as positive. Negative tags are to be given to songs evoking moods such as 'angry', ambiguous, 'annoyed', 'miserable', 'depressed', dirty, destroyed, etc., all of them spanning from 'alarmed' to 'sad' [14].

In Myanmar language, ကောင်းသော(good), သာယာသော(fine) and လှပသော (beautiful) used as positive word. ဆိုးသော, ရုပ်ဆိုးသော(ugly), မကောင်းသော(bad), ရေနစ်သည် (drown) and ပယ်ဖျက်သည်(annul) are used as negative words.

The number of positive and negative tags can be seen in Table 1.

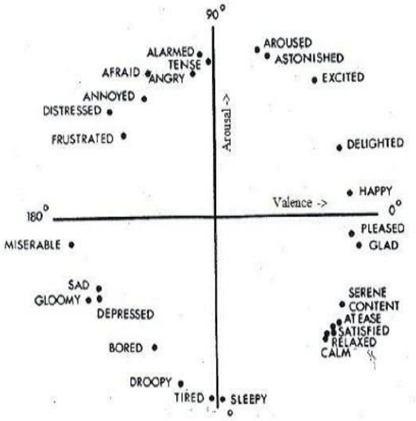

Figure 1. Russell's Circumplex Model [10] classifying 28 Affect words on the basis of positive and negative valence and arousal.

Table 1

Positive tags	Negative tags
50	50

In 100 annotated final news dataset, 50 are tagged as positive while the rest 50 are annotated as negative. This corpus can be helpful for Myanmar SentiWordNet.

4. Overview of the Proposed System

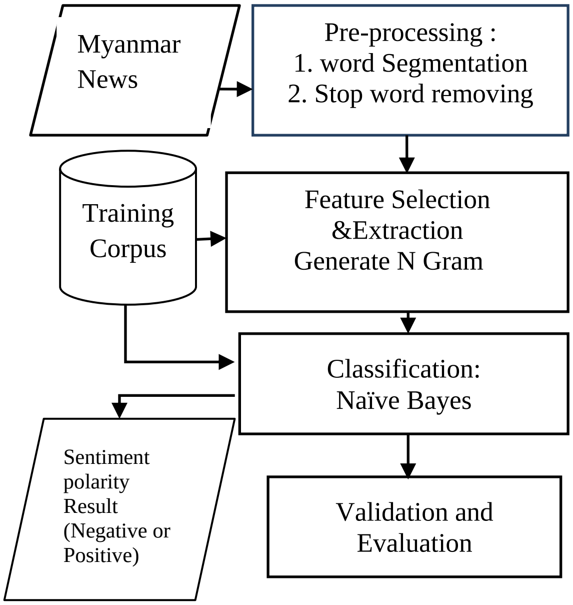

Figure 2 Proposed System Design

As shown in figure 2, there are two states in the proposed system, namely, training and testing state. In the training phase, preprocessing of collected raw data and feature selection are carried out. In the testing state, feature words from training corpus are extracted and later the classification process is done. In the proposed system, two sentiments such as positive and negative are defined.

Since the proposed system uses supervised learning approach, it needs to collect raw data to create training corpus. Therefore, news from Myanmar media websites such as news-eleven.com, 7daydaily.com and popularmyanmar.com and Duwun.com[11, 12], are manually collected for each category.

4.1 Preprocessing

Pre-processing analyzes the opinions from syntactical point of view and original syntax of sentence is not disturbed. Space delimited languages are easy to tokenize or segment words in the sentence. However, some language such Japanese, Chinese, Thai, India and Myanmar are not clearly delimited by a space. Thus, syllable segmentation is essentially needed as text preprocessing step. First, sentences in collected documents are segmented into words using maximum matching method. And then, punctuations and special characters are filtered from the collected documents. And then, stop words, punctuations and special characters are removed from the collected documents. Myanmar stop words are တွင်၊ မယ်၊ မည်၊ သည်၊ သည့်၊ များ etc.

4.2 Feature Selection

Feature selections techniques in sentiment analysis are become significant role for identifying relevant attributes and increasing classification accuracy. Feature selection is the process of selecting a subset of the words occurring in the training set and using only this subset as features in sentiment classification. The size of the vocabulary used in the experiment is selected by choosing words according to n-gram (bi gram and tri gram) method with respect to the sentiment. N-grams are very powerful and useful models and frequently the short-distance context is most important.

N-gram probabilities can be estimated by counting relative frequency on a training corpus. Table 2 shows the examples of bigram and trigram.

Table 2. Examples of Bigram and Trigram

Features	Examples
Bigram	မကောင်းသော
trigram	ညဘက်လေကြောင်းတိုက်ခိုက်မှု

4.3 Naïve Bayes Myanmar Text Classifier

Naïve Bayes method which is based on Bayesian theorem is used for analyzing 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. 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 document d for classification, the probability of each sentiment c is calculated as follows

$$P\left(\frac{c}{d}\right) = P(\frac{P(c)P(\frac{d}{c})}{P(c)} \qquad (1)$$

Where, P (d) plays no role in selecting c*. To estimate the term P(d|c), Naive Bayes decomposes it by assuming the fi's are conditionally independent given d's class as in Eq2.

$$P(c|d) = \frac{P(c)(\prod_{i=1}^{m} P\left(\frac{f_i}{c}\right)_k^{\;n_i(d)})}{P(d)} (2)$$

Where, m is the no of features and fi is the feature vector. Consider a training method consisting of a relative-frequency estimation P(c) and P (fi | c). Naïve Bayesian classifier that is simplify (naive) to assumpt 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[13].

5. Experimental Work

5.1 Data used for Experiment

The experiment is developed using data collected from Myanmar news websites. The proposed system is developed by using supervised learning. Currently, training set includes over 50 news and test set contains 30 news for each sentiment as shown in table 3. Both training and test data consist of Myanmar news which is composed of pure text data.

Table 3. Data Set

Data	Positive	Negative
No: Training Doc	50	50
No: Testing Doc	30	30

5.2 Performance Measures and Result

Table 4 results show that Naïve Bayes classifier works for this experiment. These experiments use N grams features. Precision can be looked at as a classifier's exactness while recall would give a measure of its completeness.

To measure performance, precision, recall and F-measure for each sentiment are calculated as the following equations.

$$Recall\ = \frac{\mathrm{TruePositivex}}{\mathrm{TruePositive+FalseNegative}} \quad (3)$$

$$\mathrm{Precision} = \frac{\mathrm{TruePosiitive}}{\mathrm{TruePositive + FalsePositive}} \quad (4)$$

$$\mathrm{F - measure(F1)} = \frac{\mathrm{2PR}}{\mathrm{P+R}} \quad (5)$$

Table 4. Precision, Recall and F-measure

Feature Extraction	CLASS	P(%)	R	F1
Bigram	Positive	86	52	64
	Negative	83	50	62
Trigram	Positive	90	54	67.5
	Negative	86	52	64

The fault of sentiment analysis process is caused by the amount of training corpus. In experiment, our results demonstrate that the comparison of bigram and trigram feature with Naïve Bayes. Trigram feature is more powerful than bigram.

6. Conclusion and Future work

Detection and classification of sentiments and opinion mining is a challenging task from the point of view of computational linguistics. This is because opinions are not realized uniformly in the syntax, semantics and the pragmatics of language. This system is designed to analyze sentiment of News data in Myanmar language. In this paper, training documents are collected from various news websites. And then, it classifies test documents. Since the proposed system uses supervised learning approach, it strongly relies on the size of training data. According to the result, it can work well only on well-defined sentiment To improve the performance of the proposed system, adding the more training data will be carried out. Hence, our proposed system of Myanmar news sentiment analysis can be seem to be useful and applicable for other research development in natural language processing. In the future, more and more training data must be trained. The accuracy will become higher. We have to test SVM, neural network and Random Forest for sentiment analysis system.

7. References

[1] J M Ruiz-Martínez1, R Valencia-García1, F García-Sánchez1, "Semantic-Based Sentiment analysis in financial news", Facultad de Informática. Universidad de Murcia.

[2] W N Chan[1] and T Thein[2,] "Sentiment Analysis for Twitter Stream Databy Combining Lexicon and Machine Learning Approaches", [1]*University of Computer Studies, Yangon and* [2]*University of Computer Studies, Maubin*

[3] 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

[4] B Liu, "Sentiment Analysis and Opinion Mining", April 22, 2012

[5] B Agarwal, V K Sharma, and N Mittal. ""Sentiment Classification of Review Documents using Phrase Patterns", International Conference on Advances in Computing, Communications and Informatics(ICCAI) 2013

[6] 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

[7] M Kanakaraj and R M R Guddeti, "NLP Based Sentiment Analysis on Twitter Data Using Ensemble Classifiers", 3rd International Conference on Signal Processing, Communication and Networking(ICSCN) 2015

[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] James A Russell. "A circumplex model of affect. *Journal of Personality and Social Psychology"*, 1980.

[11]www.7daydaily.com,

[12] www.news-eleven.com

[13] www.nlk.org

[14] G D Apoorva and R Mamid, "BolLy: Annotation of Sentiment Polarity in Bollywood Lyrics Dataset", Pacific Association for Computational Linguistics 2017.