ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse131862019.pdf
https://doi.org/10.30534/ijatcse/2019/131862019

Myanmar News Sentiment Analyzer using Support Vector Machine Algorithm

Thein Yu[1], Khin Thandar Nwet[2]
[1]University of Computer Studies, Yangon, Myanmar, theinyu@ucsy.edu.mm
[2]University of Information Technology, Yangon, Myanmar, khin.thandarnwet@gmail.com

ABSTRACT

Sentiment analysis is one of the natural language processing fields that combined computational linguistics and information retrieval. It is one of text classification techniques that extract opinion expressed in a positive and negative. The news, blogs and reviews obtained from social network are become important resources for sentiment analysis. This paper implements sentiment analyzer for Myanmar news. This system create sentiment annotated corpus for Myanmar news that are collected from Myanmar news web sites and ALT Treebank. Feature extraction and transformation are also needed to extract feature and transform to feature vector to get high performance. N-gram and TFIDF weighting methods are used in this system. Machine learning is combination of the techniques and basis from both statistics and computer science. This system is implemented by using machine learning algorithm Support Vector Machine compared with Logistics Regression and Naïve Bayes algorithms. We showed the comparison results of those algorithms and also showed that SVM is the more powerful than other two algorithms. User can easily know how importance about news by using this sentiment analyzer system.

Key words: Sentiment analysis, Machine Learning Algorithm, N-gram, TFIDF

1. INTRODUCTION

With the increasing growth of internet technologies in the recent years, the usage of social media has been growing. The sentiment analysis is a basically task to classify the given text into positive, negative and neutral category. Several researchers have been developing the automated techniques and algorithms for sentiment analysis and text classification. Sentiment analysis for Myanmar language has challenges due to scarcity of resources such as automatic feature extraction tools, stemming, anaphora resolution, and name entity recognition and so on. News is available from many resources and gives valuable information for user. News from Myanmar media websites and ALT data Treebank are collected, labeled and stored in the data set. N-Gram is defined as sequence of adjacent word that is used for feature selection method to increase performance of classification. TF-IDF is a feature weighting that determine important feature in the document. This paper proposed an automatic sentiment analysis system for Myanmar news. This system is implemented by using support vector machine, naïve bayes, and logistics regression learning approach.

Further structure of this paper is described as follows. Section 2 showed the related works. Section 3 defined methodology used for the system. Section 4 presented experiment of system. Finally, section 5 presented the conclusion of the paper.

2. RELATED WORKS

Many sentiment analysis systems are developed for English and other languages. Many systems use different resources using different approaches at different levels. In paper [2], authors developed sentiment analysis system for product reviews using SVM machine learning algorithm. They used different dataset for training and testing and showed better performance of SVM than others algorithms. In paper [3], authors proposed classification of polarity for twitter data using sentiment analysis. They evaluate two classifiers, linear and probabilistic for sentiment polarity categorization. They used tweets data. In paper [4], Y. M. Aye and S .S. Aung developed the Myanmar sentiment lexicon of 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. In paper [5], authors proposed sentiment analysis system using support vector machine. They used comments and tweets data. They used TFIDF method for transforming and feature selection. In paper [6], authors designed sentiment analysis system on big data. They used twitters reviews data and classified using Naïve Bayes and Logistics Regression algorithm on top of the Hadoop with Mahout. Work controller module added to automate the experiment.

3. SYSTEM FRAMEWORK AND DESIGN

In this system, preprocessing, feature extraction and transformation, and classification stages are done. System framework diagram is shown in Figure 1.

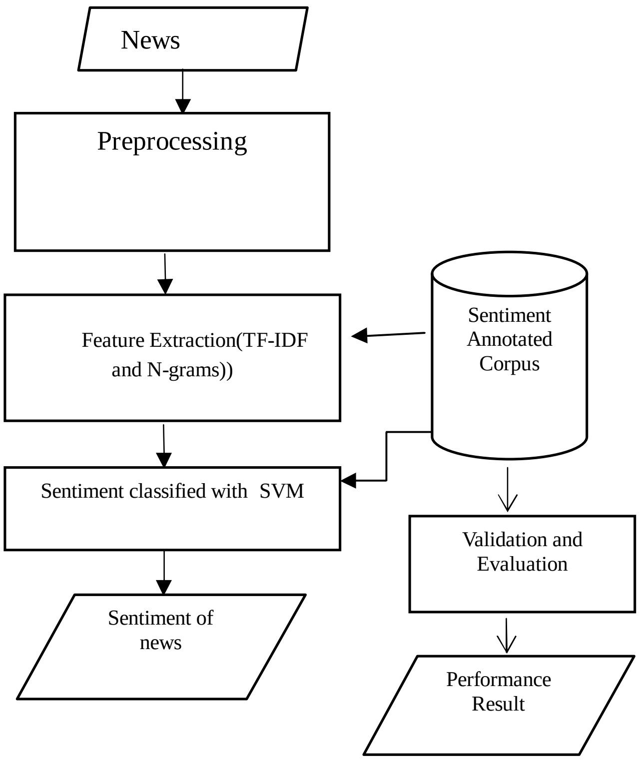

Figure 1: System Design

3.1 Preprocessing

Preprocessing is the basic step for natural language processing process and has many step. Word segmentation, tokenization, and stop words removing processes are used in the proposed system [7].

A. Word Segmentation

Myanmar word segmentation is placing spaces into textual data without other replacing or rewriting operations [8].

B. Tokenization

Tokenization is the separation of words in a sentence by locating word boundary. Tokenization is a step after segmentation process that breaks a segment into tokens to search for a match for the segment.

C. Stop Word Removal

Stop word removal is an important preprocessing technique in classification tasks so as to improve the performance of system. Stop words are most usually used words in any natural language which have very little or no significant semantic context in a sentence. Myanmar stop words are , etc.

3.2 Feature Extraction

N-gram consists of adjacent words or letters of length n. If length n is 1, it is a unigram. Then if n is 2, it refers to a bigram. And then n is 3, it also refers to a trigram. We used ngram_range features from sklearn.

Sentence 1:

Unigram:
V=[“ ”, “ ”, “ ”, “ ”, “ ”]

Bi-gram :
V=[“ ”, “ ”, “ ”, “ ”]

Tri-gram:
V=[“ ”, “ ”, “ ”]

Combination of unigram and bi-gram (unigram+bigram):
V=[“ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”]

Combination of Unigram, bi-gram and tri-gram (Unigram +Bi-gram+Tri-gram):
V=[“ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”, “ ”]

3.3 Feature Transformation

A. TF-IDF Vectorizer

TF-IDF depends on the basic of the bag-of-words (BoW) model. TF-IDF calculates weight values of term which define the importance of a term inside a document. TF-IDF can be mostly used for stop-words removing in various

subject fields including text classification and summarization. Calculation function is in (1), (2), (3).

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document) (1)

IDF (t) = loge (Total number of documents / Number of documents with term t in it). (2)

Value = TF * IDF (3)

Examples of following sentences of TFIDF values are as shown in table1.

s1**:**

s2**:**

V= [' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ']

Table 1: TFIDF Matrix Values

Features vectors	TFIDF
(0, 6)	0.37997836159100784
(0, 10)	0.534046329052269
(0, 0)	0.534046329052269
(0, 4)	0.534046329052269
(0, 4)	0.47107781233161794
(1, 6)	0.24395572500006343
(1, 1)	0.3428712594108198
(1, 5)	0.3428712594108198
(1, 8)	0.3428712594108198
(1, 7)	0.3428712594108198
(1, 11)	0.3428712594108198
(1, 9)	0.3428712594108198
(1, 2)	0.3428712594108198
(1, 3)	0.3428712594108198

B. CountVectorizer

Countvectorizer is the simplest way to tokenize text and build a vocabulary of words. It also encodes text with vocabulary and count number as in table 2.

Table 2: Count Values

Feature vector	Count values
(0, 4)	1
(0, 0)	1
(0, 10)	1
(0, 6)	1
(1, 3)	1
(1, 2)	1
(1, 9)	1
(1, 11)	1
(1, 7)	1
(1, 8)	1
(1, 5)	1
(1, 1)	1
(1, 6)	1

3.4 Classification

The classification process finds the actual mapping between features and labels (or targets). Machine learning algorithm is responsible for the automated detection of meaningful text in data. Supervised Machine Learning (SML) is the algorithms that used labeled sample data to produce general hypotheses, which then make predictions about future sample instances. In this system, support vector machine, naïve bayes, and logistics regression algorithm are applied in the proposed system.

A. Support Vector Machine

Support Vector Machine (SVM) is one of most widely used supervised machine learning algorithms that can be used for both classification and regression challenges. SVM is often defined as the classifier that makes the higher accuracy outcomes in text classification issues. Each data item is placed as a point in n-dimensional space with particular coordinate feature value. Classification is processed by searching the hyper plane that classified the two classes very well. SVM consider the hyper plane that has the maximum distance margin valuesas shown in Figure 2.

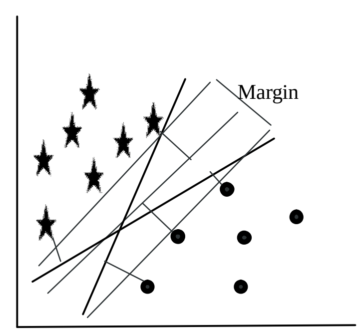

Figure 2: Maximum Margin Hyper Plane

Following equations are used to classify class in support vector machine. The decision hyper plane is in (4):

$f(\mathrm{x}) = \mathrm{w}\ \mathrm{x} + b = 0$ (4)

For positive case, (5) is used.

$f(\mathrm{x}) = \mathrm{w}\ \mathrm{x} + b > 0$ (5)

For negative case, (6) is used.

$$f(x) = w\,x + b < 0 \quad (6)$$

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

B. Logistics Regression

Logistics Regression is one of the most useful machine learning algorithms for binary classification. It's an extension of the linear regression model for classification problems. The name logistic regression is applied when the dependent variable has only two values, such as one thing or another. Logistic regression compare with discriminant analysis as a method for determining categorical-response variables. Logistic regression does not assume that the independent variables are normally distributed. Logistics regression is study of the relationship between a dependent variable (label) and a set of independent variables (feature) by calculating probabilities using its underlying logistic function [9]. These probabilities must then be transformed into binary values for prediction. The Sigmoid-Function is an S-shaped curve that can take real-valued number and map it into a value between the range of 0 and 1. This value will then be changed into either 0 or 1 using a threshold classifier. If there are many related features, logistic regression will assign a more accurate probability than naive Bayes. To calculate probability of class, (7) and (8) are used. Logistics regression process model is shown in Figure (3).

$$\ln\left(\frac{P(Y \mid X)}{1 - P(Y \mid X)}\right) = \beta_o + \beta_1 X \quad (7)$$

$$P(Y \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}} \quad (8)$$

Where

X is predicted variable (feature)

β is coefficient

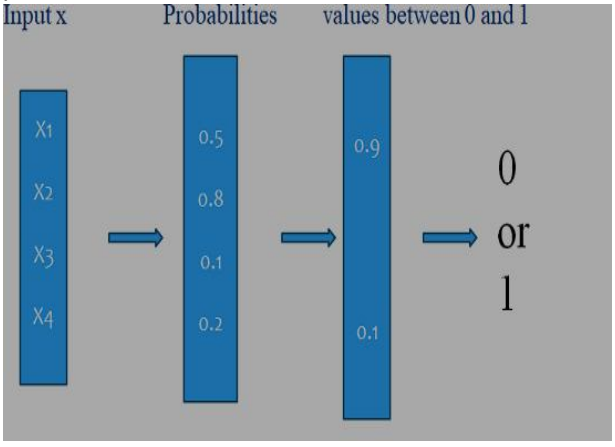

Figure 3: Logistics Regression Model

C. Naïve Bayes

Naïve Bayes classifiers are one of simple probabilistic classifiers based on Bayes' theorem that features are independent each other**.** A Naïve Bayes classifier considers each feature to contribute independently to the probability regardless of any possible correlations between different features. Naïve Bayes is a models classifiers that assign class labels to next problem that is composed of features vector, where the class labels are drawn from predefined set. It assigns to class of feature instance probabilities as in (9):

$$p\left(\frac{C_k}{x}\right) = \frac{p(C_k)\,p\left(\frac{x}{C_k}\right)}{p(x)} \quad (9)$$

Denominator of feature vector not depends on class C, so it is a constant value. Numerator is equal to joint probability and probability of class can be calculated in (10)-(13)

$$p(Ck/x1, \ldots\ldots, xn) \quad (10)$$

$$p(Ck/x1, \ldots\ldots xn) \propto p(Ck, x1, \ldots\ldots, xn) \quad (11)$$

$$\propto p(Ck)\,p(x1/Ck)\,p(x2/Ck)\,p(x3/Ck) \quad (12)$$

$$\propto p(Ck) \prod_{i=1}^{n} p\left(\frac{x1}{Ck}\right) \quad (13)$$

The final decision function of class is in (14):

$$\hat{y} = \text{argmax}_{k \in \{1, \ldots\ldots k\}}\, p(c) \prod_{i=1}^{n} p(xi \backslash Ck) \quad (14)$$

X_n means features vectors, C_k means class labeled

4. EXPERIMENT

4.1 Experimental Apparatus

This system used word segmentation tool from Natural Language Processing Lab, University of Computer Studies, Yangon. This system is implemented by using Jupyter notebook environment with python programming language. Train and test data with SVM, Logistics Regression, and Naïve Bayes machine learning algorithm.

4.2 Data

Data are collected from many web sites and ALT tree bank. This system use 2000 news data set. This system contains 200 news dataset for positive news and 1000 news dataset for negative news as shown in table 3. The 80 % of data is used for training and 20 % of data is used for testing. Unbalance data set are used because support vector machine be able to well classify unbalanced data. Training and testing data are randomly chosen by system.

Table 3: Data Set

News types	Amount
Positive	2000
Negative	1000
Total	3000

4.3 Experimental Results

In this system, we experiment SVM, Naïve Bayes, and Logistics Regression algorithm with many features such as unigram, bigram, trigram, unigram +bigram (combination of unigram and bigram), bigram+trigram(combination of bigram and trigram), and unigram +bigram +trigram (combination of unigram, bigram and trigram). We used these five features to test performance of algorithms with many features. SVM with TFIDF (unigram+bigram) feature has highest accuracy score value. Hold out evaluation method is used. Experimental results are shown in table 4-8.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (15)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (16)$$

$$\text{Recall}= =\frac{TP}{TP+FN} \quad (17)$$

$$\text{F-Measure}=\frac{2*Precision*Recall}{Precision+Recall} \quad (18)$$

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 4: Accuracy Score with CountVectorizer

Feature with CountVectorizer	Naïve Bayes	SVM	Logistics Regression
	Accuracy %	Accuracy %	Accuracy %
Unigram	82.33	82.00	84.00
Bigram	77.00	75.00	77.33
Trigram	69.50	69.00	68.83
Unigram + Bigram	81.17	82.50	82.17
Bigram+Trigram	74.17	73.83	72.33
Unigram + Bigram+ Trigram	79.83	81.17	81.33

Table 5: Accuracy Score with TFIDF

Feature with TFIDF	Naïve Bayes	SVM	Logistics Regression
	Accuracy %	Accuracy %	Accuracy %
Unigram	75.67	84.00	82.83
Bigram	69.33	74.67	68.50
Trigram	67.83	68.33	69.17
Unigram + Bigram	71.83	84.17	76.50
Bigram +Trigram	68.50	72.00	69.17
Unigram + Bigram+ Trigram	69.50	83.00	72.16

Table 6: Performance Results with SVM and TFIDF (unigram+bigram) features

	Precision %	**Recall %**	**F1-score %**
Positive	88.00	89.00	89.00
Negative	0.75	0.74	0.74

Table 7: Performance Results with Naïve Bayes and TFIDF (unigram+bigram) features

	Precision %	**Recall %**	**F1-score %**
Positive	72.00	97.00	83.00
Negative	70.00	16.00	26.00

Table 8: Performance Results with Logistics Regression and TFIDF (unigram+bigram) features

	Precision	**Recall**	**F1-score**
Positive	77.00	95.00	85.00
Negative	75.00	36.00	49.00

5. CONCLUSION

Sentiment analysis is supposed as one of the popular research topics in the domain of knowledge discovery. can be applied to analyze the useful information. This system is implemented by using SVM and compares the performance of Naïve Bayes and Logistics Regression algorithm. Support vector machine with TFIDF (unigram+bigram) features is more powerful than other two algorithms. Other machine learning algorithm such as ensemble, neural network, decision tree and maximum entropy, convolutionary neural network are attempted to implement in future. Currently, Myanmar news data set are trained and tested. In future, we will attempt to train and test Myanmar text.

ACKNOWLEDGEMENT

I specially thank to my supervisor, Dr. Khin Thandar Nwet, Lecturer, University of Information Technology, Myanmar. I also deeply thank to our rector , beloved teachers, and colleagues from Natural Language Processing Lab, University of Computer Studies, Yangon for word segmentation tool and everything.

REFERENCES

1. D.Jurafsky and J.H. Martin. **N-Grams**, **Speech and Language Processing**, 2014, ch. 4.
2. E. Tyagi and A. K. Sharma. **Sentiment Analysis of Product Reviews using Support Vector Machine Learning Algorithm**, Indian Journal of Science and Technology, Vol 10(35), September 2017. https://doi.org/10.17485/ijst/2017/v10i35/118965
3. A. S .Raghuwanshi and S. K. Pawar. **Polarity Classification of Twitter Data using Sentiment Analysis**, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 5 Issue: 6, June 2017, pp. 434 – 439.
4. Y.M. Aye and S. S.Aung. **Sentiment Analysis of Review of Restaurant in Myanmar Text**, IEEE SNPD June 26-28, 2017, Kanazawa, Japan, 2017 , pp.321-326.
5. A. Z. H. Khan, M. Atique, and V. M. Thakare, **Sentiment Analysis Using Support Vector Machine**, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, April 2015, pp.105-108.
6. A. Prabhat. **Sentiment classification on Big Data using Naïve Bayes and Logistics Regression**, International Conference on Computer Communication and Informatics, 5-7 Jan. 2017 https://doi.org/10.1109/ICCCI.2017.8117734
7. M. K. Saad. **The Impact of Text Preprocessing and Term Weighting on Arabic Text Classification**, M.S. thesis, Computer Engineering Department, The Islamic University - Gaza , 2010.
8. W. P. Pa, N. Thein. **Myanmar Word Segmentation Using a Combined Model**, e-Case 2009, January 2009.
9. www.machinelearningblog.com/tutorial
10. http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning
11. J. Khairnar and M. Kinikar. **Machine Learning Algorithms for Opinion Mining and Sentiment Classification**, International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013, ISSN 2250-3153.