Innovations in Computing Technology and Applications
Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0

Effects of Papaya Extract on Dengue Based on Sentiment Analysis in Text

Amni Abdul Muhen, Naomie Salim

[1]Information System Department, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia

[1]amnimuin@gmail.com, [2]naomie@utm.my

Abstract

The purpose of this research is to design sentiment analysis framework to identify whether the consumption of papaya leaves affect patients with dengue using sentiment analysis. Prior to this research, many have used the traditional method of consuming papaya leaf extract to counteract dengue, yet limited work has been done on whether the consumption of papaya leaf bear positive, negative, or no effect to dengue patients. There are several previous studies which have used sentiment analysis methods in determining polarity in online forums or product and services reviews, such as K-Nearest Neighbor (K-NN), Neural Networks (NN), Naïve Bayes (NB), Decision Tree (DT) and the Support Vector Machine (SVM). This project attempts to extract the sentiments from patients suffering from dengue who specifically consumes papaya leaf or papaya extract using SVM and NB algorithm. These algorithms are used to classify the sentiments extracted from reviews whether a positive, negative or neutral review. The corpus collected in this project contain 13 journals from Medline and 27 reviews from MedicineNet which is stored into CSV file. Then, synonyms and alternative terms are identified using WordNet. Data preparation and pre-processing are completed within the corpus to be ready for text feature extraction methods involving Bag of Words model and Term-Frequency Inverse Data Frequency (TF-IDF matrix). NB and SBM classification is performed to the dataset. The model was trained and applied on test data, and evaluated by calculating and comparing accuracy, precision and recall parameters. The results show that SVM outperforms NB classifier in terms of accuracy. R language is used for the implementation of this research.

Keywords: Sentiment Analysis; Text Classification; Naïve Bayes (NB); Support Vector Machine (SVM); Bag of Words; Term-Frequency Inverse Data Frequency (TF-IDF)

1 Introduction

The papaya plant has been widely-used as the treatment for several diseases. Extracts from the leaves, fruit and seeds have various beneficial effects. There are scientific studies on treating dengue using papaya leaf extracts, and as a result, increase the rate of platelets (Subenthiran, et al., 2013; Yunita, Hanani, & Kristianto, 2012) and increase the thrombocyte count (Siddique, Sundus, & Ibrahim, 2014). Since there is no definitive treatment or vaccines for the disease currently, it is important to find out if the traditional practice of using Carica papaya leaves could be used to treat dengue effectively. One of the effective ways is to determine the sentiments from patients who suffered dengue or have relatives or friends recovering from dengue who specifically used papaya extract to treat the symptoms.

The purpose of this project is to design sentiment analysis framework to identify sentiments on whether the consumption of papaya leaves affect patients with dengue. Thus, it is imperative to investigate text feature selection and text feature extraction from online reviews through sentiment analysis. In addition to that, the effectiveness and accuracy of the Naïve Bayes and Support Vector Machines classifiers have to be evaluated in classifying texts and analyzing sentiments.

2 Related Work

Machine learning algorithms, have two types which are known as supervised and unsupervised machine learning algorithms (Borele & Borikar, 2016), helps in classifying and predicting whether a document have positive or negative sentiment. In supervised learning algorithm, a labeled dataset defined as document in training set labeled with positive or negative sentiment, whereas, unsupervised learning is vice versa, whereby unlabeled dataset is texts which are not labeled with appropriate sentiments.

Furthermore, two sets of documents are required in a machine learning based classification, which is the training and the test set. A training set is used to learn the distinguished characteristics of documents, and a test set is used to validate the accuracy of the performance of the classifier.

There are three levels implemented in sentiment analysis: phrase level, document level and aspect level. Document Level classification considers a single topic or document and classifies the sentiments. Phrase level classification on the other hand, refers to a combination of two or more words and classifies the polarity of individual sentence of the document, whereas aspect level classification determines the polarity of each document by taking into consideration several aspects and components of a corpus. This research is focused primarily on applying supervised learning approach on a labeled dataset.

It is noted that several studies have carried out their research in sentiment analysis particularly for entertainment reviews, like reviews on movie ratings or tweets from Twitter, or reviews which provide useful information such as product reviews or services reviews from Amazon. Most of the approaches used for document level sentiment analysis focused on supervised learning due to its strong predictive power; the Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy algorithm are the preferred techniques by most researches in text classification analysis (Kaur & Balakrishnan, 2016). Hybrid approach are also popular in achieving higher accuracy in several of the methods.

In terms of application, the Naïve Bayes and the SVM are the two most frequently used classifiers in analyzing sentiments in documents, websites or databases. Furthermore, most of the reviews proved that both classifiers also generate comparatively high prediction accuracy compared to other classification algorithms, and if these methods are compared

against each other in terms of their performance measures, more of then not their accuracy and F-1 values are similar to each other.

Past research on DT and NN classification methods are limited in the application of sentiment analysis in TM. Neural Networks are not an efficient model to represent structural and contextual properties of the sentence, and their performance is close to the baseline Naive Bayes Algorithm (Shirani-Mehr, 2015). However, they are robust to noisy data.

Based on the previous literature, many researchers had compared the results of using K-Nearest Neighbor, Neural Networks, Decision Tree, Naïve Bayes and Support Vector Machines classification methods for analyzing sentiments. Previous studies on sentiment analysis proved that each of these models have their corresponding strengths and limitations. Most papers concluded that the use of the Naïve Bayes and SVM model produced better results than the other classification methods and are more commonly used in ML approaches. Hence, this study will apply the Naïve Bayes and SVM algorithm method to predict the effects of the papaya extract on dengue outbreak.

3 Methodology

To carry out this research, there are eight main phases of the framework with each of the phases consist of their own processes in order to complete this research effectively. Figure 1 illustrates the proposed research framework.

Text feature extraction is one of the key steps in sentiment analysis or opinion mining. The two methods used are machine-learning based using Bag of Words tokenization and manually annotated opinion lexica using TF-IDF matrix.

The process of turning a collection of text documents into numerical feature vectors is called the Bag of Words representation. Documents are described by word frequencies in a representation of text data while ignoring the relative position information of the words in the document (Bergen, 2014). In this Bag of Words model, individual words are segmented from sentences, tokenized and normalized, and given a specific subjectivity score (Taspinar, 2015). If the total score is positive the text will be classified as positive and likewise if the total score is negative.

The other technique applied is by using a pre-compiled list of words with positive and negative meaning which have been manually constructed (Hu & Liu, 2004). This list consisted around 6800 English positive and negative opinion words or sentiment words compiled over several years. Then, TF-IDF matrix, which stands for "term frequency / inverse document frequency" is combined with the sentiment polarity according to the sentiment lists to measure the relative importance of each word to the corpus.

TF-IDF stands for "term frequency / inverse document frequency" and is a popular weighting method in the field of Natural Language Processing (NLP). TF-IDF score helps in balancing the weight between most frequently words and less commonly used words. TF-IDF value also shows the relative importance of a token to a text sentence in the corpus. TF-IDF can be calculated as (Trstenjaka, et al., 2014):

$$\text{tf-idf}(t,d)=\text{tf}(t,d)\times\text{idf}(t)$$

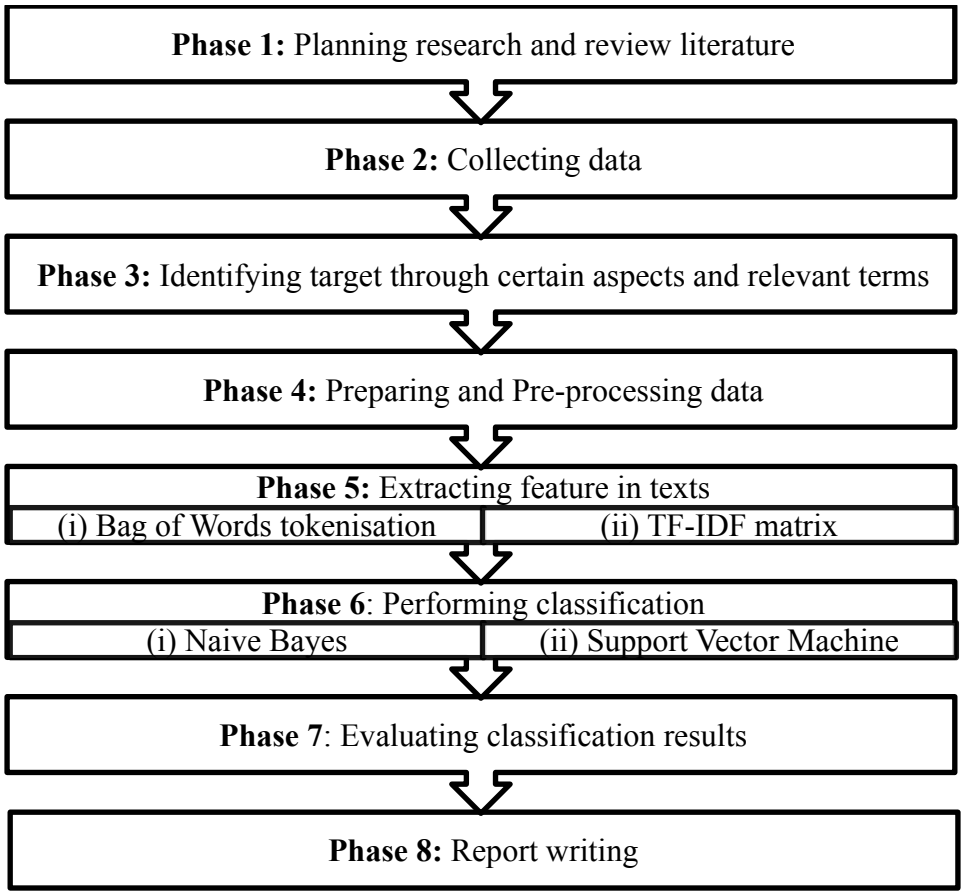

Figure 1: Research Framework

4 Experimental Results

This analysis is divided into two of validation experimentations towards the testing dataset for both Naïve Bayes and SVM classifiers, whereby dataset is randomized again to obtain variations for the performance values. In the first and second analysis, function set.seed() in R is set to 142 and 10 respectively.

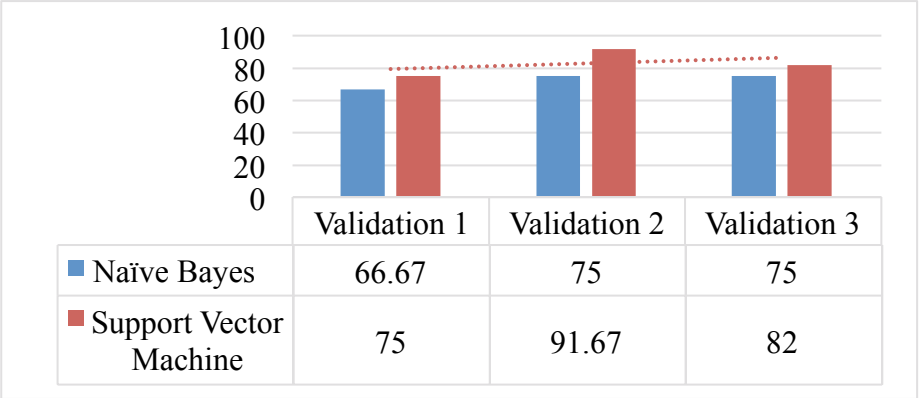

Figure 2: Overall accuracy for Naive Bayes and SVM classifier in both validation sets

Both classifiers achieve higher accuracy in the second validation set, whereby the data has been randomized in the 70:30 partition. SVM classifier however, scored highest accuracy of 91.67% out of the 6 sets. Naïve Bayes also has an increment in its accuracy level, but not as drastic as SVM – from 66.67% to 75%. However, both classifiers' accuracy levels decreased in the third validation set, which is partitioned into 80:20 ratio.

5 Discussion

Both Naïve Bayes and SVM has often been used in studies throughout the years for only two classes, which is Positive and Negative classes (Tripathya, et al., 2015; Borele & Borikar, 2016; D. & Gore, 2016). There are very few studies which employ multiple classes for sentiment analysis, and one of the few who did so is by Zhanga, et al. (2014). But even so, Zhanga, et al. (2014) do not utilize the TF-IDF feature weighting as one the steps in the framework, but rather the research was completed using the N-gram language model. Both of the studies mentioned employed Naïve Bayes and SVM classifiers to analyze sentiments.

Therefore, in order to support the results and findings of this research, the output obtained will be compared with other studies in sentiment analysis which employed Naïve Bayes and SVM methods. Table 1 shows the comparison of obtained results from other studies which use Naïve Bayes and SVM.

	(Tripathya, Agrawal, & Rath, 2015)	(Alm, Roth, & Sproat, 2005)	Output in this Study
Naïve Bayes	0.789	0.895	0.75
SVM	**0.815**	**0.940**	**0.916**

Table 1: Comparison between output obtained from this study with existing studies

Table 1 shows that SVM method obtained higher accuracy in analyzing sentiments. (Tripathya, et al., 2015) integrated feature weighting method of TF-ID matrix

as well and (Alm, et al., 2005) has multiple classes in the dataset. Thus, it can be concluded that SVM is a better approach to analyzing sentiments in texts.

6 Conclusion

In summary, two validation sets have been carried out in this chapter to measure the performance, mainly the accuracy of both Naïve Bayes and SVM methods. It is found that SVM achieved higher accuracy than Naïve Bayes does in both validation sets, thus, it can be concluded that SVM method is better at analyzing sentiment in texts in terms of accuracy.

For future work, enlarging the corpus to contain hundreds of reviews from people suffering from dengue disease and employing a medical expert to annotate and evaluate the medical terms in the reviews.

Furthermore, the framework employed in this research could also be expanded to cater to different diseases and their respective traditional treatments. Not only that, instead of focusing on only the description of the effects from treatment, future work could focus on specific preparation of Carica papaya extract that is most effective in treating dengue disease. Specific preparation in this context means the definite amount of dosage and frequency of the treatment to be consumed in a day.

There is still a need to conduct more studies, observations and investigations to gain a thorough understanding of the uses of papaya and its effectiveness in treating dengue virus.

References

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *Association for Computational Linguistics*, 579-586.

Bergen, K. (2014). Text Mining and Classification. *Stanford University*.

Borele, P., & Borikar, D. A. (2016). An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques. *IOSR Journal of Computer Engineering (IOSR-JCE) , 18*(2), 64-69.

D., B. S., & Gore, P. (2016). Sentiment Analysis on Twitter Data Using Support Vector Machine. *International Journal of Computer Science Trends and Technology (IJCST) , 4*(3), 365-370.

Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. (2016). Sentiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier .

Dhande, L. L., & Patnaik, D. P. (2014). Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3*(4), 313-320.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data . *Journal of Big Data, 2*(5), 1-14.

Hu, M., & Liu, B. (2004). Mining and Summarizing Customer Reviews. *Discovery and Data Mining*.

Kaur, B., & Kumari, N. (2016). SVM and KNN based Hybrid Approach to Sentiment Analysis. *International Journal of Technical Research & Science, 1*(5), 67-74.

Kaur, W., & Balakrishnan, V. (2016). Sentiment Analysis Technique: A Look into Support Vector Machine and Naive Bayes. *Proceedings of 2016 International Conference on IT, Mechanical & Communication Engineering*, 82-87.

Mahyoub, F. H., Siddiqui, M. A., & Dahab, M. Y. (2014). Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning. *Journal of King Saud University –*, 417-424.

Moraes, R., Valiati, J. F., & Neto, W. P. (2013). Document-level sentiment classification: an empirical comparison between SVM and ANN. *Expert Systems with Applications: An International Journal, 40*(2), 621-633.

Preety, & Dahiya, S. (2015). Sentiment Analysis Using SVM and Naive Bayes algorithm. *International Journal of Computer Science and Mobile Computing, 4*(9), 212-219.

Shirani-Mehr, H. (2015). Applications of Deep Learning to Sentiment Analysis of Movie Reviews.

Siddique, O., Sundus, A., & Ibrahim, M. (2014). *Effects of papaya leaves on thrombocyte counts in dengue — a case report.* University of Health Sciences, Karachi.

Taspinar, A. (2015, November 16). *Text Classification and Sentiment Analysis*. Retrieved from Ataspinar: https://ataspinar.com/2015/11/16/text-classification-and-sentiment-analysis/#SL_literature

Tripathya, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science, 57*, 821–829.

Trstenjaka, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF Based Framework for Text Categorization. *Procedia Engineering, 69*, 1356-1364.

Vadivukarassi, M., Puviarasan, N., & Aruna, P. (2017). Sentimental Analysis of Tweets Using Naive Bayes Algorithm. *World Applied Sciences Journal, 35*(1), 54-59.

Zhanga, L., Hua, K., Wang, H., Qian, G., & Zhang, L. (2014). Sentiment Analysis on Reviews of Mobile Users . *Procedia Computer Science, 34*, 458-465.