ISBN 978-93-84422-51-6
Proceedings of 2016 International Conference on IT, Mechanical & Communication Engineering (ICIMCE 2016)
Pattaya, January 2-3, 2016

Sentiment Analysis Technique: A Look into Support Vector Machine and Naïve Bayes

Wandeep Kaur, Vimala Balakrishnan

Department of Information Systems, Faculty of Computer Science and Information Technology
University of Malaya, Malaysia

Abstract: *Sentiment Analysis and opinion mining aims to analyze sentiments, opinions, emotions etc. towards products, services or current topics. There are various approaches applied to mine the sentiments portrayed. Supervised machine learning is one such approach that is generally applied. The aim of this paper is to investigate the current methods used to perform sentiment analysis by reviewing and comparing recently published research. The findings are discussed in hope that it would help future researchers to gain an understanding of a possible method they could adopt or even come up with a new approach to better mine sentiments from big data that is tailored to suit the need of their data source.*

Keywords: *sentiment analysis, Naïve Bayes, support vector machine, supervised machine learning*

1. Introduction

The boom of social networking system has resulted in a vast amount of textual data availability in the last few years. Realizing the opportunity that comes with studying large data, the attention has now diverted from data storage and retrieval to refining methodologies to extract and process information from raw sources [7]. Sentiment analysis (SA) is a combination of data mining and Natural Language Processing (NLP) techniques in order to computationally treat subjectivity in textual documents. It is considered a challenging NLP problem specifically for Twitter and transcribed text [9]. The primary focus of textual information retrieval technique is to mine fact from opinions. Facts consist of an objective component expressed within subjective characters. These subjective elements comprise of opinions, sentiments and emotions, which are the core of SA [18]. SA aims to classify texts as positive, negative or neutral at different levels: document-level sentiment analysis, sentence-level sentiment analysis, aspect-based sentiment analysis, comparative sentiment analysis and, sentiment lexicon acquisition [6]. Nevertheless, SA is commonly employed on three levels: sentence level, document level and aspect level [20].

Studies on SA generally focussed on supervised learning (presumes there is a defined set of class into which a document should be categorized and training data is available for each class) or unsupervised (establishes the semantic orientation of distinct phrases within a document) learning method. Naïve Bayes and Support Vector Machine aims to examine the manner in which these two techniques have been used by studies focussing on SA. This was accomplished by reviewing some of the related articles, as will be explained in the following sections. This paper is organized as follows: Section 2 discusses the literature studied for this paper. Section 3 gives a brief description of the two SA techniques mentioned above; Section 4 explains the methodology while Section 5 provides the results and discussions. Finally Section 6 presents the conclusion of the research.

2. Literature Review

Sentiment classification can approximately be divided into two key areas: lexical based approach and machine learning approach [3]. Predefined dictionaries of terms annotated with positive or negative scores are essentially used for the lexical method [14]. If lexicon matches a word marked positive in the dictionary, then the total polarity score of a text increases. A text in whole will be classified as positive given that the general polarity score is positive, else it is categorized as negative.

The machine learning approach uses a series of chosen feature vectors and a collection of tagged corpora to prepare a model which will then be used to categorize untagged corpus of text [11]. The feature selection in this approach is vital to ensure the classification success rate. A wide range of unigrams (i.e. single words from a document) or n-grams (i.e. two or more words from a document in sequential order) are essentially chosen as feature vectors. The primary asset of machine learning approach lies within the capability of its algorithm to analyse text of any sphere and produce classification models that are customized to the issue at hand [17]. Furthermore, they are not only language independent and can be successfully applied to multiple languages but can also be adapted to incorporate additional information in their decision process [9]. Machine learning approach can be further sub divided into two main classes: supervised and unsupervised methods (see Fig. 1). However, [7] discovered that most of the approaches used for document level sentiment analysis focused on supervised learning due to its strong predictive power. In particular, the Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy algorithm are the preferred techniques [3]. Lastly, the hybrid approach marries the above discussed approaches in order to achieve higher accuracy as sentiment lexicons play an imperative role in majority of the methods[1],[5],[7],[15].

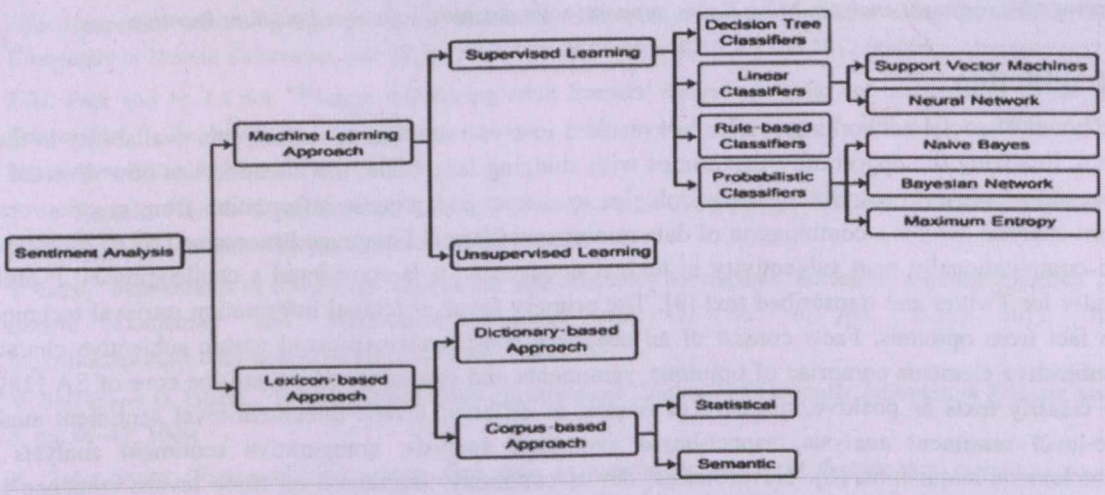

Fig 1: Sentiment classification techniques (Medhat et. al, 2014)

3. Supervised Machine Learning

A supervised machine learning approach hinges on the existence of labelled training documents. With regard to the literature study done, there are various kinds of supervised classifiers. In the following subsection, a concise description of the two most endorsed classifiers in SA will be presented [3] namely Naïve Bayes and Support Vector Machine.

3.1. Naïve Bayes

The NB is the most commonly used probabilistic classifier which utilizes the properties of Bayes theorem. The added advantage of this classifier is its need for only a small amount of training data to calculate its prediction parameters. The NB classifier works with the string of words feature extraction which is not dependent on the position of the word in the document. Therefore only a variance of a feature is computed

instead of the complete covariance matrix. Based on Bayes theorem, the probability for each class review can be calculated using the equation (1)

$$P(c \mid d) = \frac{P(d \mid c) * P(c)}{P(d)} \quad (1)$$

(Tripathy, Agrawal & Rath, 2015)

Where P(d|c) = prior probability of a label

P(c) = prior probability that given feature set is being classified as a label

P(d) = prior probability that a given feature set is occurred.

3.2. Support Vector Machine (SVM)

The SVM is a contemporary machine learning approach that is dependent on statistical learning concepts. Fig 2 shows the basic working model of SVM. An SVM model is an illustration of the examples as points in space, depicted such that the members of the independent categories are divided by a void as wide as possible [17]. New examples are then chartered into that same space and forecast to belong to one of the categories based on which side of the gap they fall in. Defining it more academically, the SVM has the ability to establish a nonlinear decision plane within the native feature by mapping data instances non-linearly to inner product space where the classes can be uncoupled directly with a hyperplane [16].

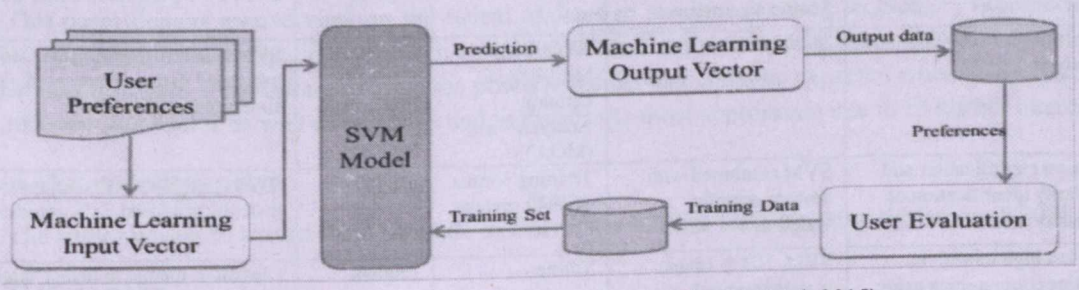

Fig 2: Basic working model of SVM (Kothari & Patel, 2015)

4. Methodology

A compilation list of articles was published previously covering publications up to 2012 [16]. Therefore, the current study focussed on scholarly articles published between 2013 and 2015. These articles listed in the tables were found through Google Scholar and Science Direct; searched using keywords such as "sentiment analysis", "multilingual sentiment analysis", "sentiment classification", "sentiment polarity", "Naïve Bayes", "support vector machine", "machine learning algorithms", "supervised machine learning" and "classifier ensembles". A basic filtering was utilized whereby only those research that used NB and SVM as their base techniques were analysed resulting in these seventeen articles. It needs to be noted that the papers listed in this tables are merely a fraction of the research done on SA and opinion mining. The articles and results of our review are presented next.

5. Results and Discussion

Results of analysis on the seventeen papers are depicted in Table I and Table II. With almost 6000 tweets tweeted every single second; Twitter is one of the largest data repositories available. SA tools such as AlchemyAPI, Lymbix and Repustate make it easier to extract and process large volume of data. Therefore, Twitter is the most favoured data source to analyse when it comes to SA. In spite of that, recent research has expanded to include TripAdvisor ([9],[11]), online movie reviews ([8],[9],[15],[20]) as well as medical disorder data ([17]). Although the dominant language seems to be English, interest to analyse sentiments from other

languages such as German, Spanish, Chinese, French, Arabic and Kannada is catching on ([2],[12],[14],[15],[21]).

NB algorithm yields a better score compared to SVM ([12],[14],[21]) for a non-English corpus. Articles [8] and [11] were examined using two sets of data; cleaned data (noise removal, hashtag removal and abbreviation clean-up) and uncleaned data. The performance accuracy produced significantly high results for when data was cleaned up. Table I and Table II also shows that a hybrid approach (combination of more than one SA approach) produced better results compared to when only one machine learning technique was used ([1],[5],[7],[15]). From the table, it can be seen that although there is an increased interest in SA of corpus whose domain language is other than English, yet a gap still persists. This could be due to the lack of a lexicon source that is non English.

The SA approach chosen is dependent on the data source and purpose of the study. If the study is related to content that requires some form of human translation and language lexicons, then NB produces better results. This could be related to the abundance of lexicons created within the database. The SVM method is statistical learning dependent hence it produces higher accuracy and precision when calculating polarity. An improved accuracy reading is recorded when the asset of both these approaches are manipulated. Hence SA method applied is dependent on the data source. If the study relates to polarity using data source consisting of non-English corpus that may require some form of translation, then using the hybrid approach may produce better results.

TABLE I: Article Summary

Ref	Objective	Technique Applied	Data Source	Language	Findings
[1]	Semantic analysis of movie review	SVM & Particle Swarm Optimization	Twitter	English	SVM-PSO performed better than SVM alone
[2]	Multilingual semantic analysis	SVM	NTCIR 8 Multilingual Opinion Analysis Task (MOAT)	English, German, French and Spanish	SVM results using English data set and translated dataset showcased almost similar results
[3]	Aspect classification and polarity identification of product review	SVM combined with domain specific lexicons	Training corpus of 1940 reviews	English	SVM combined with domain specific lexicons produced 78% accuracy
[4]	Detect high tension in online communities using computational analysis	SVM, NB & Linear logistic regression	Twitter	English	Classifier unable to detect "high tension" but able to detect some form of tension using NB & SVM
[5]	Classify tweet sentiment using ensemble classifier & lexicons	Multinomial NB, SVM, random forest & logistic regression	Twitter	English	Classifier ensemble formed by diversified compnents provide exceptional results Feature hashing results good for tweet SA
[7]	Ensemble learning to reduce noise sensitivity related to language ambiguity	Bayesian ensemble learning	Twitter	English	Ensemble technique discussed is efficient & effective
[8]	Investigate role of pre-processing data	SVM & chi square method	Online movie reviews	English	Pre-processing data / data clean up can significantly increase accuracy
[9]	Context based approach for SA	ConSent - novel approach	Twitter, TripAdvisor & Internet Movie Database	English	Performs well in data that has some context in it. Twitter results not convincing
[11]	Suggested use of standard SVM in dealing with contextual and non contexual data	SVM	TripAdvisor	English	Integration of SVM with pre-processed data produced results of higher accuracy
[12]	SA for Kannada web documents	SVM, NB	Kannada review text corpus	Kannada, English	Text corpus was translated NB performed best
[13]	SA of products and services	NB	Twitter	English	NB accuracy recorded at 90.31%

TABLE II: Article Summary

Ref	Objective	Technique Applied	Data Source	Language	Findings
[14]	Arabic sentiment lexicon tested using semi supervised method	SVM, NB	Arabic WordNet	English, Arabic	NB accuracy scored higher than SVM. Accuracy performance recorded at 97% using the Arabic lexicons
[15]	Meta classifier to develop polarity classification system	SVM, NB, Bayesian Logistic Regression, C4.5	Spanish corpus of film movie	English, Spanish	Ensemble technique improves polarity classification
[17]	Propose advanced Multi Class Instance Selection based SVM to increase efficiency of SVM	Improved SVM	Medical disorder dataset from UCI repository	English	Proposed method showed an increase in classification accuracy, ratio of selected instances and time consumption
[19]	Investigate big data reduction technique	SVM	Twitter	English	Framework proposed to cope with the problem of reducing size and dimension in big data supervised learning settings
[20]	SA movie reviews using machine learning technique	NB, SVM	Movie dataset	English	SVM produced higher accuracy results compared to NB
[21]	SA of application reviews from mobile users	NB, SVM	WeChat, iTunes	English, Chinese	Bayesian produced better results compared to SVM

6. Conclusion

This paper depicts an overview on the recent updates in machine learning techniques specifically NB and SVM. Seventeen recently published articles were read and summarized. It is hoped that from this work, researchers may gain some information on the possible results that could be expected when analysing sentiments for different data source as well as language and to decide the most appropriate one to fit his/her interest.

7. Acknowledgement

The author wishes to extend their gratitude to University Malaya for supporting this study (RP028A-14AET)

8. References

[1] Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 453-462.

[2] Balahur, A., & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis.Computer Speech & Language, 28(1), 56-75.

[3] Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment Analysis: Measuring Opinions. Procedia Computer Science, 45, 808-814.

[4] Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J. & Sloan, L. (2013). Detecting tension in online communities with computational Twitter analysis. Technological Forecasting and Social Change.

[5] da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. Decision Support Systems, 66, 170-179.

[6] Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

[7] Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian Ensemble Learning. Decision Support Systems, 68, 26-38.

[8] Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. Procedia Computer Science, 17, 26-32.

[9] Katz, G., Ofek, N., & Shapira, B. (2015). ConSent: Context-based sentiment analysis. Knowledge-Based Systems, (84), 162-178.

[10] Khan, K., Baharudin, B., Khan, A., & Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. Journal of King Saud University-Computer and Information Sciences, 26(3), 258-275.

[11] Kothari, A. A., & Patel, W. D. (2015). A Novel Approach Towards Context Based Recommendations Using Support Vector Machine Methodology.Procedia Computer Science, 57, 1171-1178.

[12] Kumar, K. A., Rajasimha, N., Reddy, M., Rajanarayana, A., & Nadgir, K. (2015). Analysis of users' Sentiments from Kannada Web Documents.Procedia Computer Science, 54, 247-256.

[13] Luke, J. (2015). Data Mining of Automatically Promotion Tweet for Products and Services Using Naïve Bayes Algorithm to Increase Twitter Engagement Followers atPT. Bobobobo. Procedia Computer Science, 59, 254-261.

[14] Mahyoub, F. H., Siddiqui, M. A., & Dahab, M. Y. (2014). Building an Arabic Sentiment Lexicon Using Semi-Supervised Learning. Journal of King Saud University-Computer and Information Sciences, 26(4), 417-424.

[15] Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M., & Ureña-López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications, 40(10), 3934-3942.

[16] Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

[17] Ramesh, B., & Sathiaseelan, J. G. R. (2015). An Advanced Multi Class Instance Selection based Support Vector Machine for Text Classification.Procedia Computer Science, 57, 1124-1130.

[18] Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences, 311, 18-38.

[19] Silva, C., Antunes, M., Costa, J., & Ribeiro, B. (2015). Active Manifold Learning with Twitter Big Data. Procedia Computer Science, 53, 208-215.

[20] Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science,57, 821-829.

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