International Journal of Pure and Applied Mathematics
Volume 118 No. 19 2018, 71-83
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

Sentiment Analysis of Product using Machine Learning Technique: A Comparison among NB, SVM and MaxEnt

Monali Bordoloi[1] and S.K. Biswas[2]
[1]Ph. D. Candidate,
Dept. of Computer Science & Engineering,
National Institute of Technology, Silchar,
Assam, India. PIN: 788010.
monali.bordoloi@gmail.com
[2]Assistant Professor,
Dept. of Computer Science & Engineering,
National Institute of Technology, Silchar,
Assam, India. PIN: 788010.
bissarojkum@yahoo.com

February 1, 2018

Abstract

Efficient sentiment analysis is of high demand in many application domains for accurate and predictive classification. Despite the wide use and popularity of some methods, a better technique for identifying the polarity of a text data is hard to find. Machine learning has recently attracted attention as an approach for sentiment analysis. Therefore this paper proposes a machine learning model for sentiment analysis and compares some popular machine learning approaches (Naive Bayes, Support Vector Machine and Maximum Entropy) in the context of sentiment classification.

1

The performances of the classifiers are measured and compared with three datasets. The performance of the classifiers is measured in terms of accuracy. From the experimental results it is observed that Naive Bayes classifier has produced better performance than Support Vector Machine and Maximum Entropy classifiers. However, all the classifiers show significant performance.

Key Words : Sentiment Analysis, Machine Learning, Text Mining, Classification, Naive Bayes, Support Vector Machine, Maximum Entropy.

1.1. Introduction

Sentiment analysis focuses on sentiment of a sentence which is expressed as positive or negative. Sentiment analysis is a process of classifying the sentiment of sentences/reviews towards a subject matter such as product, person etc. [1]. Sentiment analysis is also termed as "opinion mining" and opinions are very important in decision making because whenever someone needs to make a decision, he or she wants to know others' opinions about the same. For business, businessmen and organizations want to investigate consumer opinions about their products so that products and services can be made better for the customers. Even before buying any product dilemma is faced whether the product is good or not. Though the shopping sites have comments and star ranking for each product which can help to judge whether the product is suitable or not. But going through all those reviews is again a hectic task, as for some products it may range up to ten thousand of reviews. Not only the numbers of online users have grown but also there is an increase in the companies who provide products. In both cases, sentiment analysis becomes a powerful tool to find a reasonable solution.

The application of sentiment analysis is not confined to the business process models only, rather they are observed in many domains such as politics, security, disaster management etc. Sentiment analysis from social networking sites has recently drawn a lot of attention because of their many useful applications. There are many techniques available for sentiment analysis like machine learning tools- Nave Bayes (NB), Support Vector Machines (SVM), Maximum Entropy (MaxEnt), etc. or sentiment analysis by keyword extraction followed by polarity assignment by using Sentiment Dictionaries. Machine learning has attracted little attention as an approach for sentiment analysis. Therefore this paper proposes a machine learning model for sentiment analysis and compares some popular machine learning methods (NB, SVM and MaxEnt) in the context of sentiment classification. Hence the main contribution of this work is a comparison of performance among dominant and computationally competent approaches (NB, SVM and MaxEnt) under the same context.

1.2. Literature Survey

The growth of social networking has increased the scope of expression of opinion or sentiment on a public platform. Simultaneously the increasing interest of people to acquire more and more details about different insights and information regarding any topic available online, boost the need to generate efficient sentiment analysis techniques. The literature provides some important machine learning model for analysis of sentiment.

The pioneering work by Gaspar et al. [2] established that the social media reactions is not confined to the concept of only positive and negative or good and bad, but can interpret qualitative sentiment based on the event subjected to which the reactions are given. Boiy et al. [3] used a cascaded approach of three different machine learning techniques namely SVM, Multinomial Nave Bayes (MNB) and MaxEnt to find the sentiment associated with multilingual text representing different blog, review and forum texts using unigram feature vectors. Xia et al. [4] proposed a research model that uses POS tagging and different word relation feature sets namely unigrams, bigrams and dependency parsing pairs; to perform sentiment classification by three different ensemble methods and three different classification algorithms. Sharma et al. [5] presented a commendable comparison of different existing machine learning techniques namely NB, SVM, MaxEnt, Decision Tree(DT), K-Nearest Neighbour (KNN), Winnow and Adaptive Boosting (Adaboost) in association with different feature selection methods that can be used affectively in sentiment analysis of online movie reviews while demonstrating the superiority of SVM classifier and NB. Tan et al. [6] proposed a model for sentiment analysis of different domains namely education, movie and house, written in Chinese using different feature selection and machine learning techniques. Sidorov et al. [7] showed the superiority of unigrams along with other optimal settings like low number of classes, effectiveness of balanced and unbalanced corpus, use of best machine learning classifier etc. while performing opinion mining of text. Using different machine learning techniques; Ji et al. [8] proposed a pioneering work based on Measure of Concern (MOC) to measure public concerns via twitter data while making use of the most important unigrams. Their model proved very fruitful when the machine learning method was combined with clue-based method. Li et al. [9] used MaxEnt, SVM and NB classifiers to draw comparative results for sentiment analysis of Chinese documents using Chinese Character based Bigram (CBB), Trigram (CBT), Word-Based Unigram (WBU), Bigram (WBB) feature representations; and feature weighting schemes and feature dimensionality. Tripathy et al. [10] also reported the use of unigrams in the count vectorizer matrix along with the notion of Term Frequency-Inverse Document frequency to perform sentiment analysis of movie review using SVM and NB classifiers. Ye et al. [11] provided a great contribution to the tourism department by performing sentiment analysis of reviews regarding travel destinations with the use of SVM, NB and character based N-gram model. Dhanalakshmi et al. [12] presented the implementation of Rapid Miner (open source data analytic tool) for opinion mining of student feedback data in comparison with SVM, NB, KNN and Artificial Neural Network (ANN) classifiers. Rana et al. [13] demonstrated a model which focuses on sentiment analysis of movie reviews in order to draw a comparison between the two of the popular machine learning tools namely SVM and NB. Yan et al. [14] used MaxEnt method to perform the sentiment analysis of Tibetan sentence by using the probability difference between positive and negative. Using reviews from Amazon, Vinodhini et al. [15] proposed an integrated model which uses neural network (NN) and principal component analysis (PCA), in order to establish the superiority of neural network based models in comparison to other machine learning techniques used in the area of sentiment analysis.

Korenek and imko [16] made use of appraisal theory to determine sentiments of text data from microblogs using hybrid patterns in a deeper level rather than the simple concept of polarity assignment as different situations produce different reactions on different person. The commendable work by Che et al. [17] used Conditional Random Fields (CRF) model along with four different categories of special features to compress sentiment sentences while determining the aspect level sentiments of different product domains.

1.3. Research Methodology

The flow chart of the model is presented in fig. 1.1. The model comprises of four main stages:

1. Data Collection
2. Preprocessing
3. Feature vector extraction and
4. Classification

The raw data are collected from online shopping sites and are preprocessed to remove useless words. After preprocessing the remaining review is observed as a vector comprising features and those features are used to train the classifier for sentiment classification. All the stages are explained in details below:

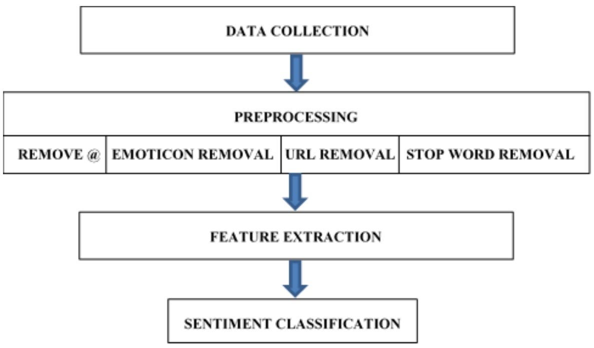

Fig. 1.1. Flowchart of the proposed model

1.3.1. Data Collection

The raw data i.e., reviews, are collected along with the ratings from Amazon and Flipkart. Before processing of those data, the raw data collected are thus assigned a class i.e., whether positive or negative, to represent as patterns. Products having rating 3 and above are assigned class 1 (indicating positive) while others are assigned class 0 (indicating negative).

1.3.2. Pre-processing

Each and every word and character is not always important for analyzing the sentiments. The patterns must first be pre-processed in order to get the optimal set of data comprising of only useful information for the processing. The preprocessing includes the following:

- Removal of stop words: The stop words such as about, almost, be, etc. that do not provide any sentiment in the review are removed.
- URL Removal: The associated URLs of different forms like *www. or http. or https. or abc.def.com or abc.def.in or abc.def.net or abc@def.net or abd.de@fgh.co.uk* are also removed from the original reviews.
- Emoticons removal: The different emoticons like ☺, :p , ☹,:* etc. are removed from the original reviews.
- Removal of @: @ is also removed from the reviews as it does not carry any meaning for sentiment analysis.

1.3.3. Feature Vector Extraction

After preprocessing, patterns do not have any stop words, URLs, emoticons and @. The patterns are now tokenized one by one and stored in an array; one array for one pattern. These arrays are observed as vectors and words in each vector are observed as features. Using the unigram feature extraction method, the final dataset to be used for classification is to be formed, where each unique feature is an attribute. Now, based on the occurrence or nonexistence of the features in a feature vector of the old patterns, 1 or 0 is assigned to each attribute for each pattern of the dataset to form the final dataset to be used in the classification. The dataset to be used is of the form as shown in table 1.1.

FEATURES	excellent	phone	os	nice	camera	working	worst	CLASS
P1	1	1	1	1	1	0	0	1
P2	0	1	0	0	1	0	1	0

Table 1.1. Final dataset to be used for classification

1.3.4. Classification

The classifier employed needs to be trained with the training data and then the trained classifier can be used for decision making processes. Here the patterns as shown in table 1.1 are given to the classifiers (NB, SVM and MaxEnt) to be trained. The classifiers are trained with positive and negative vectors and can be used for sentiment classification with new feature vector.

1.4. Results and Discussion

Three datasets of product reviews are collected along with the ratings from online shopping sites like amazon.in and flipkart.com. For each dataset 1500 unprocessed reviews are collected. The products taken into consideration are Redmi Note 1, Moto G4 and Lenovo K3 mobile phones. The description of the data used is shown in table 1.2. The reviews are manually assigned class using the ratings as can be seen from table 1.3. Reviews are distributed into training and testing sets and experiment is performed using Python 3. 90% of the dataset is used for training and 10% for the testing.

Dataset	Size	Feature Type	Number of Class
Lenovo K3	1500	Mixed	2
Redmi Note 1	1500	Mixed	2
IPHONE 5	1750	Mixed	2

Table 1.2. Dataset description

Review	Rating	Class
Redmi Note 1 is a beast	****	1
By sending tuns of email i wasted my time. Please guys do not buy or suggest this Mi products. Feeling Angry with my friend who suggest me to buy this!!!	*	0

Table 1.3. Reviews with rating and class

1.4.1. Performance Measure

The performance is measured in terms of overall accuracy of the model. Confusion matrix is used for calculating the accuracy of the classifier as it contains information regarding classifications that are predicted by a classifier and the ones that are actual values. Table 1.4 presents the confusion matrix with the following data records for a two class classifier:

(a) True positive (TP): Total "positive" reviews predicted as "positive".

(b) False positive (FP): Total "negative" reviews predicted as "positive".

(c) False negative (FN): Total "positive" reviews predicted as "negative".

(d) True negative (TN): Total "negative" instances predicted as "negative".

The accuracy is the proportion of the total number of predictions that are correct and is determined using the equation (1). [18]

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{1}$$

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table 1.4. Confusion Matrix

Table 1.5. depicts the accuracy of the three classifiers for all the datasets used. From the experimental results, it is observed that

NB classifier outperforms SVM and MaxEnt in all the datasets. If the performance of the SVM and MaxEnt is compared, MaxEnt is better than SVM because MaxEnt performs better that SVM in two datasets whereas SVM performs better than MaxEnt in one dataset. However, it can be seen from the table 1.5. that all the machine learning classifiers produce more than 70% accuracy in all the datasets. Hence it can be urged that machine learning approaches can be one of the good alternatives for sentiment classification.

Graphical representation of all the experimental results with three classifiers is shown in fig. 1.2 for better understanding and clarity. X axis of the graph shows dataset and Y axis accuracy in percentage.

Datasets	Naive Bayes	SVM	Maximum Entropy
Lenovo	81.33	78.67	72
Redmi Note 1	80	72.85	74.28
IPHONE 5	78.24	73.53	76.47

Table 1.5. Accuracy table for the different datasets

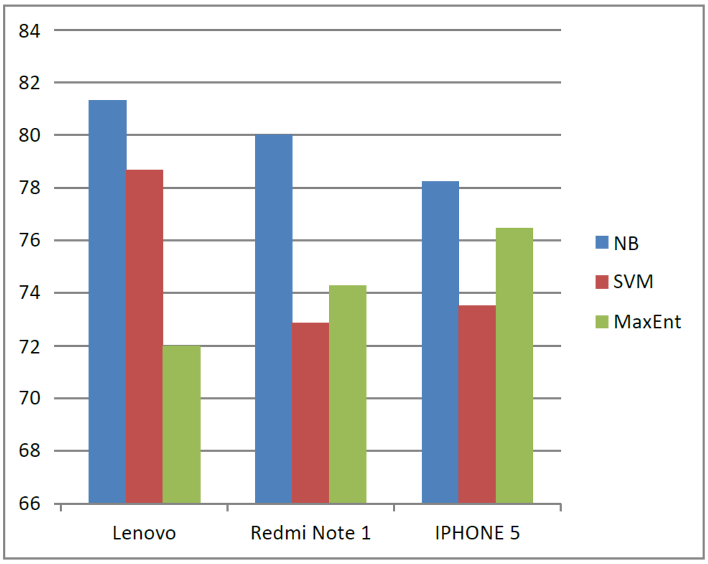

Fig. 1.2. Graphical representation of comparison of the results with different classifiers

9

1.5. Conclusions

This paper presents a machine leaning sentiment analysis model with three classifiers and compares the performance of the three classifiers. The model is divided into four stages: data collection, preprocessing, feature vector extraction and classification. Performance of the classifiers is measured in three datasets and it is observed that NB classifier has produced better performance than SVM and MaxEnt classifiers. However performance of the all the classifier is significant and hence it can be concluded that machine leaning approach has great potential for sentiment classification. Finding of this experiment would, therefore, make a meaningful contribution to understand people's perception of different products. It is believed that a larger dataset will perform better in sentiment classifications for all three algorithms. The work can be augmented by removing Internet slang and by including other machine learning classifiers.

References

[1] Wawre, S. V., & Deshmukh, S. N., *Sentimental Analysis of Movie Review using Machine Learning Algorithm with Tuned Hypeparameter*, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 6, pp. 12395-12402, (2016).

[2] Gaspar, R., Pedro, C., Panagiotopoulos P., & Seibt B.., *Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events*. Computers in Human Behavior, Elsevier, vol. 56, pp. 179-191, (2016).

[3] Boiy, E., & Moens, M. F., *A machine learning approach to sentiment analysis in multilingual Web texts*, Information Retrieval, Springer, vol. 12, pp. 526558, (2009)

[4] Xia, R., Zong, C., & Li, S., *Ensemble of feature sets and classification algorithms for sentiment classification*, Information Sciences, Elsevier, vol. 181, pp. 1138 1152, (2011).

[5] Sharma, A., & Dey, S., *A comparative study of feature selection and machine learning techniques for sentiment analysis*, RACS'12, October 23-26, 2012, San Antonio, TX, USA. ACM, 978-1-4503-1492, pp. 1-7, (2012).

[6] Tan, S. & Zhang J.., *An empirical study of sentiment analysis for Chinese documents*, Expert Systems with Applications, Elsevier, vol. 34, pp. 26222629, (2008).

[7] Sidorov, G., Miranda-Jimnez, S., Viveros-Jimnez, F., Gelbukh, A., Castro- Snchez, N., Velsquez, F., Daz-Rangel, I., Gordon J., *Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets*. I. Batyrshin and M. Gonzlez Mendoza (Eds.): MICAI 2012, Part I, LNAI 7629, pp. 114, Springer, Springer-Verlag Berlin Heidelberg, (2013).

[8] Ji, X., Chun, S.A., Wei, Z., & Geller, J., *Twitter sentiment classification for measuring public health concerns*, Soc. Netw. Anal. Min, 5:13, DOI 10.1007/s13278-015-0253-5, Springer, (2015).

[9] Li, J., & Sun, M., *Experimental Study on Sentiment Classification of Chinese Review using Machine Learning Techniques*, International Conference on Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007, DOI: 10.1109/NLPKE.2007.4368061, IEEE. (2007).

[10] Tripathy, A., Agrawal, A., & Rath, S.K., *Classification of Sentimental Reviews Using Machine Learning Techniques*, Procedia Computer Scienc. 3rd International Conference on Recent Trends in Computing 2015, Elsevier, vol. 57, pp. 821 829, (2015).

[11] Ye, Q., Ziqiong Zhang, Z., & Law, R., *Sentiment classification of online reviews to travel destinations by supervised machine learning approaches*, Expert Systems with Applications. Elsevier, vol. 36, pp. 65276535, (2009).

[12] Dhanalakshmi, V., Dhivya, B., & Saravanan, A. M., *Opinion mining from student feedback data using supervised learning algorithms*, 3rd MEC International Conference on Big Data

and Smart City, DOI: 10.1109/ICBDSC.2016.7460390, IEEE. (2016).

[13] Rana, S., & Singh, A., *Comparative Analysis of Sentiment Orientation Using SVM and Nave Bayes Techniques*, 2nd International Conference on Next Generation Computing Technologies, IEEE, (2016).

[14] Yan, X., & Huang, T., *Tibetan sentence sentiment analysis based on the maximum entropy model.*, 10th International Conference on Broadband and Wireless Computing, Communication and Applications, IEEE, (2015).

[15] Vinodhini, G. & Chandrasekaran, R.M., *A comparative performance evaluation of neural network based approach for sentiment classification of online reviews*, Journal of King Saud University, Computer and Information Sciences , Elsevier, vol. 28, pp. 212, (2016).

[16] Korenek , P. & imko, M., *Sentiment analysis on microblog utilizing appraisal theory*, World Wide Web, Springer, vol. 17, pp. 847867, (2014).

[17] Che, W., Zhao, Y., Guo, H., Su, Z. & Liu, T., *Sentence Compression for Aspect-Based Sentiment Analysis*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, IEEE, vol. 23, No. 12. (2015).

[18] Biswas, S. K., Bordoloi, M., Heinsnam, R.S., Purkayastha, B.: *A Neuro-Fuzzy Rule-Based Classifier Using Important Features and Top Linguistic Features.* International Journal of Intelligent Information Technologies, vol. 12, issue 3, July-September, IGI Global, (2016)