Comparative Analysis of Supervised Learning for Sentiment Classification

Afusat O. Muyili(✉) and Oladipupo A. Sennaike

Department of Computer Sciences, University of Lagos, Lagos, Nigeria
muyiliafusat@gmail.com, osennaike@unilag.edu.ng

Abstract. Sentiment analysis is an active research area which deals with information extraction and knowledge discovery from text using Natural Language Processing and Data Mining techniques. Sentiment analysis, also known as opinion mining, plays a major role in detection of customer's attitude, response and opinion towards a product or service. The aim of this paper is to perform sentiment analysis on a particular service to discover how users perceive the service automatically. Data is extracted from twitter, pre-processed and classified according to the sentiment expressed in them: positive, negative or neutral using five supervised learning classifiers-The Naïve Bayes, Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Linear Support Vector Machine (SVM) and Decision Tree classifiers. Finally, the performance of all the classifiers is compared with respect to their accuracy. In addition, the results from the classifiers show that supervised learning classifiers perform excellently in sentiment classification.

Keywords: Sentiment analysis · Twitter · Feature extraction · Naïve Bayes
Decision Tree · Linear Support Vector Machine · Multinomial Naïve Bayes
Bernoulli Naïve Bayes

1 Introduction

Social media have completely changed the way people communicate and have become a large part of daily lives in most societies. This development has led to the creation of huge amounts of data which is useful for analysis of users' opinion such as evaluating a written or spoken language to determine if the expression is favourable, unfavourable or neutral and to what degree.

Over the years, business organizations have experienced exponential growth in the use of online resources, particularly social media and microblogging websites like Facebook, Twitter, Tumbler, YouTube, etc. Such organizations highly depend on these resources as a rich mine of marketing knowledge unlike the conventional methods (interviews, questionnaires and survey) which are highly expensive and time consuming in gaining insight and feedbacks into how customers perceive their products or services due to poor design and environmental factor. Hence, there is need of a system that can automatically generate users' opinions (sentiment analysis) from huge amount of data.

V. Odumuyiwa et al. (Eds.): AFRICOMM 2017, LNICST 250, pp. 272–278, 2018.
https://doi.org/10.1007/978-3-319-98827-6_26

Sentiment Analysis (SA) is the process of classifying the emotion conveyed by a text as negative, positive or neutral. It has a wide variety of applications in e-business and e-government as it extracts people's opinions, sentiments, appraisals, attitude towards entities such as products, services, organizations and their attributes using various machine learning techniques and natural language processing (NLP).

This paper applies sentiment analysis to analyze customers' opinions and reviews about two companies: Arik Airline and Guarantee Trust Bank using comparative analysis of supervised learning approach. To achieve this, specified tweets about these companies are extracted from twitter and pre-processed. The system architecture applied in this paper is shown in Fig. 1. Section 2 discusses the existing methods in sentiment analysis. Section 3 presents our research methodology. Section 4 covers discussion of results and Sect. 5 contains the conclusion.

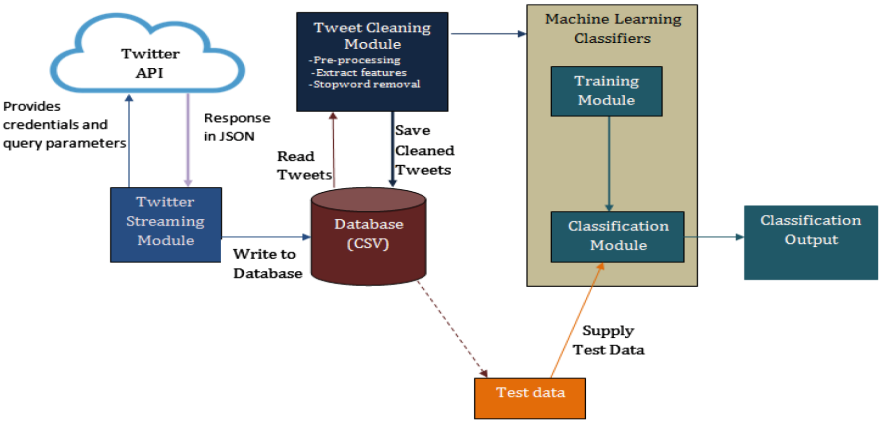


Fig. 1. Proposed system architecture

2 Data Mining and Sentiment Analysis

The objective of data mining is to extract information or knowledge from dataset and transform it into a structure that can be understood Dodd in [1] pointed that data mining focuses on discovering patterns in data while sentiment analysis focuses on discovering patterns in text that can be analyzed to classify the sentiment in that text.

2.1 Twitter Sentiment Analysis

Twitter is a micro-blogging service which enables users to send and read short text messages usually in 140 characters or less, known as "**tweet**". Twitter begun as a backup project for a failed project and today it is one of the highest growing social media websites in the world with huge amount of accumulated unstructured data written in natural language. Sentiment analysis on twitter posts is the process of accessing tweets for a particular topic as these tweets give us a rich and varied source of

opinions on product reviews or the individual state of mind with the help of different machine learning algorithms.

2.2 Sentiment Analysis Classification Techniques

Generally, researches on sentiment analysis require very fast and concise information to make accurate decisions which mostly depends on machine learning algorithms. Machine learning algorithms consist of two approaches: Supervised and Unsupervised machine learning approach.

Supervised Machine Learning Approach: This approach derives a function from labelled training examples consisting of a large set of examples about a particular topic. Each training example occurs as a pair of input and output (target) value. The algorithm analyzes the data and generates an output function which maps a new dataset to its appropriate class [2]. Naïve Bayes, Support Vector Machine, Maximum Entropy and Decision Tree are the most commonly used supervised machine learning techniques. Some of the work carried out on supervised machine learning approach can be found in [3–5, 6].

Unsupervised Machine Learning Approach: This approach is used when it is difficult to find labeled training documents. Major works carried out on the unsupervised machine learning approach can be found in [7, 8]. K-means, Spectral Clustering, Hierarchical Clustering, Partitioning Clustering and Semantic Orientation are commonly used unsupervised machine learning techniques.

3 Methodology

The workflow of the proposed system can be seen in Fig. 2 which consists of the following steps:

Step 1: Tweet Collection

The training data for the proposed system were collected from twitter with the help of the twitter streaming API. The dataset consists of roughly 6000 short messages collected on daily basis for the month of April 2017. They were stored and pre-processed for mining.

Step 2: Tweet Pre-processing

The aim of pre-processing tweet is to remove any piece of information within the tweet that will not be useful for the machine learning algorithm in assigning class to the tweets. The main pre-processing steps adopted in this paper include:

i. **Lowercase Conversion:** All tweets were converted to lower case.
ii. **Removal of URLs, @user and retweets:** All URL links (E.g. https://t.co/99GCMKVxHx), usernames (@omojuwa which indicates the user name) and special words (e.g. RT meaning Retweet) were removed with regular expressions to increase the accuracy of our classifiers.

iii. **Removal of stop-words:** Stop-words like *the, and, before, to, on, while* were eliminated. We built a custom list and added it to the list of stop-words available in the NLTK library.
iv. **Removal of duplicates and repeated characters:** People sometimes repeat letters to stress their emotion. Words like hunggrryyy, are used in place of 'hungry', haapppyyy instead of 'happy'. Such repeated letters were replaced by only two occurrences.
v. **Punctuation and whitespace Removal:** Punctuations in each word, words which do not start with an alphabet were removed while multiple whitespaces were replaced with single whitespace.

Step 3: Feature Extraction

We used the bag-of-words model to create the feature vector. Each tweet in the training dataset was split into words and each word added to the feature vector, some of the words which do not indicate the sentiment of a tweet were filtered out.

Step 4: Classifiers Used

Naïve Bayes, Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), Linear Support Vector Machine (SVM) and Decision Tree classifiers were used.

Naive Bayes

Naïve–Bayes are probabilistic classifier, it relies on the application of Bayes theorem given as

$$p = \left(\frac{c}{d}\right) = \frac{p(c) * p(c/d)}{p(d)}. \tag{1}$$

We used the Naïve Bayes Classifier and its variants from NLTK to train and test the data.

Support Vector Machine

The SVM is a non-probabilistic binary linear classifier. It plots the training data in a multidimensional space and discovers a hyper plane which separates the documents as per the sentiment, and the margin between the classes. A python package known as the Linear SVM from the Sci-kit learn was utilized to classify the tweet as NLTK does not provide libraries for SVM.

Decision Tree

A decision tree classifier is a tree whose nodes are labeled by the features as it categorizes a document starting at the tree node through the branches until it reaches the leaf using the information gain. The edges that leave these nodes are labeled by the class. The information gain measures how the input values will be organize once they are divided with a given feature using the formula

$$H = -\sum p(l) \times \log_2(l). \tag{2}$$

The decision tree ensures that each feature from the training set is checked in a particular order and to achieve this, the decision tree classifier in the NLTK library was used (Table 1).

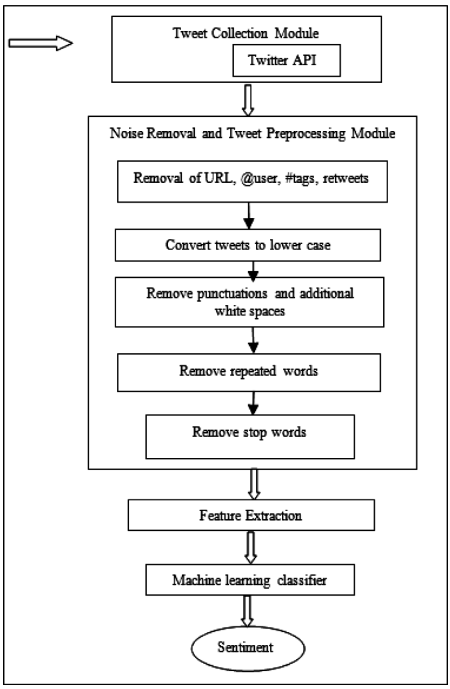


Fig. 2. Sentiment analysis process flow

4 Discussion of Results

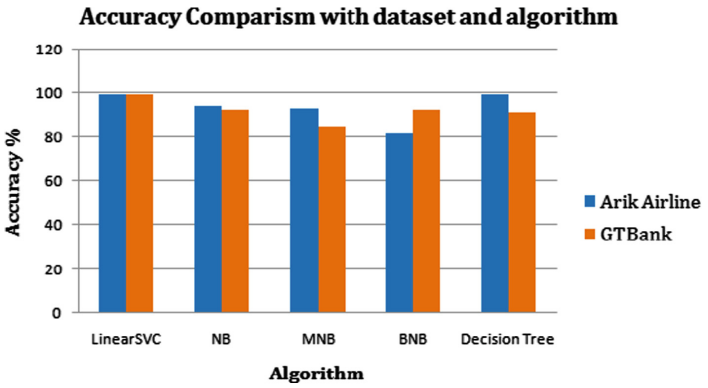
The supervised machine learning model was designed using the Hold Out validation method which separates the dataset into two sets called the training set and the test set. We trained with 3600 tweets (60%) and test with 2375 tweets (39.5%). Each tweet is classified to be positive, negative or neutral based on a query term and polarity classification, the percentage accuracy of each classifier was calculated using a python library and confusion matrix. Table 2 depicts the confusion matrices and accuracies of our classifiers.

Table 1. Confusion matrix

		Predicted		
		Positive	Negative	Neutral
Actual	Positive	True Positive (TP)	False Positive (FP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)	False Negative (FN)
	Neutral	False Neutral (FNeu)	False Neutral (FNeu)	True Neutral (TNeu)

Table 2. Confusion matrices with classifier accuracies

	GTBank			Arik Airline		
	Positive	Negative	Neutral	Positive	Negative	Neutral
Linear SVM						
Positive	557	0	5	106	0	0
Negative	0	319	2	0	57	0
Neutral	0	0	1039	0	0	285
Classifier accuracy = 99.62				99.99		
Naïve Bayes						
Positive	550	5	46	102	0	8
Negative	7	304	75	0	53	10
Neutral	0	10	925	4	4	267
Classifier accuracy = 92.56				94.19		
Multinomial Naïve Bayes						
Positive	500	301	58	97	1	8
Negative	47	301	150	6	54	12
Neutral	10	9	838	3	2	265
Classifier accuracy = 92.61				92.86		
Bernoulli Naïve Bayes						
Positive	550	3	60	97	5	26
Negative	4	315	71	6	50	40
Neutral	3	1	915	3	2	219
Classifier accuracy = 85.28				81.69		
Decision Tree						
Positive	104	0	0	548	4	65
Negative	1	57	0	1	315	92
Neutral	1	0	285	8	0	889
Classifier accuracy = 91.16				99.55		



5 Conclusion

The results of the analysis show that, the machine learning classifiers works correctly. Evaluation of the different algorithms shows that the Linear Support Vector Machine had the highest accuracy on all the datasets with short processing time. The Decision Tree classifier had the second highest.

References

1. Dodd, J.: Twitter sentiment analysis (2014). http://trap.ncirl.ie/1868/1/johndodd.pdf. Accessed 18 Mar 2017
2. Garg, P.: Sentiment analysis of Twitter data using NLTK in Python, June 2016. http://dspace.thapar.edu:8080/jspui/bitstream/10266/4273/4/4273.pdf. Accessed 15 Feb 2017
3. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC (2010)
4. Read, J.: Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Proceedings of the ACL Student Research Workshop, pp. 43–48. Association for Computational Linguistics (2005)
5. Parikh, R., Movassate, M.: Sentiment analysis of user-generated Twitter updates using various classification techniques (2009)
6. Kharde, V., Sonawane, S.: Sentiment analysis of Twitter data: a survey of techniques. Int. J. Comput. Appl. **139**(11), 5–15 (2016)
7. Suresh, A., Bharathi, C.R.: Sentiment classification using decision tree based feature selection. IJCTA **9**(36), 419–425 (2016)
8. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)