Analyzing Sentiment of Twitter Data using Machine Learning Algorithm

Dr. Shubhra Saxena
Associate Professor
S. K. I. T. Jaipur, Rajasthan.

Received Dec. 08, 2017 Accepted Jan. 14, 2018

ABSTRACT *A huge amount of data is generated every minute for social networking and content sharing via Social media sites that can be in a form of structured, unstructured or semi-structured data. Such data can be further used for business purposes that make possible to analyze such huge amount of data in distributed computing environment. One of the largest used social media sites is Twitter, where each and every day millions of data generated in the form of unstructured tweets. Tweets or opinions of the people can be used to extract sentiments of the people. Sentiment analysis is beneficial for organizations to improve their products and make required changes on demand to increase their profit. This paper shows the usefulness of text preprocessing with a selection of attributes that increases the accuracy of our model. Here a framework is proposed which is used to analyze sentiments of twitter with data mining tool thus provide some prediction of business intelligence. A comparative analysis of machine learning algorithm shows that Support Vector Machine (SVM) give overall better performance than Decision tree (DT) and Naive Bayes (NB).*
Keywords – *Sentiment Analysis, Text Preprocessing, Feature Selection, Results.*

1. INTRODUCTION

Twitter, the largest used social media site, has now become a very popular trend over the world for people who want to share an opinion about their social, political and economic interest. User opinion can be related to various aspects like gadgets, politics, products, services etc. that can directly convey the viewpoint of the user and helps in making predictions of a consumer market. Such kind of opinions or sentiments of huge people around the world is capable of performing analysis and future predictions.

Sentiment analysis (SA) is a process of detecting the contextual polarity of text in terms of positive, negative or neutral. Organizations across the world widely adopted the ability to extract insights from these sentiments of various social media sites. It helps organizations to make predictions of a certain product, reviews, and other decision-making processes that will ultimately increase the profit. So ultimately SA is beneficial for organizations and individuals to improve their profit as per user or market demand. Usually, tweets contain incomplete, poorly structured, noisy, irregular expressions, ill-formed words and non-dictionary terms [1]. Also, messages or tweets are short and have 140 lengths of limitations. So it requires preprocessing done on our collected datasets to reduce noise in tweets by removing stop-words, removing URLs, replacing negations etc. Sentiment dictionary contains all forms of a word with each word's polarity strength that can save more time.

In this paper we use three machine learning algorithms SVM, DT and NB for classifying our data also helps in evaluating the performance of our training dataset. We focused on comparing outcomes of these algorithms to identify best machine learning method which gives most accurate and efficient results for classifying twitter data.

2. SENTIMENT ANALYSIS

Social media sites become popular now a day as people over the world depends upon them to communicate with their relatives, friends, and rest of the world. SA also known as opinion mining, is a most popular trend in today's world which is the process of identifying and categorizing opinions on the web, determines the writer attitude towards a particular topic or product. It tells about what author wants to communicate and defines his state of mind in terms of emotions, feelings, and subjectivities about an event or topic. Natural Language Processing (NLP) is the interaction between the computers and the human/natural language [6]. NLP technique facilitates easy pre-processing of text i.e. NLP cleans and normalizes text for sentiment analysis. Analysis of sentiments can be based on single phrase or sentence, where the sentiment of the whole sentence is calculated. It contains following steps:

- Tweets posted on twitter are freely available through a set of APIs of twitter. At first, we collected a corpus of positive, negative, neutral and irrelevant tweets from twitter API.
- Then preprocessing done by removing stop words, negations, URL, full stop, commas etc. to reduce noise from tweets and to prepare our data for sentiment classification.
- After that, we apply machine learning algorithms to our dataset and compare their results.
- Results help us to identify which machine learning algorithm is best suited for classification of SA.

Applications of Sentiment analysis are broad and powerful that provide us easier and quicker social media monitoring like in: Consumer market for product reviews; Marketing to know consumer trends and attitude; Social media to find general user opinion about current topics; Movie to know whether released movie is liked or not, etc [11].

3. RELATED WORK

Zhao Jianqiang et al. [1] explored the effects of various preprocessing methods for classifying sentiments of twitter datasets. They presented some preprocessing methods for removing unwanted data such as removal of URLs, numbers, stop words; replacing negations, reverting repeated words and expanding acronym. Two feature models namely word n-gram and prior polarity score are used on five twitter datasets to identify the polarity of tweet sentiment. Four popular supervised classifiers are used to analyze the effect of preprocessing such as Naive Bayes, SVM, Logistic Regression, and Random Forest. Their experimental result shows that by removing stop words, numbers, and URLs from sentiments noise level can be reduced but it can't improve performance much more. By replacing negations and expanding acronym efficient results can be obtained also it can improve the accuracy of classification methods.

Akrivi Krouska et al. [2] described the role of preprocessing techniques for classification methods. They used 3-different sub-datasets to investigate the performance of some machine learning classification based algorithms with different preprocessing options. For preprocessing purpose they presented TF-IDF weighting scheme, Stemming, Stop-word removal and Tokenization technique. To increase classification accuracy and reduce training time they used feature extraction model. For classification of tweets, four machine learning algorithms are used by them namely Naive Bayes, SVM, K-Nearest Neighbor and Decision Tree. Their experimental result shows that appropriate feature selection and representation can improve performance for various classification methods.

Shruti Wakade et al. [13] evaluated the impact of sentiment words with emoticons based tweets to represent, label and classify twitter training datasets. They used iPhone and Microsoft related tweets for sentiment analysis process. A methodology is proposed by them for sentiment classification which consists following steps such as Data Collection layer, here tweets related to iPhone and Microsoft corpus data is collected; Preprocessing layer, used to remove unwanted data like: stop words, URLs from tweets and Snowball Stemmer is used to reduce words to its normal form; Feature Extraction layer, used to select subject-related words with their frequency distribution; and last Sentiment Labeling layer, used to create labeled training tweets. The experimental result indicates that decision tree algorithm outperform than Naive Bayes algorithm. It is also indicated that document filtering and indexing technique with proposed approach gives the effective tweet system analysis.

Hemalatha et al. [3], defined terms twitter and social network analysis on twitter. They proposed a model of sentiment analysis to make a business intelligence view. This model include three layers as Data Collection layer, used to collect tweets from networking sites; Preprocessing or data mining layer, used to remove repeated words, numbers, URLs, stop words, emotion icons, and special symbols from tweets; last SA layer, in which machine learning algorithms are used for data classification and polarity distribution to reviews. Two algorithms namely Naive Bayes and Maximum Entrophy are used by them for SA. Their experimental result presents that the size of files is gradually decreased from its original size when we apply preprocessing techniques. Also, they investigated that these two machine learning algorithms achieve higher accuracy for sentiment classification.

Manisha Rani et al. [5] introduced SA with Natural Language Processing (NLP) technique and its classes as positive, negative and neutral. They proposed an architectural diagram for SA which consists three layers such as: Data Collection, Preprocess, Feature Extraction and classification layer. The author described Naive Bayes, SVM and Lexicon-based approach for classifying sentiments. The Lexicon-based approach further classified as Dictionary-based, corpus-based and used to determine polarity. They concluded that NB and SVM algorithms perform better and provides higher accuracy rates than lexicon-based algorithm.

R. Nivedha et al. [14] introduced Twitter as a knowledge discovery tool which contains health and non-health related tweets that can be useful for patients, doctors, and researchers. Also, tweets contain noisy information that can be removed by preprocessing without it can't be classified further. They presented a methodology to classify social media data which consist following steps: first Data Extraction, at first data, is extracted from easily available Twitter API, this kind of dataset contain unlabeled and noisy information by nature; second Preprocess stage, used to remove unwanted data from raw datasets; last

Classification or data mining stage, used to classify data for that they found Decision Tree algorithm most popular and easy to use. Their experimental result shows that size of file or words can be reduced with preprocessing technique to classify plain text into health and non-health data they used CART algorithm. They found that Naive Bayes performs better than CART by achieving accuracy rate 84.21%.

4. PROBLEM DEFINITION

As users on social media sites are rapidly growing and producing a large amount of data every day, so there is a need to classify and analyze these messages to find out its polarity about some topic or event. Emotions and opinions can be expressed in many ways. Classifying sentiments that have few relative classes such as "positive", "negative", or "neutral", is the most complicated task. SA is a popular topic and lots of research has been going on from a long time. Many researchers used supervised learning algorithms also with various automatic classifiers for classification of the polarity of sentiments. The problem is in assigning the strongest polarity of sentiments and in finding the best algorithm which provides most accurate results.

5. EXPERIMENTAL FRAMEWORK

This paper presents a model presented in figure-1, which consists of three layers for analyzing sentiments. First Data Collection layer, used to collect tweets from twitter APIs; Second Data preprocessing layer with a selection of attributes which is used to reduce noise level from tweets, and last SA or Data Mining layer used to apply machine learning algorithm [2].

5.1 Data Collection

At first, we obtain training data of twitter sentiments from 2- different twitter API. First, dataset taken from "Twitter Sentiment System for SemEval 2016", (denoted by "SE-T") contains approx 13541 tweets with 2-attributes namely: class and content [15]. Second dataset is taken from "Sanders Analytics twitter sentiment corpus" (denoted by TS), which contains 479 instances with class and text two attributes [16].

Table-1 Statistics of two twitter datasets used

Dataset	Total Tweets	Positive	Neutral	Negative
E-Twitter **(SE-T)**[15]	13541	5232	6242	2067
Twitter-sanders**(TS)**[16]	479	163	-	316

Table-1 shows statistics of both datasets with their attributes they contain. We store these datasets in .csv format then convert it into a .arff format for further analysis.

5.2 Preprocessing Setup

For getting accurate results by classifiers we have to make sure that these datasets processed efficiently by removing unrelated contents and thus related contents are accurately extracted. As most researchers consider that URL doesn't have any information regarding sentiments, so by removing short URLs from tweet contents can be refined. People often use emotional words that contain repeated letters to express their sentiments which are very common trends like "coooool". Also, numbers are not used for analyzing sentiments so tweet contents can be refined by removing them [1]. The polarity of the word will be changed when they are preceded by a negation or negation can change/reverse the meaning of words. By checking negations, Removing of URLs, emotions, numbers and Repeated Word; noise in tweets can be reduced.

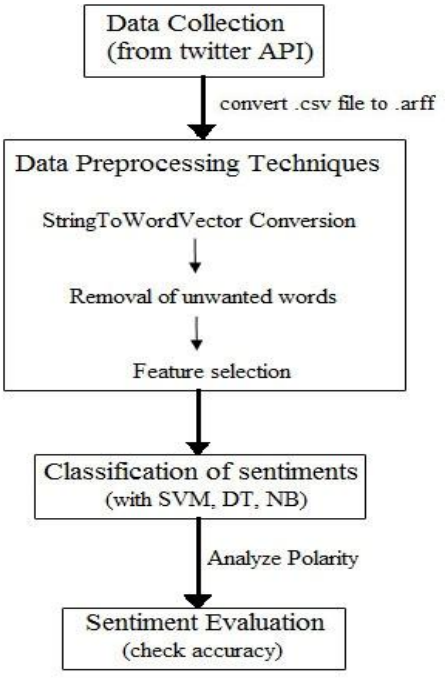

Figure-1 System Architecture using Machine learning algorithms

To perform preprocessing with WEKA we use filter "StringToWordVector" that is used to convert string attributes into a set of attributed representing word occurrences information from the text contained in the strings. This filter provides us options to do configuration with our dataset which includes following steps:

Stemming: It is used to remove suffix from the word according to some grammatical rules. Here we apply most popular Snowball Stemming library.

Stop Word Extractor: Some words that don't have polarity so they don't need to be further

analyzed like: able, are, both, which, has, become, after etc. So after elimination of these words, our result will not be affected. We used Rainbow list for our experiment.

Tokenization: It is used to split a document into a word or terms and make a word vector. Here we used NGramTokenizer.

Feature Selection: This process decreases the number of attributes into a better subset which can increase accuracy also it brings a reduction in training time. It is done by using Filters and Wrappers. WEKA provides "AttributeSelection" filter to choose an attribute evaluation method. We use "cfsSubsetEval" method which considers the individual predictive ability of each feature to evaluate the worth of an attribute.

5.3 Sentiment Classifier

To classify sentiments machine learning (ML) algorithms are used i.e. a branch of Artificial Intelligence (AI) concerned with the study of classification and pattern analysis, allows the computer to learn behaviors of empirical data taken from sensors or database. ML algorithm allows us to automatically recognize complex patterns and make intelligent decisions based on data. In this paper, we used various machine learning algorithms such as Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT).

5.3.1 Naïve Bayes Classifier

It refers to counting the frequency of words that are related to the sentiments in the message. As Bayes theorem based on probabilistic classifier so it allows us to capture uncertainty about the model to determine the probability of the outcome. Explicit probabilities can be calculated by it for the tested dataset and it helps to reduce noise robustly. It is numerical based approach with easy, fast and high accuracy features.

5.3.2 Support Vector Machine (SVM)

It yields more accurate results when it is used for classifying text. The basic idea behind it is to find the hyperplane (or vector w), which is responsible for separating one class document vector from the vector in other class [6]. It is successfully employed in text classification and various other sequence processing applications as it is a type of linear classifier.

5.2.3 Decision Tree

It is a flowchart used to output labels for certain features, act as input values. It categories a document as by, starting from the tree root (labeled as features), followed downward by branches (labeled as features weight) and last reached a leaf node (labeled by categories).

6. EVALUATION

6.1 Experimental Setup

We use Waikato Environment for Knowledge Analysis (WEKA) to implement data mining algorithms for preprocessing, classification, clustering, and analysis of results. This environment includes java libraries that implement algorithms and provide the best environment to researchers for classifying datasets. We apply "StringToWordVector" filter and done lots of preprocessing with our datasets. Using n-gram tokenizer option and attribute selection method different number of attributes are created. With attributes selection method 50 attributes are taken for testing out of 1613 words from first dataset SE-T [15] and 105 attributes out of 2065 words are taken from second dataset TS [16]. This method increases accuracy rate of our training dataset also it brings a reduction in execution time. Following table-2 shows reduction in size of file after preprocessing:

Table-2 Compare file size of both dataset with preprocessing techniques

	SE-T[15]	TS[16]
Size of file before preprocessing	1.7 MB	93.7 KB
Size of file after preprocessing (feature selection)	181 KB	9.9 KB

To evaluate performance we apply 10-fold cross validation technique which splits the original set into training sample to train the model and a test set to evaluate results.

For computing sentiments quickly of tweets without compromising accuracy, an approach known as "Information Retrieval Metrics" can be used to evaluate experimental results in terms of precision, recall, f-measure, and accuracy with the use of following formulas [9]:

Precision: TP/ (TP+ FP)

Recall: TP/ (TP+FP)

F-measure: 2* Precision* Recall/ (Precision+ Recall)

Accuracy: TP+TN/ (TP+ TN+ FP+ FN)

Here (TP= True Positive; TN= True Negative; FP=False Positive; FN= False Negative)

6.2 Experimental Results

We observed that our classification results improved in terms of time and accuracy using processed and small features data than simple datasets. For example in first SE-T dataset, time taken to build a model for NB algorithm takes 10.56 seconds, accuracy 53.73% and after processing time taken to test model on training data is reduced at 0.35 seconds only, accuracy

improved by 57.46%. Table-3 demonstrates the accuracy of classifiers on both datasets after applying various preprocess methods.

Table-3 Accuracy Criteria for both datasets

Evaluation Criteria	Dataset	SVM	DT	NB
Correctly classified instances	SE-T	8335	8118	7776
	TS	419	412	355
Incorrectly classified instances	SE-T	5206	5423	5765
	TS	60	67	124
Accuracy (%)	SE-T	61.55	59.95	57.42
	TS	87.47	86.01	74.11
Error	SE-T	0.38	0.40	0.42
	TS	0.13	0.14	0.26

The number of correctly classified instances and accuracy rate is greater in both cases with SVM algorithm. In our experiment obtained accuracy using SVM algorithm is 61.64% and 87.47% respectively (with 50 feature SE-T and 105 feature TS datasets) which is greater than other two algorithms.

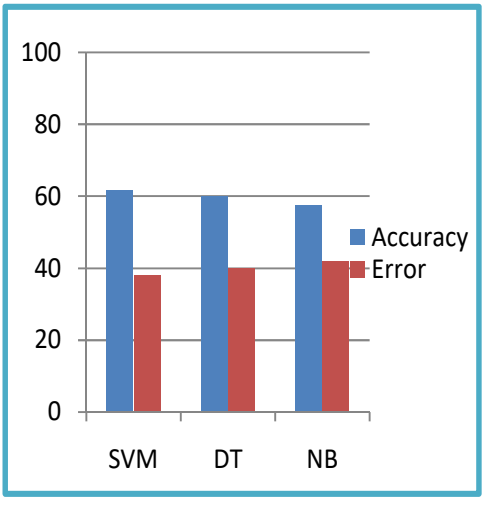

Figure-2 Accuracy measured by SE-T dataset

Figure-2 and 3 demonstrate the accuracy measured by both datasets that we obtained from our experimental results as shown in table-3.

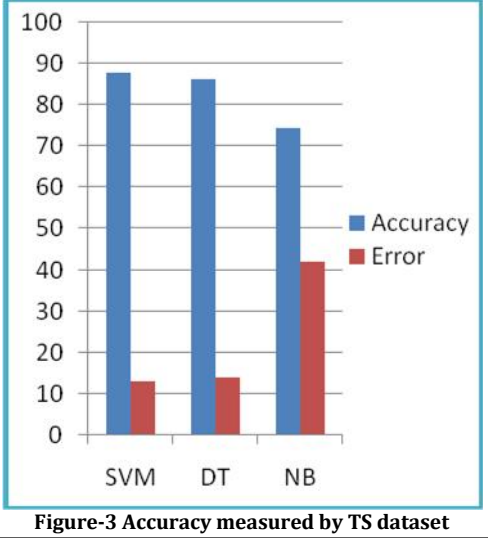

Figure-3 Accuracy measured by TS dataset

Following performance measures are reported in table-4 by our experimental result using first SE-T and second TS dataset, after conducting 10-fold cross validation technique:

Table-4 Performance measured by algorithms in both datasets

		TP Rate	FP Rate	Precision	Recall	F-Measure	
SVM	SE-T	0.352	0.036	0.859	0.352	0.500	positive
		0.906	0.562	0.580	0.906	0.707	neutral
		0.404	0.070	0.510	0.404	0.451	negative
	TS	0.644	0.006	0.981	0.644	0.778	positive
		0.994	0.356	0.844	0.994	0.913	negative
DT	SE-T	0.34	0.048	0.821	0.349	0.489	positive
		0.914	0.603	0.564	0.914	0.698	neutral
		0.284	0.054	0.486	0.284	0.359	negative
	TS	0.601	0.006	0.980	0.601	0.745	positive
		0.994	0.399	0.828	0.994	0.904	negative
NB	SE-T	0.437	0.186	0.597	0.437	0.505	positive
		0.780	0.479	0.582	0.780	0.667	neutral
		0.300	0.063	0.461	0.300	0.364	negative
	TS	0.252	0.006	0.953	0.252	0.398	positive
		0.994	0.748	0.720	0.994	0.835	negative

Our experimental result shows that same preprocessing methods on a different dataset affect similarly the classifiers performance. After analyzing results of Table-4 it is observed that SVM provides 64.96% and 91.25% overall precision which is better than other two algorithms. Also, overall Recall and F-measure rate of SVM is greater than NB and DT in both datasets which is demonstrated in figure-3 and figure-4 also.

Also time taken to build a model is greatly reduced by applying feature selection method. Time taken to build model in first SE-T datasets is 0.45, 29.43, 4.47 seconds respectively with NB, SVM, and DT algorithm; in second TS dataset, it is 0.01, 0.06, 0.01 seconds with NB, SVM and DT algorithms respectively.

7. CONCLUSION & FUTURE SCOPE

In this paper, we discuss Sentiment analysis which can tell us the thought of writers about the particular entity. These days, it becomes a routine task to find people sentiments about a real-world entity from social media sites like Twitter, facebook or blogs etc. To efficiently analyze this large amount of datasets it is essential to accurately classify it.

In this paper, we have presented a methodology of data mining using Weka tool for classifying sentiments of twitter. We use three machine learning algorithms SVM, DT, and NB for classifying sentiments of twitters data. We conduct an experiment on two twitter's datasets to verify the effectiveness of pre-processing. Our experimental results indicate that by removing unwanted words and selecting features in the preliminary phase of preprocessing, time to build model is reduced and also it provides more accurate results in applied algorithms. The result may be affected by the choice of features for training and choice of algorithm for sentiment classification. The performance of SVM, DT, and NB algorithms improve on datasets after removing unwanted words. Therefore, removing unwanted words is useful to improve the performance of sentiment classification. We discuss the comparative analysis of three algorithms and calculate overall performance measures in terms of precision, recall, and f-measure. Our experimental results indicate that SVM provides more accurate results than other algorithms.

However, it is important to further study current available preprocessing techniques that help us to improve results of various classifiers. A method should be found to automatic incorporate feature selection at time of model building according to any language.

REFERENCES

[1] Zhao Jianqiang, Gui Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis", IEEE Access, DOI 10.1109/ACCESS. 2017, 2672677

[2] Akrivi Krouska et al., "The effect of preprocessing techniques on Twitter SA" , DOI: 10.1109/IISA.2016.7785373, 7th International Conference on Information, Intelligence, Systems & Applications (IISA), Research Gate Conference: 2016

[3] Bholane Savita D. et al, "Sentiment Analysis on Twitter Data Using Support Vector Machine", International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 3, May - Jun 2016

[4] Hemalatha et al, "Sentiment Analysis Tool using Machine Learning Algorithms", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 2, Issue 2, March-April 2013

[5] Alexander Pak et al, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining ", pages 1320-1326, Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, FRANCE

[6] Manisha Rani et al, "A Review of Data Analysis of Twitter", Volume 6, Issue 5, May 2016 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[7] Luciano Barbosa et al, "Robust Sentiment Detection on Twitter from Biased and Noisy Data", Coling 2010: Poster Volume, pages 36–44, Beijing, August 2010

[8] Bo Pang et al, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of EMNLP 2002, pp. 79–86.

[9] G. Vinodhini et al, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6 , 2012

[10] Hana Anber et al, "A Literature Review on Twitter Data Analysis", 2016, International Journal of Computer and Electrical Engineering

[11] Snehal. A. Mulay et al, "Sentiment Analysis and Opinion Mining With Social Networking for Predicting Box Office Collection of Movie", International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-5, Issue-1)

[12] Yeliz Yengi , Sevinç İlhan Omurca, "Distributed Recommender Systems with Sentiment Analysis", European Journal of Science and Technology Vol. 4, No. 7, pp. 51-57, June 2016

[13] Shruti Wakade et al, "Text Mining for Sentiment Analysis of Twitter Data", The University of Akron, Department of Computer Science

[14] R. Nivedha and N. Sairam, "A Machine Learning based Classification or Social Media Messages", Indian Journal of Science and Technology, Vol 8(16), DOI: 10.17485/ijst/2015/v8i16/63640, July 2015

[15] https://github.com/WladimirSidorenko/SemEval-2016 , SemEval-2016 Task 4: Sentiment Analysis on Twitter, Training + Dev dataset

[16] https://github.com/guyz/twitter-sentimentdataset, Sanders Analytics twitter sentiment corpus

[17] Raymond Hsu et al. , "Machine Learning for Sentiment Analysis on the Experience Project", http://www.experienceproject.com

[18] I.Hemalatha et al., "Automated Sentiment Analysis System Using Machine Learning Algorithms", IJRCCT, Vol 3, Issue-3, March-2014

Florian B¨utow et al., "Semantic Search: Sentiment Analysis with Machine Learning Algorithms on German News Articles", http://www.aot.tu-berlin.de/