International Journal on Recent and Innovation Trends in Computing and Communication
Volume: 5 Issue: 6

ISSN: 2321-8169
434 – 439

Polarity Classification of Twitter Data using Sentiment Analysis

Arvind Singh Raghuwanshi
M.Tech, Computer Science and Engineering Department
Samrat Ashok Technological Institute
Vidisha(M.P.), India
e-mail: arvindraghuwanshi1411@gmail.com

Satish Kumar Pawar
Asst. prof., Computer Science and Engineering Department
Samrat Ashok Technological Institute
Vidisha(M.P.), India

Abstract—Crowd source information is of vital importance these days, since we relay much on information available from internet. Thus, Sentiment analysis or opinion mining becomes one of the major tasks of NLP (Natural Language Processing) and has gained much attention in recent years. Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, specifically to determine whether the user's attitude towards a specific area or product in case of ecommerce, etc. is positive, negative, or neutral. Sentiment analysis application are broad and powerful. It can be helpful in many ways like it helps marketers to evaluate the success of an ad campaign, in new product launch, to determine which versions of a product or service are popular and it also identifies which demographics like or dislike product features. This paper evaluates two classifiers, one is linear and other is probabilistic for sentiment polarity categorization. Data used in this study are the tweets collected from twitter.com. We further represent a comparative study of three different algorithms, Naïve Bayes, SVM (Support Vector Machines), and Logistic regression and how they vary on the same data set.

Keywords*-component; Logistic regression, Natural language processing, Sentiment analysis, Support vector machines, Twitter.*

*****

I. INTRODUCTION

SENTIMENT analysis or opinion mining is the process of determining the emotion behind a series of words, generally for use in social media, opinions and emotions expressed over an online platform [1]. It is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a topic, product, etc. is positive, negative, or neutral. Sentiment analysis application are broad and powerful. Opinion mining or sentiment analysis is widely used in areas such as reviews and survey responses, online and social media, and health care materials for applications that range from marketing to customer services to clinical medicine. Sentiment analysis is extremely useful in monitoring of social media as it allows us to gain an overview of the wider public opinion [1].

Opinion mining extracts and analyzes people's opinions about an entity whereas sentiment analysis identifies the sentiment expressed in a text then analyzes it [2]. Therefore, the target of sentiment analysis is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Fig. 5.

Sentiment analysis usually involves classification levels in analysis process that is: document-level, sentence-level, and aspect-level analysis. Document-level analysis aims to classify an opinion document as expressing a positive or negative opinion or sentiment. Sentence-level analysis aims to classify sentiment expressed in every single sentence. Aspect-level or entity level analysis aims to classify the sentiment with respect to the specific aspects of entities [2].

However, data available online have several flaws that potentially hinder the process of sentiment analysis. First, since people can freely post their own content, the quality of their opinions cannot be guaranteed or justified. For example, instead of sharing topic-related opinions, online spammers post spam on forums. [3][4]

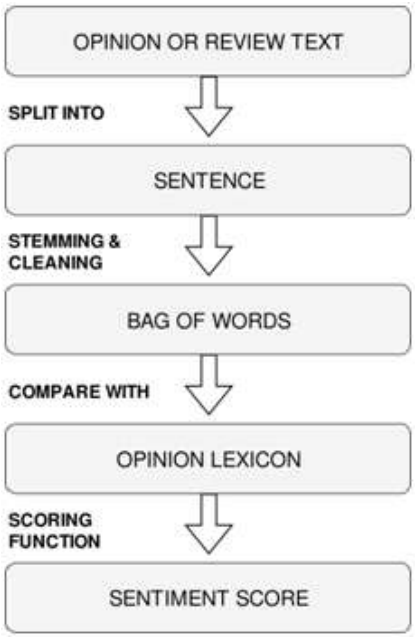

Fig. 1. Process of sentiment analysis

 ISSN: 2321-8169
434 – 439

Other important problem with this data is that the ground truth of such online data is not always available. A ground truth is more like a tag of a certain opinion, indicating whether the opinion is positive, negative, or neutral [5]. This paper deals with the fundamental issue of sentiment analysis, called sentiment polarity categorization.

The rest of this paper is organized as follows: 'Background and Literature Survey', provides a brief review towards some related work on sentiment analysis. Details of experiment and analysis is covered in 'Experiment and Preliminaries' section. Software package and classification models used in this study are presented in the sub-section 'Methods'. Results evaluation and analysis are presented in section 'Result Evaluation'. Discussion and future work is presented in section 'Conclusion and Future Work' which highlights the area still to be address in sentiment analysis.

II. BACKGROUND AND LITERATURE SURVEY

One of the fundamental problem while performing sentiment analysis is the categorization of sentiment polarity i.e. positive, negative, and neutral. In a posted piece of written text, it can be review, tweet or opinion, the problem is to categorize the text into one specific sentiment polarity out of three, positive or negative (or neutral). Depending up on the area we choose for analysis, there are three levels of sentiment polarity in which they categorized namely the document level, the sentence level, and the entity also called aspect level [6].Minqing Hu and Bing Liu[7] in their work summarized a list of positive words and a list of negative words, respectively, based on reviews collected from users which include misspelled words also. To classify different categories of sentiment is specifically a classification problem, where the features that contain opinions or sentiment information other than general construct need to be identified before the classification. Bo Pang and Lillian Lee[8] in their work suggested separation of objective sentences by extracting subjective ones [8] with a text-categorization technique that can identify subjective content using minimum cut.

Gann et al in his work [9] represents 6,799 selected tokens on Twitter data (tweets), and where each token is assigned a sentiment score, also called TSI (Total Sentiment Index). Whose main feature is to tag itself as positive token or a negative token. This classifier defines the main area to hit while performing sentiment analysis of any data (structured or unstructured data) available from any source. This paper also aims to explore polarity categorization.

Sentiment analysis is one of the fastest growing research areas in computer science, making it challenge to keep track of all the activities in the area. We present a computer-assisted literature review, where we utilize both text mining and qualitative coding, and analyse 6,996 papers from Scopus. We find that the roots of sentiment analysis are in the studies on public opinion analysis at the beginning of 20th century and in the text subjectivity analysis performed by the computational
linguistics community in 1990's. However, the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts on the Web. Consequently, 99% of the papers have been published after 2004. Sentiment analysis papers are scattered to multiple publication venues, and the combined number of papers in the top-15 venues only represent ca. 30% of the papers in total. We present the top-20 cited papers from Google Scholar and Scopus and a taxonomy
of research topics. In recent years, sentiment analysis has shifted from analysing online product reviews to social media texts from Twitter and Facebook. Many topics beyond product reviews like stock markets, elections, disasters, medicine, software development and cyberbullying extend the utilization of sentiment analysis.

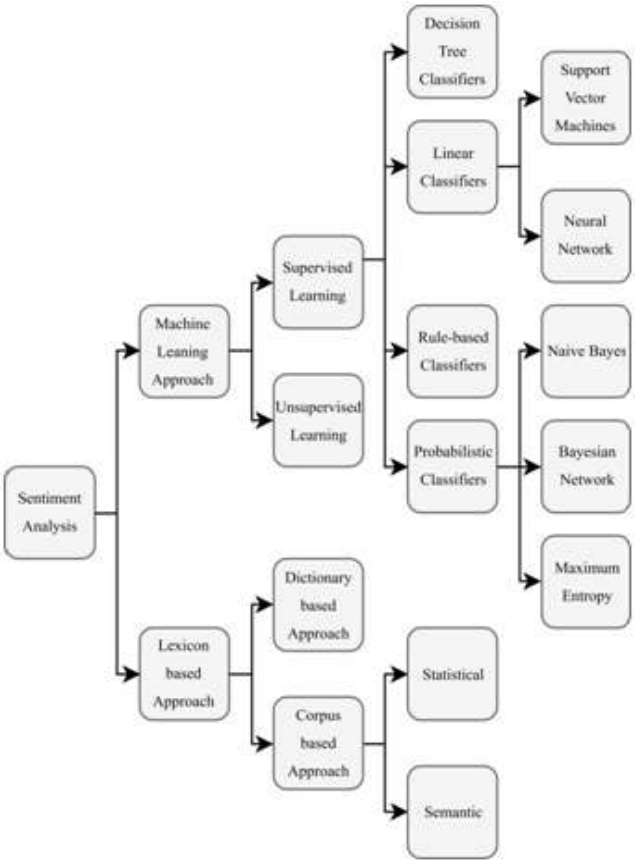

Fig. 2. Classification

III. EXPERIMENT AND PRELIMINARIES

This paper represents a competitive model that compares linear and probabilistic approach. This paper deals with classifier part in sentiment analysis of any data. The algorithm we used to represent linear model is SVM (Support Vector Machines) and for probabilistic model it is

 ISSN: 2321-8169
434 – 439

Logistic Regression and Naive Bayesian Classifier.

*Data Set:*Data set used are the tweets collected from twitter.com.

*Working:*It includes following steps- gathering data, analyzing their sentiments, assigning tweets to different categories using classifier and finally visualizing the results and making them more consumable and understandable.

It includes following steps- gathering data, analyze their sentiment, assign tweets to different category using classifier, and finally visualizing our results and make them more consumable and understandable.

To give a detailed overview of the process, steps were taken in the following order:

a. Tokenizing- Splitting sentences and words from the body of text.
b. Part of Speech tagging.
c. Machine learning with algorithms and classifiers.
d. Tie in scikit-learn (sklearn).
e. Training classifiers with dataset.
f. Performing live, streaming, sentiment analysis with twitter.

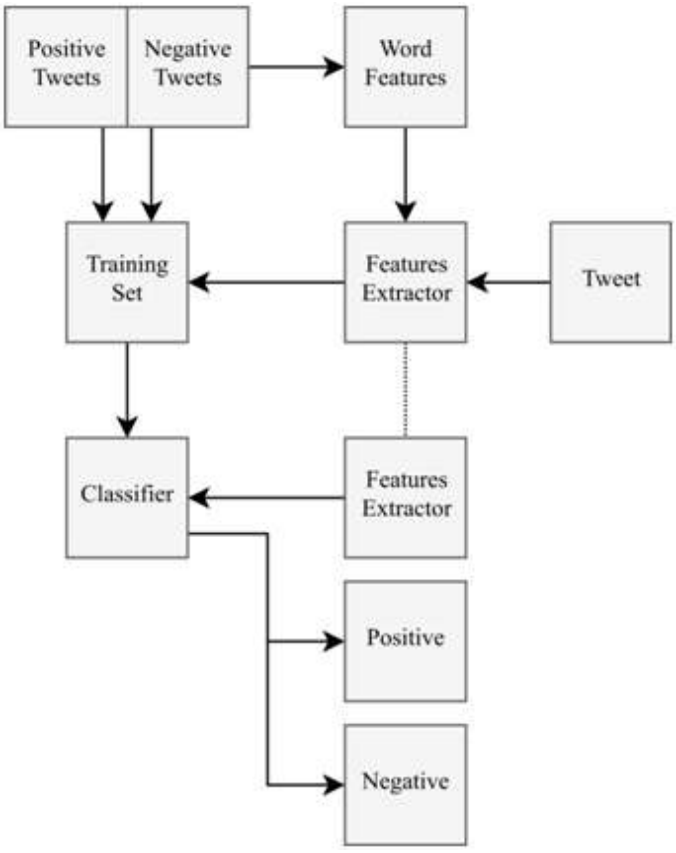

Fig. 3. Workflow

METHODS

Software used for this study is scikit-learn, an open source machine learning software package in Python. The classification models selected for categorization are- Naïve Bayesian, Logistic Regression, and Support Vector Machine.

A. *Support Vector Machines*

Support Vector machines represents a linear model classifier. Support vector machines (SVM) is a method by which we can classify both linear and nonlinear data. For linearly inseparable data, the SVM searches for the linear optimal separating hyper plane (the linear kernel), which works as a boundary for the decision that separates data of one class from the other class. Mathematically, a separating hyper plane can be written as: $W \cdot X + b = 0$, where Wa weight vector and$W = w_1, w_2, \ldots w_n$. X is a training tuple. b is a scalar. To optimize the hyper plane, the problem essentially transforms to the minimization of $||W||$, which is eventually computed as:$\sum_{i=0}^{n} \alpha_i y_i x_i$ where α_i are numeric parameters, and y_i are labels based on support vectors,x_i. That is: ify_i=1, then $\sum_{i=0}^{n} w_i x_i \geq 1$; if y_i=-1 then $\sum_{i=0}^{n} w_i x_i \geq -1$[10] [11].

And for linearly inseparable data, the SVM uses nonlinear mapping to transform the data into a higher dimension. It therefore solves the problem by finding a linear hyper plane. Functions to perform such transformations are called kernelfunctions. The kernel function selected for our experiment is the Gaussian Radial Basis Function-(RBF):

$$K(X_i, X_j) = e^{-\gamma \|X_i - X_j\|^2 / 2}$$

Where X_i are support vectors, $X_j$are testing tuples, and γ is a free parameter that uses the default value from scikit-learn in our experiment. [14]

The other advantages of support vector machines are:

a. Effective in high dimensional spaces.
b. Still effective in cases where number of dimensions is greater than the number of samples.
c. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
d. Versatile: different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

a. If the number of features is much greater than the number of samples, the method is likely to give poor performances.
b. SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see scores and probabilities, below).

B. Logistic Regression

Logistic regression represents probabilistic model. Logistic regression is used to find the probability of event=Success and event=Failure. Logistic regression is used when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can be represented by following equation:

$$Odds = \frac{p}{(1-p)} = \frac{\text{probability of event occurrence}}{\text{probability of not event occurrence}}$$

$$\ln(odds) = \ln(p/(1-p))$$

$$\text{Logit}(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$$

Above, p is the probability of presence of the characteristic of interest.

To work with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution and, here it is logit function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression). [12]

Logit Function

A logit function is simply a function of the mean of the response variable Y that we use as the response instead of Y itself.

All that means is when Y is categorical, we use the logit of Y as the response in our regression equation instead of just Y:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_k X_k$$

The logit function is the natural log of the odds that Y equals one of the categories. For mathematical simplicity, we're going to assume Y has only two categories and code them as 0 and 1. [13]

C. Naïve Bayesian Classifier

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above. As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects. [16]

Fig. 4. Naïve Bayes Classification

Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

The Naïve Bayesian classifier works as follows: Suppose that there exist a set of training data, *D*, in which each tuple is represented by an *n*-dimensional feature vector, $X=x_1,x_2,..,x_n$, indicating *n* measurements made on the tuple from *n* attributes or features. Assume that there are *m* classes, C_1, C_2, ..., C_m. Given a tuple *X*, the classifier will predict that *X* belongs to $C_i$if and only if: $P(C_i|X) > P(C_j|X)$, where $i, j \in [1, m]$ $and i \neq j$. $P(C_i|X)$ is computed as:

$$P(C|X) = \prod_{k=1}^{n} P(x_k|C_i)$$

Advantages of Naive Bayes Classifier:

a. It is easy and fast to predict class of test data set. It also performs well in multi class prediction.
b. When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
c. It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

Disadvantages of Naive Bayes Classifier:

a. If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

 ISSN: 2321-8169
434 – 439

b. On the other side, Naive Bayes is also known as a bad estimator, so the probability outputs are not to be taken too seriously.
c. Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

IV. Result Evaluation

Experiment is carried out using 10-fold cross validation. A 10-fold cross validation is applied as follows: The dataset is partitioned into 10 equal size subsets, each of which consists of 10 positive class vectors and 10 negative class vectors. One of the 10 subsets are selected, and that single subset is retained as the validation data for testing the classification model along with others, and the remaining 9 subsets are used as training data.

Performance of each classification model is estimated by generating confusion metric with the calculation of precision and recall. Using precision and recall value, F1-score is calculated and result are compared on it. [15]

$$F1_{avg} = \frac{\sum_{i=0}^{n} \frac{2 \times P_i \times R_i}{P_i + R_i}}{n}$$

Here P_i is the precision of the i^{th} class, R_i is the recall of the i^{th} class, and n is the number of classes.

Results obtained by applying different methods are given in TABLE I:

TABLE I : RESULTS

Algorithm	Accuracy
SVM	78.82530122684955
Logistic Regression	76.18072290722891
Naïve Bayes	71.53614465930557

"Accuracy is in percentage", SVM = Support Vector Machines

Following graph represents the polarity of tweets (positive or negative):

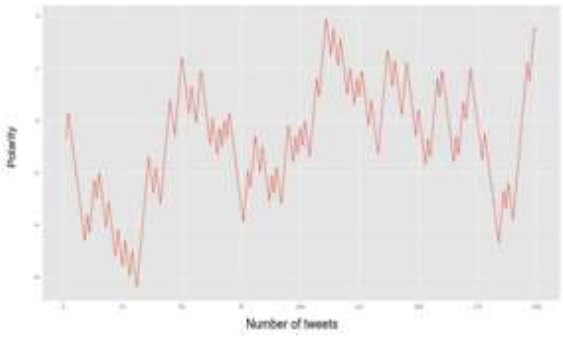

Fig. 5. Polarity of tweets

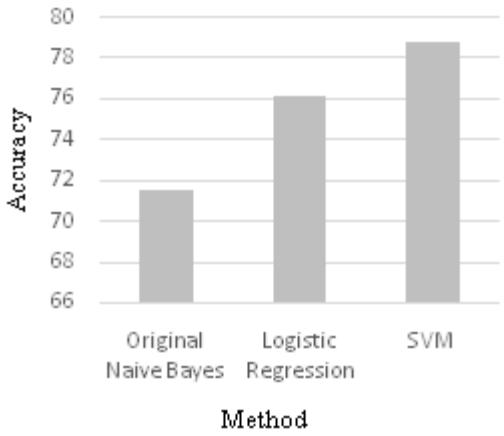

Fig. 6. Comparison of Original Naive Bayes, SVM, and Logistic Regression

Following graphs shows the performance of various methods that we have chosen for evaluation.Out of which,SVM turns out to be best among all. It can work with linear or non-linear data.

V. Conclusion and Future Work

Data is the essence of entire work since people relay more and more on information available over internet these days, which involve natural language. This paper tackles a fundamental problem of sentiment analysis, sentiment polarity categorization. Online twitter datasets are selected as data used for this study which ultimately represents a comparatively model of linear, probabilistic and discriminative classifier.

To make data available for processing and extracting exact emotions are two major area to work in this field. We further need more efficient machine learning, deep learning algorithm for better classifier. Also, there is a lot way to go to deal Spam post/tweets. Better mining techniques will help to deal natural language processing more efficiently.

References

[1] https://www.brandwatch.com/blog/understanding-sentiment-analysis/
[2] WalaaMedhat, Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093-1113, May 2014.
[3] Bing Liu, "The Science of Detecting Fake Reviews", May 2012. Available: http://content26.com/blog/bing-liu-the-science-of-detecting-fake-reviews/
[4] Nitin Jindal, Bing Liu. (2008, Feb.). Opinion spam and analysis. Presented at WSDM '08 Proceedings of the 2008 International Conference on Web Search and Data Mining,

Palo Alto, California, USA. Available: http://dl.acm.org/citation.cfm?id=1341560
[5] Stanford Sentiment 140. Available: http://www.sentiment140.com
[6] Bing Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, May 2012. Available: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf
[7] Minqing Hu, Bing Liu. (2004, Aug.). Mining and summarizing customer reviews. Presented at KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA. Available: http://dl.acm.org/citation.cfm?id=1014073
[8] Bo Pang, Lillian Lee. (2004, Jul.). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. Presented at ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain. Available: http://dl.acm.org/citation.cfm?id=1218990
[9] Gann W-JK, Day J, Zhou S. (2014). Twitter analytics for insider trading fraud detection system. Presented at second ASE international conference on Big Data.
[10] B. Pang, L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, Jan. 2008.
[11] Patetta M., "Categorical Data Analysis Using Logistic Regression Course Notes," Copyright © 2002 by SAS Institute Inc., Cary, NC 27 13, USA.
[12] James Jaccard, "Interaction Effects in Logistic Regression," SAGE Publications, Inc., 2001. Available: https://us.sagepub.com/en-us/nam/interaction-effects-in-logistic-regression/book11268
[13] http://www.theanalysisfactor.com/what-is-logit-function/
[14] Scikit-learn, 2014. Available: http://scikit-learn.org/stable
[15] Alexander Pak, Patrick Paroubek. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Presented at LREC 2010, Seventh International Conference on Language Resources and Evaluation, Valletta, Malta. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf
[16] Han J, Kamber M, Pei J, "Data Mining: Concepts and Techniques," Second Edition (The Morgan Kaufmann Series in Data Management Systems), 2nd ed., Morgan Kaufmann, San Francisco, CA, USA. 2006.