

SENTIMENT ANALYSIS OF TWITTER DATA IN R USING LEXICON, NAÏVE BAYES AND LOGESTIC REGRESSION

¹Sharat Babu Jami, ²Ajit Kumar Pasayat

¹M. Tech Student, ²Asst. Professor

¹Department of Computer Science and Engineering,

¹Centurion University of Technology and Management, Bhubaneswar, India

Abstract : *Social media is an internet-based form of communication. Social media platforms allow users to have conversations, share information and create web content. Billions of people around the world use social media to share information and make connections. Facebook, Twitter, Instagram, etc. are the mostly used social media applications which we use daily. We post, share, tweet in different applications in situations we go through. We also respond to many posts or tweets which were posted by popular personalities in the society. Sentiment Analysis is growing exponentially due to the importance of the automation in mining, extracting and processing information in order to determine the general opinion of a person. The problem that this paper proposes to address is to determine what methods are more suitable to extract subjective impressions in real time from Twitter, since the opinions collected from Twitter are limited to certain amount of characters and it will happen in a real-time environment, this provides an interesting scenario; we will test using both the Machine Learning Approach and the Lexicon-based Approach, It investigates the most popular document ("tweet") representation methods which feed sentiment evaluation mechanisms.*

Index Terms - Sentiment Analysis, R programming, Lexicon analysis, Naïve Bayes, Logistic regression

I. INTRODUCTION

With the advancement in technology, communication has grown. It is now easier and cheap to communicate and connect with people across the world. The issue of distance is no longer an excuse for lack of communication. Communication systems have grown from wired devices to wireless devices. Social networks allow people to keep and manage accounts. Social media can be seen as a perfect replacement to conferencing makes it possible to reach many people in a very short time[1]. Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Opinion mining (sometimes known as sentiment analysis or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral. It aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. Analysing such huge data for a person is nearly impossible and for doing that we are developing a system based analysis process which makes the analysis of the textual data provided by the user or public and analysing the data and finding the sentiment of the given arguments or comment.

The major contributions of this work are: the extended comparison of sentiment polarity classification methods for Twitter text; the inclusion of combination of classifiers in the compared set, and; the aggregation and use of a number of manually annotated tweets for the evaluation of the methods [7]. Especially regarding the latter, we consider it to be a main contribution in the sense that from past experience the automated annotation of tweets based on the detection of features like the emoticons ("☹", "😊", etc.) has been problematic since it does not always reflect the case about the overall sentiment expressed by the author, especially when one considers the expression of no-sentiment ("neutral") through the text.

The rest of this report is structured as such: Section 2, defines the problem of sentiment analysis. Section 3 provides details about the representation models that are commonly met in the literature. Section 4, provides details about the experiments that were conducted and the results. Finally, Section 5, highlights the main conclusions from this work and reports on possible future directions for research and experimentation.

II. PROBLEM FORMULATION

According to Pang and Lee Sentiment is “... given an opinionated piece of text, wherein it is assumed that the overall opinion in it about single issue 0 item, classify the opinion as falling under one of the two opposing sentiment polarities, or locate its position on the continuum between these two polarities.” Later there is room for third category of words which do not support both the positive and negative sentiment known as neutral. In this context the problem of document-level sentiment analysis is addressed. In this problem it is assumed that documents (in contrast to sentences or features) are opinionated regarding a particular topic. In the case of Twitter, the document is referred to as a "tweet" and it has a very specific form: a text message containing at most 140 characters.

The aim is to build a program which automatically reads the sentences and identify whether the author is expressing a positive sentiment or negative sentiment or a neutral sentiment on a particular topic. We need to convert the tweeter data in such a way that the algorithm will use it as input and classify the text's sentiment. Then we need to build a function which performs the operation of the

sentiment analysis[4]. The related work in representation models and classification algorithms that influenced this research is presented in the next section.

Definition of lexicon analysis :-

Lexical analysis is the first phase of a compiler. It takes the modified source code from language pre-processors that are written in the form of sentences. The lexical analyser breaks these syntaxes into a series of tokens, by removing any whitespace or comments in the source code.

If the lexical analyser finds a token invalid, it generates an error. The lexical analyser works closely with the syntax analyser. It reads character streams from the source code, checks for legal tokens, and passes the data to the syntax analyser when it demands.

Graph:-

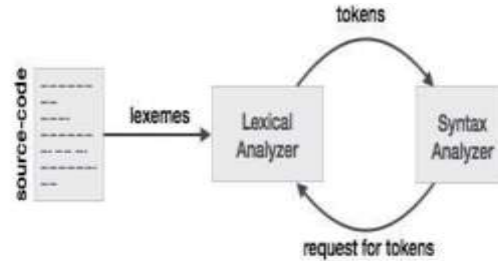


Figure: Lexical Graph

Tokens

Lexemes are said to be a sequence of characters (alphanumeric) in a token. There are some predefined rules for every lexeme to be identified as a valid token. These rules are defined by grammar rules, by means of a pattern. A pattern explains what can be a token, and these patterns are defined by means of regular expressions.

In programming language, keywords, constants, identifiers, strings, numbers, operators and punctuations symbols can be considered as tokens. [16]

Definition of naïve base classifier:-

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors[17]. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability
Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Above,

- $P(c/x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of class.
- $P(x/c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

Application of naïve bayes algorithm:-[17]

- **Real time Prediction:** Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi class Prediction:** This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
- **Recommendation System:** Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not

Definition of logistic regression:--

Command: Statistics
 └── Regression
 └── Logistic regression

Description

Logistic regression[18] is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of presence of the characteristic of interest:

$$\text{logit}(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$\text{odds} = \frac{p}{1-p} = \frac{\text{probability of presence of characteristic}}{\text{probability of absence of characteristic}}$$

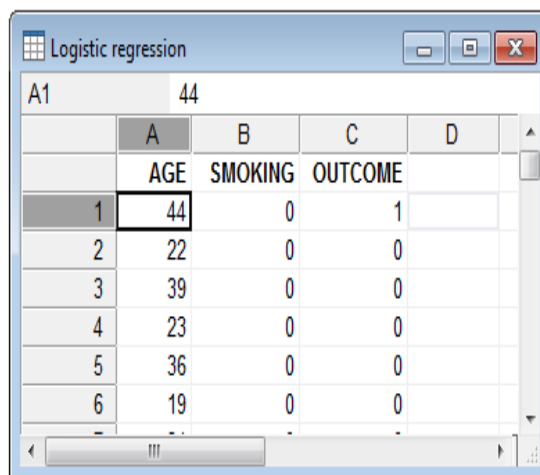
and

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

How to enter data

In the following example there are two predictor variables: AGE and SMOKING. The dependent variable, or response variable is OUTCOME. The dependent variable OUTCOME is coded 0 (negative) and 1 (positive).



	A	B	C	D
	AGE	SMOKING	OUTCOME	
1	44	0	1	
2	22	0	0	
3	39	0	0	
4	23	0	0	
5	36	0	0	
6	19	0	0	

Figure : Sample Dataset



Logistic regression dialog box configuration:

- Dependent variable: OUTCOME
- Independent variables: AGE, SMOKING
- Method: Enter
- Enter variable P/P: 0.05
- Remove variable P/P: 0.1
- Classifier table cutoff value: 0.5
- Graph: Show graph, History

Figure : Required input

III. RELATED WORK

The main problem when researching Sentiment Analysis problem is the translation of the textual data into a format that the computer can understand and process. For that a number of methods have been developed in the years, In this paper we a going to discuss the Bag of Words method.

A bag-of-words[15] model is a way of extracting features from text so the text input can be used with machine learning algorithms like neural networks. The bag-of-words model is one of the simplest language models used in NLP. It makes an unigram model of the text by keeping track of the number of occurrences of each word. This can later be used as features for Text Classifiers. In this bag-of-words model you only take individual words into account and give each word a specific subjectivity score. Later the scores are summed together and the sentiment of the given line is classified as a positive if the sum is more than or equal to 1 and if the sum is -1 or less then it is

classified as a negative number and is the sum reaches 0 that is considered to be a neutral sentiment. We represent the sentiment score in form of a graph.



Figure: Lexicon model

The next step is to analyze the pre-processed dataset using various machine learning algorithms and categorization tools. In this paper Naïve Bayes classifier, Logistic Regression, Lexicon and SVM[6] are examined; in isolation but also in combination.

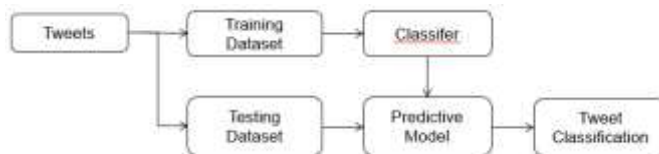


Figure: Learning Based approach

In Naïve Bayes method we use the probabilistic model in-order to maximize their accuracy with no consideration to the size of the produced tree. Furthermore, Logistic Regression and Support Vector Machines [2] try to find a mathematical function that can predict the correlation between the variables and the class. The Logistic Regression is using various logistic functions in order to calculate a function [3] with a graphical representation that has the minimum distance from each of the training data points in the multidimensional space. The Support Vector Machines [6] use other mathematical functions, such as the minimum square function, in order to create another function with a similar graphical representation.

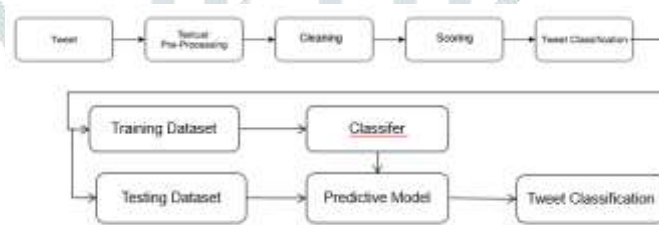


Figure: Combination of Lexicon and Naïve Bayes approach

Finally, the Multilayer Perceptron's are an implementation of artificial neural networks. They consist of a number of nodes, grouped in different fully interconnected layers. Each node is mapping its input values into a set of output values, using an internal function. This way each variable will be processed by one or more of node trees. The outputs of the first layer will become the inputs of the second and so on until at least one node of each layer has been activated. That way each one of the variables can contribute into the final classification decision with an intelligently calculated weight.

IV. EXPERIMENT

For the experiments we ran, we used a tweeter dataset of 1000 tweets, that were discussed on a different topic, from the healthcare system to everyday life and politics. These tweets were rated manually by number of researchers, according their sentiment polarity towards theirs subject. That way tweets were assigned to positive, negative and neutral categories which help to train the machine learning algorithms we use.

Firstly, we set the working directory for saving or for locating the R files. Later we include the libraries which are essential for the Sentiment Analysis like "twitterR", "ROAUTH", "RCurl", "plyr", "SentimentAnalysis", etc., the twitterR is used to communicate with the twitter application and perform a handshake with the application. The ROAuth is used authenticate the application by using the consumer key, consumer secret and access token and access secret for which say that you are a legal user of the tweeter. We get all the information like the following figure



Figure: Twitter App creation

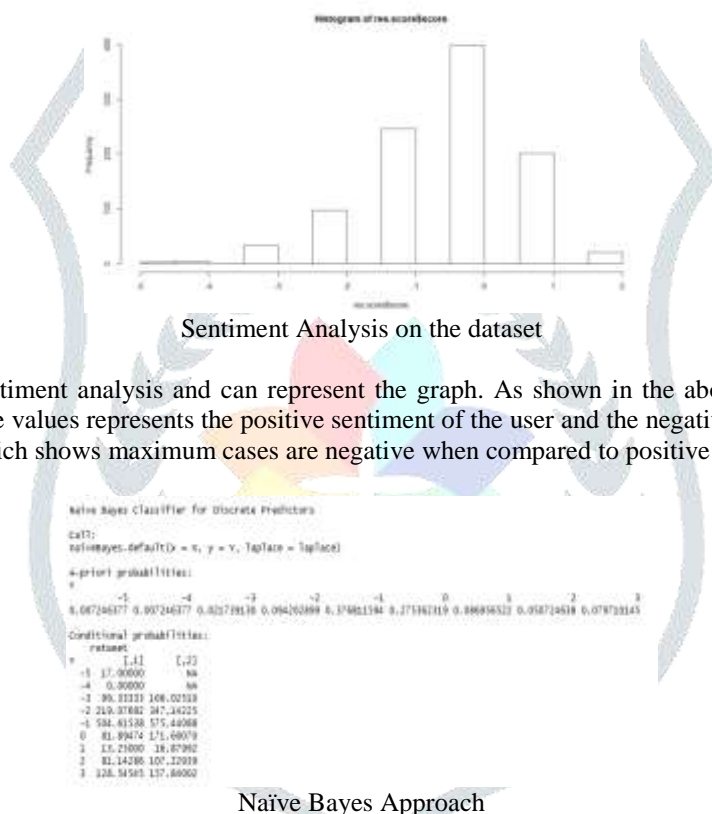
In the above page we click on the create new app and then we create the app and get the information as below



Figure: Application setting and Keys

Using this information we can get the data from twitter for the processing.

We need to connect API for that url from where to download data and we also assign the requestURL and accessURL and also authorizeURL to connect API and later on we setup the oauth and connect. We search a topic and get all the data into a vector and later we convert it to data frame. Later we clean the text which we get in the twitter search. Then we construct a function in which we perform the data analysis and represent it in a graphical representation shown below



In this way we can perform the sentiment analysis and can represent the graph. As shown in the above graph the 0 represent the neutral sentiment of the user and the positive values represents the positive sentiment of the user and the negative shows the vice-versa. By using our dataset, the above graph produce which shows maximum cases are negative when compared to positive and neutral sentiment.

As shown in the above figure the probability of sentiment to a particular tweet can be categorized according to the retweet count of a tweet and by applying the Naïve Bayes approach we categorized the probability of every sentiment score accordingly.

```

Call:
glm(formula = bvote ~ retweet, family = binomial, data = mysubdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8666 -0.8616 -0.5969 -0.1290  2.3967

Coefficients:
(Intercept) -0.785950  0.241321 -3.257  0.00113 **
retweet      -0.003420  0.001323 -2.584  0.00976 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 144.51  on 137  degrees of freedom
Residual deviance: 132.06  on 136  degrees of freedom
AIC: 136.06
    
```

Logistic regression Approach

When we are using the logistic regression we find that the regression process needs independent and incremental value to perform the regression, so we cannot perform the logistic regression on this particular data set.

V. CONCLUSION AND FUTURE WORK

In this paper we have discussed few very prominent methods which were used to perform Sentiment Analysis of Twitter. In this we have seen that the logistic regression has no much impact as it requires another incremental value to construct the graph it is not a better approach for the sentiment analysis, Naïve Bayes show superior or better results than the logistic and the combination of the Lexicon and Naïve Bayes show the better results because of the errors or the incorrect valuation in the Lexicon approach the Naïve Bayes shows some incorrect results which give the approach less accuracy compared to that of the Naïve Bayes approach. The work can be expanded in the future on linguistic, emoji and action using other machine learning approaches.

REFERENCES

- [1] Angel Cambero, Joe Geigel, A Comparative study of twitter Sentiment Analysis Methods for live Application, Aug 2016
- [2] Salazar, D.A., Vélez, J.I., Salazar, J.C., 2012. Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? Rev. Colomb. Estad. 35, 223–237.
- [3] John, G.H., Langley, P., 1995. Estimating Continuous Distributions in Bayesian Classifiers, in: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 338–345..
- [4] Medhat, W., Hassan, A., & Korashy, H. Sentiment Analysis algorithms and applications: A survey. AinShams Engineering Journal, 5(4), 1093-1113. 2014.
- [5] Mukherjee, S. Sentiment Analysis: A Literature Survey. Roll No: 10305061. Bombay, India. June, 2012
- [6] Mullen, T., Collier, N., 2004. Sentiment analysis using support vector machines with diverse information sources, in: In Proceedings of Conference on Empirical Methods in Natural Language Processing
- [7] Evangelos Psomakelis, Konstantinos Tserpes, Dimosthenis Anagnostopoulos, Comparing Methods For twitter Sentiment Analysis, Greece, Jan 2014.
- [8] Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 347–354. doi:10.3115/1220575.1220619
- [9] Lee, Lillian and Pang, Bo. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135. 2008.
- [10] Villena-Román Julio, Janine García-Morera, Miguel A. García Cumbreñas, Eugenio Martínez Cámara, M.Teresa Martín Valdivia, and L. Alfonso Ureña López, eds. 2015. *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*. CEUR WS Vol 139
- [11] Liu, Bing. 2012. *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1):1-167.
- [12] Turney, Peter D. 2002. *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 – ACL '02, 417, Philadelphia, Pennsylvania.
- [13] Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31, 102–107.
- [14] Amir, S., Ling, W., Astudillo, R., Martins, B., Silva, M. J., & Trancoso, I. (2015). Inesc-id: A regression model for large scale twitter sentiment lexicon induction. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 613–618). Denver, Colorado: Association for Computational Linguistics.
- [15] Godbole, N., Srinivasaiah, M., Skiena, S., 2007. Large-scale sentiment analysis for news and blogs. ICWSM'07.
- [16] TutorialPoint,[Online]:https://www.tutorialspoint.com/compiler_design/compiler_design_lexical_analysis.htm
- [17] AnalyticsVidya,[Online]:<https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
- [18] MEDCALC easy-to-use statistical software [Online]:https://www.medcalc.org/manual/logistic_regression.php