International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

Twitter Sentiment Analysis

Yasir Abdullah R[1], Bhargavi G[2], Priya K[3], Vasan S[4], Mohana Prasad P[5]

[1]*Professor , B.Tech Computer Science and Business Systems , Sri Krishna College Of Engineering and Technology, Coimbatore, TamilNadu, India.*

[2,3,4,5] *B.Tech Computer Science and Business Systems , Sri Krishna College Of Engineering and Technology, Coimbatore, TamilNadu, India*

Abstract - Opinion Mining also called as Sentiment analysis is the method of identifying and evaluating the emotions behind the combination of texts, which is used to understand the individuals opinions , emotions and attitudes delivered in an online platform. Social media like Twitter sentiment analysis will help to evaluate the feelings originated from social media posts. We can observe customer sentiment or interest towards the brand and evaluate the sentiment analysis score that is generated by some social media campaigns or online services. By knowing the users' emotions, we can get a better vision of their experience and so better customer service can be provided, which finally leads to a decrease in customer churn.The applications results in broad and powerful analysis. This ability to extract insights from social data widely adopted practice by organizations around the world. Natural language processing (NLP) technique is used in this analysis to determine whether emotion i.e,, data is negative, neutral or positive. Here, we provide a survey and the implementation details of Twitter data sentiment analysis using existing techniques like machine learning along with evaluation metrics using algorithm like Stochastic Gradient Descent (SGD), LGR, Multinomial Naive Bayes classifier (MNB).

Key Words: **Machine learning, Twitter , Natural Language Processing (NLP), Sentiment analysis (SA), Opinion mining, Multinomial Naive Bayes classifier (MNB).**

1. INTRODUCTION

Internet age has emerged as the tool for the people to express their opinions and views. Nowadays . millions of people are using online forums, blog posts in social media, like Twitter to unfold their ideas and emotions, and take into their daily lives. In this competitive world it is important to understand what people think and react in any fields like business , politics etc. For example , to develop their strategies, business and marketing field use some technique to understand how people react to their campaigns or product launches and why consumers are not interested in buying some. For instance, SA can be used to keep track and find consistent or inconsistent statements and actions at the political level. To picturize the general mood of the blogosphere , we can monitor and analyze social phenomena, and identifying very dangerous situations. Deriving emotional information from words techniques mainly relies on the following : process, search and analyze the facts present. Some textual contents like opinions, appraisals, sentiments and attitudes thus forming the base of Sentiment Analysis while facts have a subjective component but, there are which express subjective characteristics. Because of the tremendous increase in the availability of information or data on platforms or sources like online blogosphere and social networks these findings offer many challenging opportunities to develop new applications. For instance, by making use of this analysis some items can be predicted by considering opinions such as positive or negative in a recommendation system.

2. RELATED WORKS

In the paper published by Wei-Hai Lin et al [1] the authors have discussed sentiment analyzed classified data techniques. This article includes three major segments like emotion detection (ED) , learning transfer and resources building. Publication about the sentiment analysis using Naive-Bayes Strategy from Twitter Tweets proposed by Pablo Gamallo [2] had picturized the estimated work is straightly synchronized to opinion mining from textual data. In a publication Zhaopeng Tu et al[4] proposed document level sentiment analysis that measured various linguistic structures determined as tough observations for textual level of classifying the sentiments issue, comparatively to adapt syntactical structural units by not declaring linguistical facts explicitly. Kernels like Partial Tree (PT) and Subset Tree (SST) for reliance parse tress and component correspondingly are explored. Thus, by combining these kernels, gained a notable improvement of 1.55 point in accuracy.

3. METHODS AND MATERIALS

We created a web application using Python and a framework named Flask. The system have dashboard and a user registration system. To get Twitter text based on the entered keyword lively, the users can enter keywords which can be

further analyzed for one's feelings, emotions and sentiments. Graph visualization can be applied on this analyzed twitter data. In our project, we mined the data using Tweepy API. In real time , this Tweepy API will connect to Twitter and gets metadata together with the text from that platform. This technique helps to understand the user sentiment towards a particular product and help for growth. SA helps to analyze a tremendous amount of data, very fast and efficiently for system users. An easy-to-use library for Python is Tweepy. This can be installed in a Python IDE environment using pip commands. For pre-processing, we have used machine learning pre-existing libraries. We can visualize the collected data clearly and effectively by segregating customer emotions on a scale of -1 to 1, where 1 depicts a positive reaction strongly and -1 depicts a negative sentimentality strongly towards the keywords.

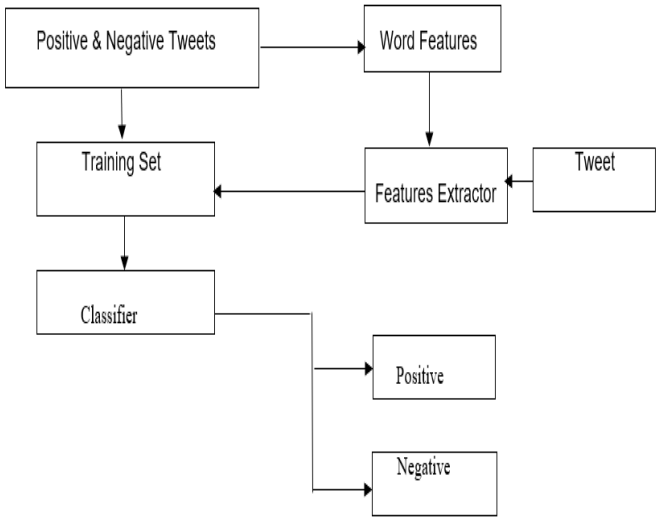

FIGURE 1: Architecture Diagram

For businesses to reach a broad audience , Twitter hosts three hundred and thirty million active users every month, to connect with customers without intermediate fields quickly and efficiently. On the flipside, it's hard for brands to quickly detect negative social media mentions affecting their business. This is why because sentiment data analysis includes tracking opinions in the words exchanged on social media platforms like twitter, have become an important technique in social media field. Listen and analyzing the way people react and exploring Twitter paves way for business wizards to understand their customers, to keep track of what's being said about their competitors, and their brand, and identify new trends in the industry. Thus quick actions can be taken by the companies. This technique could be extended to normal users, where hashtags and local other key texts could be used to analyze emotions from the words on social media platform. Tweepy is an open source wrapper written in python and that helps to twitter data using tokens we can read and write the tweets. To use Tweepy one has to sign in dev.twitter.com and create an application as a pre-requisite.

4. DESIGN SYSTEM

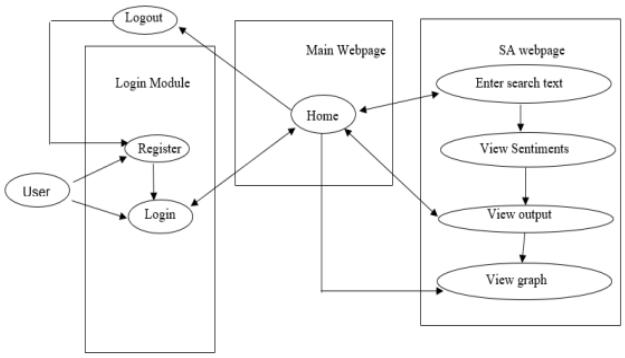

FIGURE 2: Design Of Twitter SA App

The dataset is first pre-processed to dissolve anomalies like redundancies and invalid records. Then it is applied to the machine learning algorithms and the resulting accuracies are analyzed. The following pre-trained models are used :

- Stochastic Gradient Descent
- Local Gaussian Regression
- Multinomial Naïve Bayes
- Random Forest Classifier

Stochastic Gradient Descent : SGD is a significant method for linear dataset used for the understanding of classifiers that are linear. It is a simple optimization algorithm. It uses convex loss functions with loss boundaries in machine learning applications. It fits well between the actual and predicted results. The goal was to find the optimal values for the intercept and the slope i.e., Predicted Height = intercept + slope * Weight. Then we can use the sum of the squared residuals as the Loss Function to determine how well that initial line fit the data. The sum of the squared residuals is just one of many different Loss Functions that can evaluate how well something fits the data. To find the optimal values for the intercept and slope , we plugged the equation for the Predicted Height into the sum of the squared residuals. Then we plugged the values from the observed data into the derivative with respect to the intercept and the initial guess for the slope 1, we did the math , plugged the slopes into the step size formulas and multiplied by the Learning rate , which we set to 0.01. We calculated the new intercept and new slope by plugging in the old intercept and the old slope along with the step sizes. Finally ending up with a new intercept and a new slope. The main advantage of this model is that it is easy to implement. The accuracy we got in our implementation is around only 60%.

Local Gaussian Regression : LGR is very efficient computationally which uses Gaussian progress regression. It fits regression models in small localized regions using only nearby data. These points contribute to the fit. Local lines of best fit from weighted least squares. Use the local line of best fit to predict at this x. The generalized addictive models so far focusses on non-linearity for a single predictor at a time. Sampling from a Gaussian process includes :

- Points x where we want to sample
- Compute covariance matrix X
- Can only obtain values at those points

Latent variables v,w drawn from a Gaussian Process , observations y are corrupted with noise and observations y are drawn from Gaussian process. Given the train data , we can visualize an infinite number of functions passing through them , a Gaussian Process can be seen as a distribution over these functions. Unlike traditional models , using Gaussian regression we can tell the confidence score related to a particular prediction.

Input: new data point $\{\mathbf{x}, y\}$.
for $k = 1$ **to** number of local models **do**
  Compute distance to the k-th local model:
    $w_k = \exp(-0.5(\mathbf{x} - \mathbf{c}_k)^T \mathbf{W}(\mathbf{x} - \mathbf{c}_k))$
end for
Take the nearest local model:
  $v = \max(w_k)$
if $v > w_{gen}$ **then**
  Insert $\{\mathbf{x}, y\}$ to nearest local model:
    $\mathbf{X}_{\text{new}} = [\mathbf{X}, \mathbf{x}]$
    $\mathbf{y}_{\text{new}} = [\mathbf{y}, y]$
  Update corresponding center:
    $\mathbf{c}_{new} = \text{mean}(\mathbf{X}_{new})$
  Compute inverse covariance matrix and prediction vector of local model:
    $\mathbf{K}_{\text{new}} = \mathbf{K}(\mathbf{X}_{\text{new}}, \mathbf{X}_{\text{new}})$
    $\boldsymbol{\alpha}_{\text{new}} = (\mathbf{K}_{\text{new}} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}_{\text{new}}$
else
  Create new model:
    $\mathbf{c}_{k+1} = \mathbf{x}$,
    $\mathbf{X}_{k+1} = [\mathbf{x}]$, $\mathbf{y}_{k+1} = [y]$
  Initialization new inverse covariance matrix and new prediction vector.
end if

FIGURE 3: LGR algorithm

The bias variance tradeoff here is the large span that contributes to smooth estimate and the small span contributes to very noisy estimate also for large span has more bias and less variance but small span has less bias and more variance. To be precise , a Gaussian process can be seen as a distribution over functions. These functions are seen as infinitely long vectors containing the value of the function at every input. Let the input space be X and f:X-> be a function fitting the train data. Then f is a Gaussian process if for any vector of inputs x such that X for all I , the vector of outputs f(x) is gaussian distributed. A gaussian process is specified by a mean function such that mean(x) is the mean of f(x) and a covariance/kernel function k : X * X -> R such that k(x1,x2) is the covariance between f(x1) and f(x2). One drawback of Gaussian processes is that it scales very badly with the number of observations. Bayesian Linear Regression (BLR) which can be seen as a special case of GP with the linear kernel , has complexity of only O(d^3) to find the mean weight vector , for a d dimensional input space. To make a prediction at any point , Gaussian process requires O(Nd) where d is the complexity of evaluating the kernel), while BLR only requires O(d) computation. Another drawback in most machine learning models is that the overfitting due to less ground truth data and high dimensional spectral data. The accuracy we got in our implementation is around only 57%.

Multinomial Naïve Bayes : It is a widely used Natural Language Processing method which calculates the text tags more accurately comparing the previous algorithms. It uses the Bayes theorem. This model tremendously reduces the complexity in the classification of text data. In our implementation , for analysis this model acts as the baseline Solution. It is mainly used because the resulting polarity possibly has three outcomes : positive , negative or neutral. For multinomially distributed information , multinomial naïve bayes is efficient and useful. For example , the probability we see the word “Dear , hey , friend , bad , great” in a textual data. We calculate the probabilities of discrete , individual words and not the probability of something continuous like weight or height , these probabilities are also called as Likelihoods. Also adding a count to each word did not change the number of messages in the training dataset. Multinomial Naïve Bayes ignores all the rules because keeping track of every single reasonable phrase in a language would be impossible. That said , even though Naïve Bayes is naïve , it tends to perform surprisingly well when separating normal text from emotional behind the texts. The accuracy we got in our implementation is around only 62%.

Random Forest Classifier : This is an efficient and significant model that uses during training time , multiple decision trees. The main advantage of this model is that it takes very little time for training than other models. It is also easy to implement and understand. Decision tree can be defined as a learning approach in machine learning that is

supervised. In businesses, it is used as a blackbox model. The working of this model is that from the training set, it selects random data points and then construct the decision tree with the subsets. It selects the n for each decision tree. Finally it evaluates the predictions based on the actual data points from the cluster that gains the most number of votes. The major advantage in random forest is that there is no overfitting because of the use of multiple trees reduce the risk of overfitting. And the training time is less. High accuracy is due that It runs efficiently on large database. For large data , it produces highly accurate predictions. Random forest can maintain accuracy when a large proportion of data is missing. We have to frame the conditions that split the data in such a way that the information gain is the highest. The accuracy we got in our implementation is around only 64%.

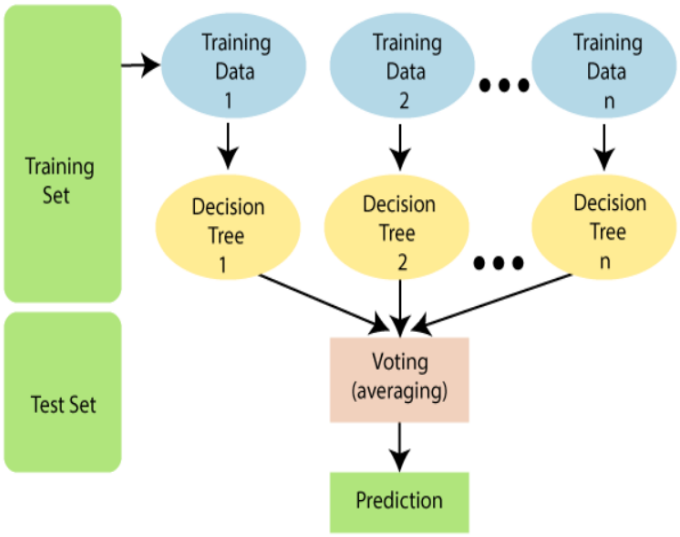

FIGURE 4: RFC model

4. CONCLUSION

Thus, we provided a comparison study of already available solutions for sentiment analysis along with ML and some metrics for evaluation in this paper. These results research show that machine learning methods together with cross domain and cross-lingual models like Support Vector Machine and NB has better accuracies and could be referenced as the core of understanding models, meanwhile other models are regularized stable in many points but requires some efforts in labelled dataset .We have explored and compared the results of different advantages on the classification segment. Thus, concluding that better results could be gained with more cleaner data. Using trained model provides better sentiment. The users can get information or data from online if it is a positive , or negative polarity or can be neutral. Eventually, reviews from online like social media platforms come under particular polarity. In future , a web application can be created that fetches the twitter data in real time and uses any of the mentioned models and the output predictions can be represented in graphical visualization like pie-charts.

REFERENCES

[1] Wei-Hai Lin et al, "Which Side are You on? Identifying Perspectives at the Document and Sentence Levels", In Proceedings of the Tenth Conference on Natural Language Learning (CoNLL'06).

[2] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-17.

[3] Ronen Feldman. 2013. Techniques and applications for sentiment analysis. Communications of the ACM, 56(4): 82-89.

[4] Zhaopeng Tu et al, "Identifying High-Impact Sub-Structures for Convolution Kernels in Document-level Sentiment Classification", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. July 2012.

[5] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, T. Wilson, Sem Eval-2013 Task2:Sentiment Analysis in Twitter (Vol.2,pp. 312-320 , 2013.