TEXT CATEGORIZATION ON TWITTER DATA

S.SUNIL KUMAR[1], B.PALLAVI[2], B.SHRUTHI[3], T.ROHINI[4]

ASSISTANT PROFESSOR[1], UG SCHOLAR[2,3&4]

DEPARTMENT OF INFORMATION TECHNOLOGY MALLA REDDY ENGINEERING COLLEGE FOR WOMEN (UGC-AUTONOMOUS) MAISAMMAGUDA, HYDERABAD-500100

ABSTRACT: Sentiment analysis is a classification problem where the main focus is to predict the polarity of words and then classify them into positive or negative sentiment. Two types of Classifiers are mainly used, namely lexicon-based and machine learning based. The first one include Senti WordNet and Word Sense Disambiguation while the second one include Multinomial Naive Bayes (MNB), Logistic Regression(LR), Support Vector Machine(SVM) and RNN Classifier. In this paper, existing datasets have been used, the first one from "Sentiment140" from Stanford University, consisting of 1.6 million tweets and the other one originally came from "Crowd flower's Data for Everyone library", consisting of 13870 entries, and both datasets are already categorized as per the sentiments expressed in them. Textblob, Senti wordnet, MNB, LR, SVM and RNN Classifier are applied on the above dataset and a comparison is drawn between the results obtained from above mentioned sentiment classifiers, classifying tweets according to the sentiment expressed in them, i.e. positive or negative. Also, along with the machine learning approaches, an ensemble form of MNB, LR and SVM has been performed on the datasets and compared with the above results. Further the above trained models can be used for sentiment prediction of a new data.

Index Terms – Sentiment Analysis, Classifiers, lexicon-based, machine learning, Multinomial Naive Bayes, Logistic Regression, Textblob, SentiWordNet.

INTRODUCTION

Sentiment Analysis which means to analyze the underlying emotions of a given text using Natural Language Processing (NLP) and other techniques to extract a significant pattern of information and features from a given large corpus of text. It analyses the sentiment and attitude of the author towards the topic of the subject mentioned in the text. This text can be a part of any document, post on social media or from any database source. Sentiments can be classified as objective or subjective, positive or negative or neutral. This classification can be either lexicon-based or machine learning based. Lexicon based classification makes use of already existing dictionary which has pre-assigned

scores to each word and those scores are used to calculate the overall sentiment expressed in the sentence whereas in machine learning based classification, a model is trained using some ML algorithm using some labeled data and then use that model to predict a class for a new text. Twitter is nowadays easily one of the most popular micro blogging platforms and millions of users express their views publicly on Twitter making it a rich source of information on public opinions and thus, helpful in sentiment analysis on any topic. Here, lexicon and machine learning based approaches have been incorporated to reveal the prevailing sentiments of tweets. Textblob, SentiWordNet and Word Sense Disambiguation are giving the correct sense of a word in a given context for Lexicon-based Sentiment analysis while among the machine learning based algorithms MNB, LR, SVM and RNN Classifier have been used. In this paper, a comparison has been presented in terms of accuracy in predicting the sentiment of a given tweet. An ensemble approach has also been implemented on the datasets which involve majority voting of MNB, LR and SVM and the results are being compared with the rest of the approaches.

EXISTING SYSTEM

Sentiment analysis is a classification problem where the main focus is to predict the polarity of words and then classify them into positive or negative sentiment. Classifiers used are of mainly two types, namely lexicon-based and machine learning based. The former include SentiWordNet and Word Sense Disambiguation while the latter include Multinomial Naive Bayes(MNB), Logistic Regression(LR), Support Vector Machine(SVM) and RNN Classifier.

PROPOSED SYSTEM

Lexicon and machine learning based approaches have been employed to reveal the prevailing sentiments of tweets. Textblob, SentiWordNet and Word Sense Disambiguation are giving the correct sense of a word in a given context for Lexicon-based Sentiment analysis while among the machine learning based algorithms MNB, LR, SVM and RNN Classifier have been used. The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well-ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are:

Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

RELATED WORK

Sentiment analysis is an application of natural language processing. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. It helps in human decision making. To perform sentiment analysis, one has to perform various tasks like subjectivity detection, sentiment classification, aspect term

extraction, feature extraction etc. This paper presents the survey of main approaches used for sentiment classification.

The wide spread of World Wide Web has brought a new way of expressing the sentiments of individuals. It is also a medium with a huge amount of information where users can view the opinion of other users that are classified into different sentiment classes and are increasingly growing as a key factor in decision making. This paper contributes to the sentiment analysis for customers' review classification which is helpful to analyze the information in the form of the number of tweets where opinions are highly unstructured and are either positive or negative, or somewhere in between of these two. For this we first preprocessed the dataset, after that extracted the adjective from the dataset that have some meaning which is called feature vector, then selected the feature vector list and thereafter applied machine learning based classification algorithms namely: Naive Bayes, Maximum entropy and SVM along with the Semantic Orientation based WordNet which extracts synonyms and similarity for the content feature. Finally we measured the performance of classifier in terms of recall, precision and accuracy.

One of the most important parts of running business successfully is analyzing customer's opinion and sentiments. In this paper, the paragraph of sentences given by the customer is accepted and after extracting each and every word, they are checked with the stored (database has been maintained here) parts of speech, articles and negative words. After checking against the database, CFG is used to validate proper formation of the sentences. Each sentences are delimited by `.' or `?' or `!'. Emotions are detected as - positive, negative or neutral sentence. There are 3 types of cases:

- If the paragraph contains more positive sentences than negative, then overall result will be positive.
- If the number of negative sentence is greater than positive sentence, then the overall result is negative.
- If there are same numbers of positive and negative sentences in the input paragraph, then the result is neutral and if a sentence has been entered that is a normal statement neither positive nor negative, that will be also considered as neutral.

METHODOLOGY

Two algorithms are used here. They are:

- Linear regressor

- Decision tree regressor

Linear regressor:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable The linear regression model provides a sloped straight line representing the relationship between the variables.

linear Regression

Mathematically, we can represent a linear regression as:

$y= a0+a1x+ \varepsilon$

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value). ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Decision tree regressor:

Decision Tree algorithm has become one of the most used machine learning algorithm both in competitions like Kaggle as well as in business environment. Decision Tree can be used both in classification and regression problem. A decision tree is arriving at an estimate by asking a series of questions to the data, each question narrowing our possible values until the model get confident enough to make a single prediction. The order of the question as well as their content is being determined by the model. In addition, the questions asked are all in a True/False form. This is a little tough to grasp because it is not how humans naturally think, and perhaps the best way to show this difference is to create a real decision tree from. In the above problem x1, x2 are two features which allow us to make predictions for the target variable y by asking True/False questions. For each True and False answer there are separate

branches. No matter the answers to the questions, we eventually reach a prediction (leaf node). Start at the root node at the top and progress through the tree answering the questions along the way. So given any pair of X1, X2. One aspect of the decision tree I should mention is how it actually learns (how the 'questions' are formed and how the thresholds are set). As a supervised machine learning model, a decision tree learns to map data to outputs in what is called the training phase of model building.

MODULES

User: Users to determine whether a product, service, news, article, etc. is generating positive, neutral or negative responses. Not just the polarity, but also the depth of the feeling towards the particular service or product is taken into account. Alternatively, Text classification also helps the consumers get a better idea of the pros and cons of the product or service.

Admin: The aim of admin is to approve the machine learning users. The entire data must be gathered to admin. Multinomial Naive Bayes Classifier uses Bayes Theorem to predict the probability of a given set of features belonging to a particular class label. It uses the assumption that the probability of different events is independent of each other. SVM is a classification technique which tries to find the most optimal hyper plane with the maximum margin, between the classes, that separates them in space

CONCLUSION

Various techniques for both lexicon-based and machine learning based, have been applied in this project and the results are compared. It has been observed that for a totally new data/text machine learning based models trained over a related data are much more accurate than the classification based on standard dictionaries. This is because of the fact that the text that's being observed i.e., the tweets are highly informal and do not use the standard grammar rules or the spelling and thus the data here is highly unstructured. The comparison results can be clearly observed among different machine learning algorithms also. As of now, among the algorithms used, RNN is observed to have the highest accuracy.

REFERENCES

[1] Dr. Priyanka Harjule, Astha Gurjar, Harshita Seth, Priya Thakur, "Text Classification on Twitter Data",978-1-7281-1683-9/20/$31.00 ©2020

[2] A. Weiler, M. Grossniklaus, M. H. Scholl et al., "Survey and experimental analysis of event detection techniques for twitter," The Computer Journal, vol. 60, no. 3, pp. 329–346, 2017.

[3] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Sentiment analysis for modern standard Arabic and colloquial," 2015.

[4] O. Loyola-González, A. López-Cuevas, M. A. Medina-Pérez et al., "Fusing pattern discovery and visual analytics approaches in tweet propagation," Information Fusion, vol. 46, pp. 91–101, 2018.

[5] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari. "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", International Journal of Information Engineering and Electronic Business, 2016.

[6] P.Kalaivani, "Sentiment Classification of Movie Reviews by supervised machine learning approaches" Indian Journal of Computer Science and Engineering (IJCSE) ISSN: 0976–5166 Vol. 4 №4 Aug-Sep 2013.