RESEARCH ARTICLE OPEN ACCESS

Analysing Sentiments Using Twitter Data

Sruthimol E.K[1],Alfiya M.A[2],Soofiya V.H[3],Thasni T.A[4],Jisha Jamal[5]
Student [1][2][3][4], Assistant Professor [5]
Dept of Computer science
KMEA Engineering College, Kerala

ABSTRACT

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text. Sentiment analysis is also known as opinion mining,is a field with natural language processing that builds system that try to identify and express opinions within the text. In this paper we are using twitter data for analyzing sentiments.Tweets are collected from twitter using twitter API, analyses and classifies these tweets. Naive bayes classifier is used for classifying tweets in order to obtain the result as positive, negative or neutral.

Keywords :— Sentiment analysis,naïve bayes,polarity,tokens,unigram,chi-square

I. INTRODUCTION

Sentiment analysis is the automated process of understanding an opinion about a given subject from written language. Currently, sentiment analysis is a topic of great interest and development since it has many practical applications. Since publically and privately available information over internet is constantly growing, a large number of text expressing opinions are available in review sites, forums,blogs and social media. Sentiment analysis is a branch of machine learning which a subset of Artificial Intelligence is also. With the help of sentiment analysis system, unstructured information could be automatically transformed into structured data of public opinions about products, services, brands, politics, or any topic that people can express opinions about. This data can be very useful for commercial applications like marketing analysis, public relations, product reviews, net promoter scoring, and product feedback customer service.

II. LITERATURE SURVEY

[1]Proposed a system for a sentence level sentiment analysis. It includes three phases in which the first phase is caption of dataset using tweets of electronic products. The second phase is pre-processing of tweets, Feature vector is created using relevant features for classification and finally they classifies the tweets as positives and negatives by using different classifiers such as naïve Bayes classifier,svm classifier, maximum entropy classifier and ensemble classifier.

[2]Specifies a classifier model for sentiment analysis. Data retrieval module is used for retrieving tweets from twitter.The pre-processed tweets are segregated domain wise.Tweet classification module uses machine learning techniques to classify tweets as positive and negative. .[3]Describes machine learning techniques such as support vector machine model, naive bayes classification model, maximum entropy model and artificial neural network model. [4]Sentiment of uber is analysed during the period of July 2016 and July 2017 from Facebook users.Analysis is done using machine learning techniques.[5]Twitter messages are retrieved in real using streaming API.Streaming API requires a persistent HTTP connection and authentication.Streaming API offers possibility of filtering tweets according to several categories such as location,language,hashtags or words in tweets.

III PROPOSED SYSTEM

We are proposing a system that determines the ratio of polarity of different kinds of sentiments in a group of text. Sentiment analysis is completely based on using text classification technique to determine document level or sentence level polarity of sentiments. Polarity of sentiments can be achieved by using naïve bayes classifier.

A. Sentiment Analysis Process

In this process, the algorithm needs to be trained with var ious data sets that shows it what a positive sentiment is, what a negative one is and what a neutral sentiment looks like.This training data sets can be in a plain text file, the algorithm can recognize and understand the data formats.This can be done by reading a text file, splitting the sentences into words and temporarily storing them in an array.After the training sets are provided in distinct arrays, sentiment analysis can be performed using the Naïve Bayes Algorithm to score a sentence based on the distribution of the words in it.

B. Tweet collection

The API requires us to register a developer account with Twitter and fill in parameters such as consumer Key, consumer Secret, access Token access, and Token Secret. This API allows getting all random tweets by using keywords. API supports to retrieve tweets which match a specific criterion defined by the developer. We used this to retrieve tweets related to specific keywords which are taken as input from users.

C. Processing of tweets

Tweet processing involves tokenization which is the process of splitting the tweets into individual words called tokens.

Tokens can be split using whitespace or punctuation characters. It can be unigram or bigram depending on the classification model used.In our project we use unigram features. It is just a collection of individual words in the text to be classified, so, we split each tweet using whitespace.

D. Filtering of data

A tweet acquired after processing has a portion of raw information which may or may not find useful for our application. Thus, these tweets are further filtered by removing stop words, numbers and punctuations.

a.Stop words:Tweets contain stop words which are extremely common words like “is”, “am”, “are” and holds no additional information.

b.Non-alphabetical characters: In some tweets there will be symbols like “#@” and numbers hold no relevance in case of sentiment analysis and these symbols needs to be removed using pattern matching.

c.Stemming: It is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

d.Feature Selection: The word scores of the features are tested based on Chi-square method. It creates a list of all positive and negative words. Then build frequency distribution of all words and then frequency distribution of words within positive and negative labels. Finally, the number of positive and negative words as well as the total number of words and the dictionary of word score based on Chi-Square test is found. This test method gives good result for both positive and negative classes and it is also used to select feature from high dimensional data. So that word scores are found and the best number of words based on word scores are also extracted.

IV. CLASSIFIER

Naïve Bayes classifier is the classifier used for classifying into positive, negative or neutral. It is used because of its easiness in both during training and classifying steps. Preprocessed data is given as input to train input set using Naïve Bayes classifier and that trained model is applied on test to generate either positive or negative sentiment. The Bayes theorem is as follows.

$$\frac{H}{P(X)} = \frac{\;\frac{H}{P(X)\,P(H)}\;}{P(X)} \qquad (1)$$

where X- Tuples, H-Hypothesis, P(H|X) represents Posterior probability of H conditioned on X i.e. the Probability that Hypothesis holds true given the value of X, P(H) represents Prior probability of H i.e the Probability that H holds true irrespective of the tuple values, P(X|H) represents posterior probability of X conditioned on H i.e. the Probability that X will have certain values for a given Hypothesis, P(X) represents Prior probability of X i.e the Probability that X will have certain values. The proposed system understands whether the tweet is positive or negative based on the dictionary methods of score. An experiment result of accuracy is evaluated using following information retrieval matrices. Accuracy is the performance evaluation parameter and it is calculated by number of correctly selected positive and negative words divide by total number of words present in the corpus. The formula is given as below

$$\text{Accuracy}=\frac{\sum\text{True Positive}+\sum\text{True Negative}}{\sum\text{Total number of words}} \qquad (2)$$

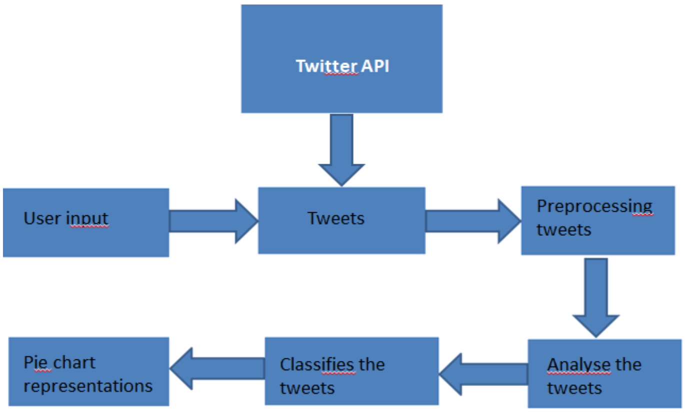

FIG 1: STAGES IN PROPOSED SYSTEM

V. RESULTS AND COMPARISON

In the proposed system, PHP is used for implementation. The twitter data programmatically can be accessed by creating an application in twitter that interacts with the Twitter API. In order to search for particular tweets, Oauth protocol is used for authentication. The first step is to login the Twitter and register a new application. The name and a description for the app are chosen. From this, a consumer key and a consumer secret are received. These are the application settings that should always be kept private.

From the configuration page of the app, an access token and an access token secret are also required. Similarly to the consumer keys, these strings must also be kept private. They provide the application access to Twitter on behalf of the user account.Using these four keys, the only text data from the twitter are filtered based on the keyword in the particular location and languages.

A number of tweets is collected based on the search from user. The content of a tweet is embedded in the text to analyze by breaking the text down into words. Tokenization is used to split a stream of text into smaller units called tokens.The preprocessed tweets are appended for feature selection process in each list.Naïve Bayes classifier is used for classifying dataset to positive,negative,neutral.This is represented graphically using pie chart.

VI. CONCLUSION

In this paper we discussed about analyzing tweets.Each word represents a token.These tokens are classified using naïve bayes classifier. Sentiment analysis have wide range of scope in different levels.For example, document level: to obtain sentiment of entire document and sentence level: to get sentiment of a single sentence.In this paper we focused on both document as well as sentence level.We tentatively conclude that sentiment analysis for Twitter data is not that different from sentiment analysis for other genres. In future work, we will explore even more in sentiment analysis.

REFERENCES

[1] Neethu M S & Rajasree R, “Sentiment analysis in twitter using machine learning techniques,”.4th ICCCNT 2013 July 4 - 6, 2013, Tiruchengode, India

[2] Manju Venugopalan & Deepa Gupta, “Exploring sentiment analysis on twitter data ,” in Proceedings of the 11th International Conference on Electronic Publishing, pp. 349–360, 2007.

[3] Harpreet Kaur,Veeenu Mangat & Nidhi“A survey on sentiment analysis technique,”978-1-5090-3243-3/17/$31.00 ©2017 IEEE

[4] Boiy, Erik, and M. F. Moens. "A machine learning approach to sentiment analysis in multilingual Web texts." Information Retrieval Journal12.5(2009):526-558.

[5] Davidov, Dmitry, O. Tsur, and A. Rappoport. "Semi-supervised recognition of sarcastic sentences in twitter and amazon." Conll(2010):107-116.

[6] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani,Veselin Stoyanov, SemEval-2016 Task 4: Sentiment Analysis in Twitter.

[7] Gamallo, Pablo, and M. Garcia. "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets." International Workshop on Semantic Evaluation 2014:171-175.

[8] Kang, Hanhoon, S. J. Yoo, and D. Han. "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews."Expert Systems with Applications An International Journal39.5(2012):6000-6010.