(IJRST) 2018, Vol. No. 8, Issue No. II, Apr-Jun **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

ANALYSIS OF TWITTER DATASET TO ENHANCE ACCURACY BY EMPLOYING KNN

Jenifer Ghai

ABSTRACT

Social networks impacts day to day activities of human behaviour. Lots of data are generated in social networks. Managing such data in the form of classification, clustering and maintaining that data is a critical issues faced by data base managers. Twitter is one kind of social network used by internet users. In this work, the sentiments of tweets are classified using Senti value and sentiment. The performance of Senti value is appreciable than sentiment. The sent value has only 3 classes but sentiment has 13 classes. The experimental results show that the classes with numerical classification produces more accurate results than text classes.

INTRODUCTION

Social networks such as blogs, twitter, face book, what's app and various other medias are used to produce a large amount of data in the form of text, image and other numerical formats on the World Wide Web[1]. The large amount of data used in this network provides services in various forms such as behavioural analysis, weather analysis, market profit and loss prediction, disease prediction, and classification of customer, commercial and scientific dealings. Sentiment analysis is a recent research topic of internet and social media users who posts their feelings, emotions, predictions, decisions, analysis and behaviours. The process of collecting reviews, opinions, tweets and emotions about a particular topic is collected for analysis and information drawn on this topic. The sentiments are classified as positive, negative and neutral. Positive tweets will create happy environment. The negative feelings or emotions is considered to be serious and direction is needed to handle such a feeling. The main aim of this paper is to classify tweets based on the sentiment [9] and senti value using knn algorithm. Two approaches for sentiment analysis was performed. First one, using sentivalue and the other one with sentiment using knn. K-Nearest Neighbors(KNN)[2] is a supervised machine learning algorithm. Based on the classes by the knn algorithm nearest points is classified. The majority of voting class is determined by the value of k. However to improve this algorithm weights are assigned to each of the k points according to their distance from the test point. This research paper is organised in to providing literature review in chapter2, methodology followed was described in section 3, the experimental results are analyzed in chapter 4 and chapter 5 concludes the work.

REVIEW OF LITERATURE

Different algorithms and models performed by the author to solve polarity, opinion and product reviews using dual sentiment analysis [3]. Presented 4 way sentiment classification approach on Arabic social sentiments data. The author collected 10, 000 tweets from twitter data set and performed experiments. The classification results are based on the objective, subjective positive, subjective negative, and subjective mixed [4]. Twitter data set is collected from twitter API. The cleaning process was carried out by removing the stop words and classified the tweets into positive and negative. The word cloud is generated to predict the current public emotions [5]. The sentiment classification models using svm, Twitter-specific lexicon and DAN2 machine learning approach was developed by the

author and concluded that dan2 produces more accurate results than svm [6]. The author performed sentiment analysis to handle large volume of data using hadoop on a hadoop cluster faster in real time [7]. This work is to observe the effect of pre-processing on twitter data of sentiment classification. The n-gram method was used to find the classification which clearly indicates the improvements in accuracy of classification [8].

METHODOLOGY

KNN is a simple Classification algorithm, the output is determined by the class with the highest frequency. The class with the most choice is taken as the prediction value. The k value is chosen to be odd when the number of classes is odd to avoid tie. The algorithm steps are

1. Calculate the Euclidean distance between the points.

2. Arrange distances in increasing order.

3. Choose k be a positive integer, pick the first k distances from this order.

4. Locate those k-points related to these k distances.

5. Let ki denotes the number of points belonging to the ith class among k points i.e. $k \geq 0$

6. If ki >kj $\forall$ i $\neq$ j then put x in class i. Two kinds of classification are performed.

The first classification is based on the sentivalue. Each tweet in the twitter is assigned a value 0(neutral), 2(negative) and 4(positive). The second classification is performed with sentiment attribute. The levels of sentiment in twitter dataset is anger, boredom, empty, enthusiasm, fun, happiness, hate, love, neutral, relief, sadness, surprise and worry. The number of times the tweets occurred in the twitter data set is given in table 1. The levels of sentivalue attribute in this data set is 0, 2 and 4. The occurrence of frequency of tweets is shown in figure 1.

Tweets	Frequency
anger	3
boredom	6
empty	19
enthusiasm	17
fun	18
happiness	39
hate	48
love	39
neutral	230
relief	11
worry	280
sadness	238
surprise	52

Table 1. Frequency of tweets

(IJRST) 2018, Vol. No. 8, Issue No. II, Apr-Jun e-ISSN: 2249-0604, p-ISSN: 2454-180X

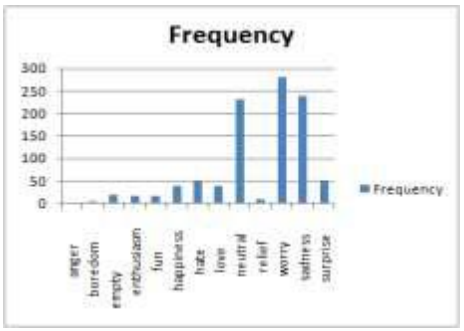

Figure 1. Frequency of tweets

EXPERIMENTAL RESULTS

The data set is divided into training set and test set. Training set dimension is 701 with 5 attributes. Testing set contains 299 samples with the same 5 attributes. The twitter data set is downloaded from the internet and pre-processing was performed. The feature selection algorithm was applied on this dataset to identify the important attribute. Then the knn algorithm was implemented for classification. There are 701 samples with 4 predictors and 3 classes. The classes are 0 denotes neutral, 2 represents negative and 4 is positive. The training data set is used to implement knn algorithm and performance was analyzed. The repeatedcv method in the train control is repeated for 3 times with 10 fold cross validation. The Summary of sample sizes: 632, 631, 630, 631, 630, 630. The Resampling results across tuning parameters is given in table 2.

K	Accuracy	Kappa
5	0.3240059	0.01692654
7	0.4634009	0.25456321
9	0.6844772	0.38179620
11	0.7227341	0.36703561
13	0.8126190	0.60310613
15	0.7414471	0.42536492
17	0.6909569	0.28371246
19	0.6523044	0.17717036
21	0.5891942	0.00000000
23	0.5891942	0.00000000

Table 2. Resampling results

Accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 13. The accuracy is 81% and is shown in figure 2.

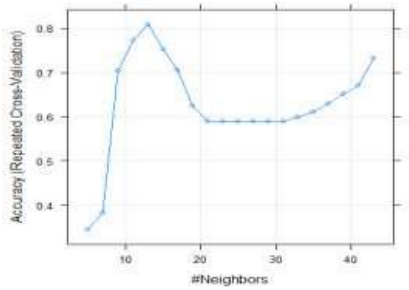

Figure 2. senti value performance

The confusion matrix of test prediction of sent value with 3 classes 0,2,4 are shown in table 3.

Prediction	Class 0	Class 2	Class4
0	74	1	0
2	8	164	25
4	0	0	27

Table 3. Confusion matrix

The overall statistics of accuracy is 88% and kappa is 79% using sentivalue as a classification attribute. The next classification was performed with same training set of 701 samples, 4 predictors and 13 classes, 'anger', 'boredom', 'empty', 'enthusiasm', 'fun', 'happiness', 'hate', 'love', 'neutral', 'relief', 'sadness', 'surprise', and 'worry'. The bootstrapped resampling was applied in this dataset. The Summary of sample size is 701, resampling results across tuning parameters.

Accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 5 with accuracy 23%. In order to improve the accuracy, the bootstrapped resampling method is applied with same samples of tune length 15 . The Summary of sample sizes are 701, 701, 701, 701, 701, 701, . The resampling results across tuning parameters. Accuracy was used to select the optimal model using the largest value. The final value used for the model was k = 25.

CONCLUSION

From the experimental results, the accuracy of sentiment is 36% because this attribute is text based. Hence the knn algorithm classifies into 13 classes. The feature selection algorithm ranked sentivalue as the first attribute for classification. But sentiment is a second ranked attribute, therefore it produces 36% . The tuning is again needed for such work.