CLUSTERING AND CLASSIFICATION BASED APPROACH FOR EMOTION ANALYSIS IN ONLINE SOCIAL NETWORK DATA

[1]T.SURESHKUMAR,[2]G.VADIVEL

[1]Research scholar,[2]Assistant professor in BCA department
[1]M.phil computer science,
[1]SNMV college of arts and science, Coimbatore,India

Abstract : Social media plays a significant role in explore opinions and emotions of the users based on the day-to-day activities. The data mining techniques for social media data analysis and emotion mining based analysis are ranging from unsupervised to semi supervised and supervised learning methods. Mining of social network data about user's opinion and emotion is necessary to understand the user's behaviour and mentality. This research work proposed hybrid data mining approach using K-Means clustering and Naive Bayes classification techniques to analyse the emotions in the tweets. In the reprocessing phase, K-Means clustering process performed in the tweet emotion data set using Euclidean distance as distance function. The clustered data classified in the classification phase using Naïve Bayes classifier with 10 fold cross validation. Emotion type and cluster attributes used as the class variables to classify the clustered tweet emotion data. The experimental result shows Naïve Bayes classification in cluster data produces higher accuracy than the classification of tweet emotion data set without clustering.
Keywords: Emotions, Hybrid Data mining approach, K-Means, Naïve Bayes, Social network data

I. INTRODUCTION

Data mining is considered as the process of extracting significant patterns from a given database and it always act as a valuable tool for converting data into usable information. Data mining classification and clustering techniques can be used in the variety of areas like marketing field, banking sector, educational research, surveillance, telecommunications fraud detection, and scientific discovery. The social media data mining is one of these domains in which the primary concern is the evaluation and, in turn, enhancement of social media related services based on social media domain.

Basically data mining is a five-step process. It consists of Identifying the information source, selection of data points that are need to be analyse, extraction of patterns from the data, identify the key information from the extracted patterns or results and finally interpreting and reporting the results.

Data mining is one among the steps of Knowledge Discovery in Databases (KDD) process. KDD is a multi-step process that includes selection of data and pre-processing, transformation of the data, clustering and classification, extraction of patterns and knowledge discovery for conversion of data to useful information. Data mining used to extract the patterns from the transformed data in KDD.

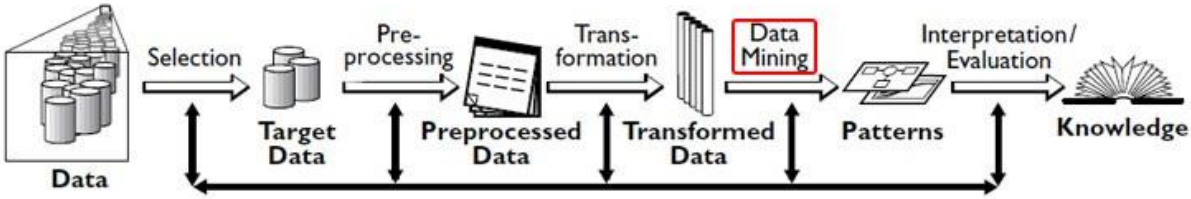


Fig.1. KDD Process

1.1. SOCIAL MEDIA DATA ANALYSIS

Social media is an internet based communication tool that empowers people to share information. To understand better the term social media, social indicates associating with people and spending time in order to develop their relationships whereas media indicates tool for communication such as internet, TV, radio, newspaper so on, here our focus is internet. Social media is stated as an electronic platform for socializing people. Some example of social media sites are Facebook, Twitter, YouTube, LinkedIn, Digg etc. Initially people involved in social media to associate with their friends and lost friends, gradually they improved to the status of updating and consuming any information on social media, these led to vast generation of user data which could be further processed for future development.

Social network portal allows the users to interact with others through post comments, text messages, images and videos. Nodes and Links are used to represent data in the social media's data graph. Nodes denote entities like friends, relatives, etc., and links denotes the relationship among the users.

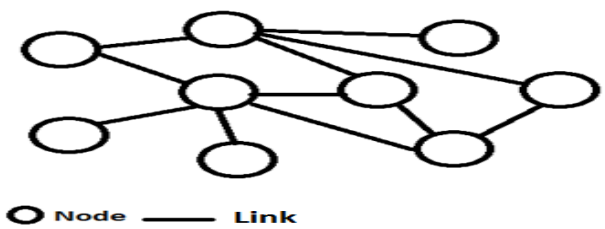


Fig.2. Links and nodes in social media

1.2. DATA MINING TECHNIQUES

The data mining techniques for social media analysis are generally classified in the categories of link analysis, change and deviation detection, sequential pattern mining, tracking patterns, classification, association, outlier detection, clustering, regression and prediction.During the past few years, social media has seen tremendous growth in the user's count. For example, there are more than 336 million active members belonging to the Twitter network. Generally social media can be classified into social networking web sites, social book marking sites, video sharing portals, photo sharing portals, Wikis, Blogs and Micro blogs and Reviews and Ratings based community sites. Twitter is a one of the example for micro blogs.For the social media analysis, three types of learning models are used for classification; they are supervised learning model, semi-supervised learning model and unsupervised learning model. Sentiment analysis in used for opinion mining process in the social media data like tweets. It helps to analyse and monitor the social phenomena to determine the mood and mind-set of the users. Most of the social media contents are available in the form of unstructured data. Sentiment analysis performs pre-processing, extraction of sentiments from the tweets, classify the extracted sentiments based on sentiments and subjectivity like emotion type and summarization of the opinions in the tweets. In the pre-processing phase, removes the symbols, numbers and stop words, replace the emotions with their sentiments, remove the non-English words and expand the acronyms.The sentiment analysis performed data set is taken to this research from the github data set repository portal and perform the clustering and classification process to analyse the performance based on the emotion type.In this research work, K-Means clustering is used to cluster the tweet_emotion data set and Naïve Bayes classifier is used to classify the clustered data based on the emotion types. Clustering and classification process in the tweet_emotion data set are to be explained in third chapter.

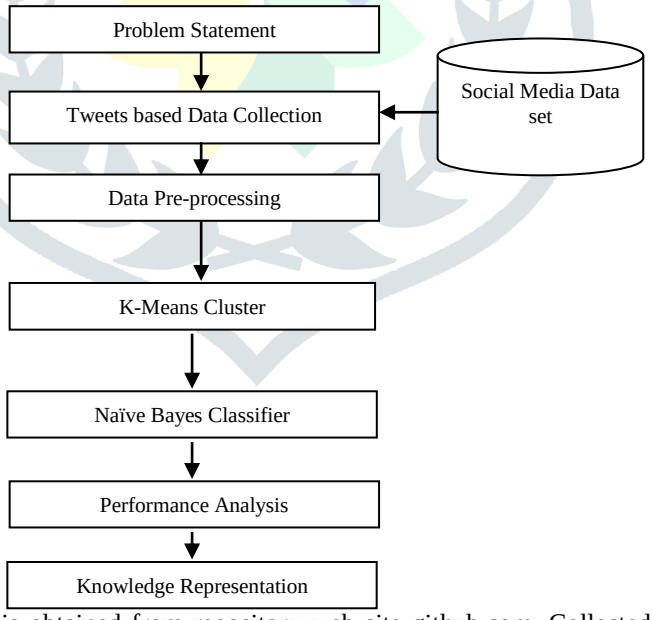
II. EXISTING WORKS

The existing works reveals that the sentiment analysis and opinion mining based process in unstructured data from the social media plays vital role to identify the behaviour of the user. Most of the approaches deals with extraction of features from the tweets or posts in the social media wall and perform the sentiment analysis. But emotion based classifications are not performed in the existing research works. This research work focus the emotion type based classification and Cluster based classification process performed in the tweet emotion data set using Naïve Bayes classifier and analyse the performance of the hybrid data mining approach using performance measures.

III. DATA MINING IN SOCIAL MEDIA DATA USING CLASSIFICATION

Social media based data mining refers to extracting or "mining" knowledge from large amounts of unstructured social media data set using text mining. Currently, the tweet based data are stored in the tweet_emotion excel file in the format CSV, these data set contain the useful information to cluster and classify the emotions. In this work, the K-Means used to cluster the data in the pre-processing phase and Clustering and Naïve Bayes classifiers is used to classify the data based on the cluster and also sentiment. The following figure 3.1 represents working methodology in the tweet_emotion dataset based on the framework. The work methodology begins with problem definition, data collection from github and data pre-processing that includes group the data into 13 clusters, classify the data set based on sentiment and also cluster using Naïve Bayes classifier, analyse the performance and discovering knowledge.

Fig. 3.1 Data mining work methodology

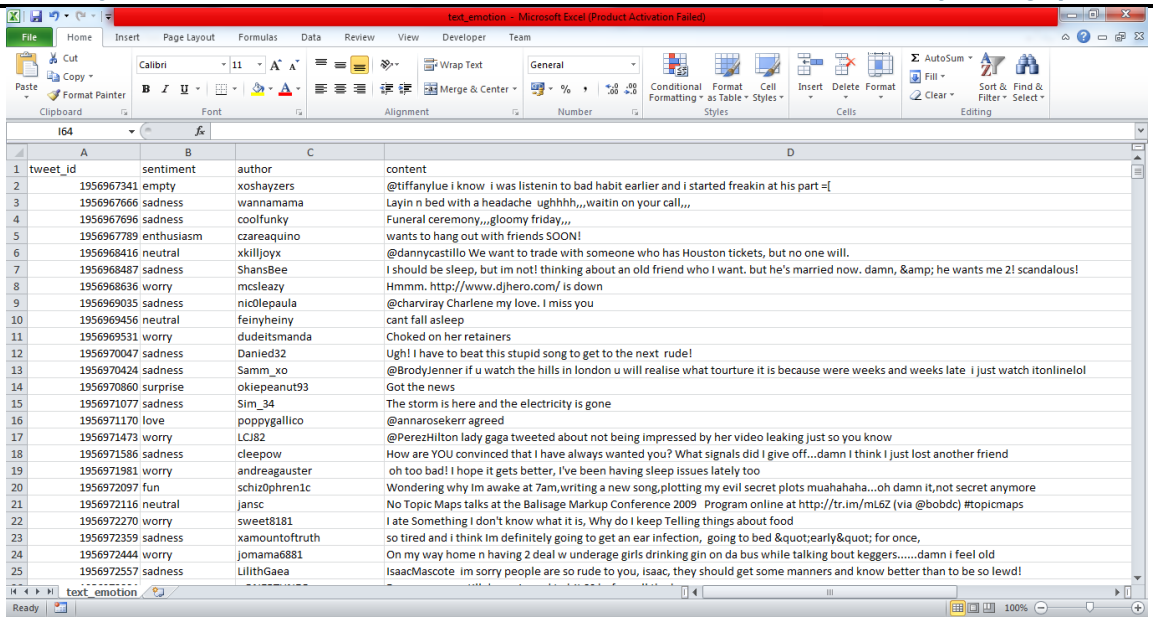


3.1.1. DATA COLLECTION

The data set used in this study is obtained from repository web site github.com. Collected data set consists of 39985 instances with tweet id, sentiment, author and content posted by author attributes.

Collected data set consists of contextual meaning of the tweets posted by the author based on the subjective meaning extracted from the source content. Emotion type is the subjective meaning of the content. Anger, boredom, empty, enthusiasm, fun, happiness, hate, love, neutral, relief, sadness, surprise and worry are the sentiment category based on the content posted by the author.

Total size of the data set is 39985 with 4 attributes. Collected all details are stored in Comma Separated value file format (.CSV). It is used to cluster and classify the data based on the sentiment and cluster using K-means clustering and Naïve Bayes classification techniques.



tweet_id	sentiment	author	content
1956967341	empty	xoshayzers	@tiffanylue i know i was listenin to bad habit earlier and i started freakin at his part =[
1956967666	sadness	wannamama	Layin n bed with a headache ughhhh,,,waitin on your call,,,
1956967696	sadness	coolfunky	Funeral ceremony,,,gloomy friday,,,
1956967789	enthusiasm	czareaquino	wants to hang out with friends SOON!
1956968416	neutral	xkilljoyx	@dannycastillo We want to trade with someone who has Houston tickets, but no one will.
1956968487	sadness	ShansBee	I should be sleep, but im not! thinking about an old friend who I want. but he's married now. damn, & he wants me 2! scandalous!
1956968636	worry	mcsleazy	Hmmm. http://www.djhero.com/ is down
1956969035	sadness	nic0lepaula	@charviray Charlene my love. I miss you
1956969456	neutral	feinyheiny	cant fall asleep
1956969531	worry	dudeitsmanda	Choked on her retainers
1956970047	sadness	Danied32	Ugh! I have to beat this stupid song to get to the next rude!
1956970424	sadness	Samm_xo	@BrodyJenner if u watch the hills in london u will realise what tourture it is because were weeks and weeks late i just watch itonlinelol
1956970860	surprise	okiepeanut93	Got the news
1956971077	sadness	Sim_34	The storm is here and the electricity is gone
1956971170	love	poppygallico	@annarosekerr agreed
1956971473	worry	LCJ82	@PerezHilton lady gaga tweeted about not being impressed by her video leaking just so you know
1956971586	sadness	cleepow	How are YOU convinced that I have always wanted you? What signals did I give off...damn I think I just lost another friend
1956971981	worry	andreagauster	oh too bad! I hope it gets better, I've been having sleep issues lately too
1956972097	fun	schiz0phren1c	Wondering why Im awake at 7am,writing a new song,plotting my evil secret plots muahahaha...oh damn it,not secret anymore
1956972116	neutral	jansc	No Topic Maps talks at the Balisage Markup Conference 2009 Program online at http://tr.im/mL6Z (via @bobdc) #topicmaps
1956972270	worry	sweet8181	I ate Something I don't know what it is, Why do I keep Telling things about food
1956972359	sadness	xamountoftruth	so tired and i think Im definitely going to get an ear infection, going to bed "early" for once,
1956972444	worry	jomama6881	On my way home n having 2 deal w underage girls drinking gin on da bus while talking bout keggers.....damn i feel old
1956972557	sadness	LilithGaea	IsaacMascote im sorry people are so rude to you, isaac, they should get some manners and know better than to be so lewd!

Fig.3.2. Collected Sentiment Analyse Data set

3.1.2. PREPROCESSING AND DATA SELECTION

3.1.2.1. PREPROCESSING

Tweet ID, Author, Content and Sentiment are the attributes of the data set with 39985 instances. In the initial phase of the pre-processing, single quote, double quote and ASCII characters and symbols are removed from the content attribute data. After removing the unwanted characters, data set is passed to the simple K-Means algorithm to cluster the data.

K-Means clustering process performed in the tweet emotion data set using Euclidean distance as distance function and K-means++ as initialised method. The total 39985 instanced are grouped under 13 clusters.

K-Means clustering is added by using the Add Cluster from the attribute in unsupervised category in Weka.

K-means Clustering Method:

In this twitter_emotion dataset, K-means algorithm can be executed in the following steps:

- Partition of objects into 13 non-empty subsets based on the selection of objects using K-Means++ selection
- Identifying the cluster centroids (mean point) of the current 13 non-empty subsets.
- Assigning each point in these subject to a specific cluster
- Compute the distances using Euclidian distance from each point in the subset and allot points to the cluster one among 13 clusters based on nearby points for the centroid.
- After re-allotting the points from the subsets, find the centroid of the new 13 clusters are formed.

3.1.2.2. DATA SELECTION

After completion of clustering process in the pre-processing phase, the clustered data set is selected for the classification and knowledge discovery process. The pre-processed data set consists of 13 clusters in the tweet emotion data set.

K-Means clustering is one of the simplest and accurate unsupervised machine learning technique. Similar data objects are grouped together and dissimilar data objects fall under another group in the k-means clustering.

3.1.3. NAÏVE BAYES CLASSIFICATION

Naïve Bayes classifier estimates the class conditional probability by assuming that the attributes are conditionally independent, given the class label y.

Naïve Bayes Classification Method:

- Scan the clustered tweet emotion dataset
- Calculate the probability of each attribute value based on cluster / sentiment. [n, n_c, m, p]
- Apply the formulae P(attribute value(ai)/subject value vj)=(n_c + mp)/(n+m)
- Multiply the probabilities by p
- Compare the values and classify the attribute values to one of the predefined set of class such as Sentiment or cluster. Both are used as the nominal class variable to classify the tweet_emotion data set.

The clustered Tweeter_emotion data set is classified using Naïve Bayes classifier with 10 cross fold validation based on the sentiment and also cluster as nominal class variables.

Naive Bayes in each cluster

- Calculates the prior probability for the target attribute in the tweet_emotion data set
- Calculate the conditional probability for the remaining attributes attribute in the tweet_emotion data set

CROSS FOLD VALIDATION

In this methodology, 10 Cross fold validation is used to measure the stability of the performance of the Naive Bayes with K-Means model in the tweet emotion data set. The sensitivity, specificity and accuracy are calculated in the tweet emotion data based on classification. These are calculated by the values of Positive, Negative, True Positive and True Negative values.

Sensitivity = True Positive / Positive, Specificity =True Negative / negative

Accuracy =(True Positive + True Negative) / (Positive + Negative)

IV. RESULTS AND DISCUSSION

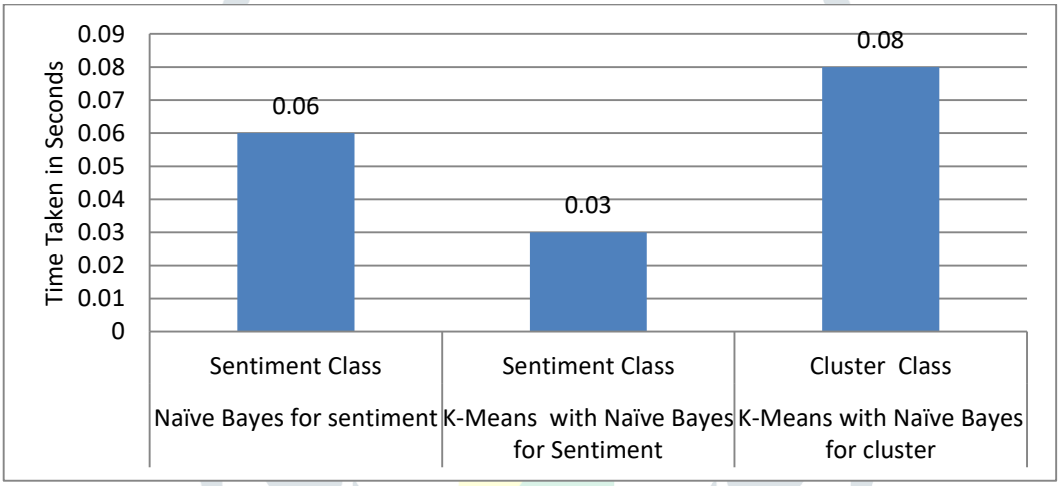
4.1 COMPARISION OF TIME TAKEN TO BUILD THE MODEL

Table 4.1. Time taken to build the model in tweet_emotion data set

Method	Nominal Variable	Time Taken
Naïve Bayes	Sentiment	0.06 Seconds
K-Means Clustering with Naïve Bayes	Sentiment	0.03 Seconds
K-Means Clustering with Naïve Bayes	Cluster	0.08 Seconds

The above table reveals that the Naïve Bayes classifier in clustered data takes least time to build the model with 0.03 seconds for Sentiment as nominal variable and its takes highest time to build the model with 0.08 seconds for cluster as nominal variable.

Chart 4.1. Time Taken to Build Model in tweet emotion data



4.2. COMPARISON OF NAÏVE BAYES CLASSIFIER BASED ON CORRECTLY CLASSIFIED INSTANCES

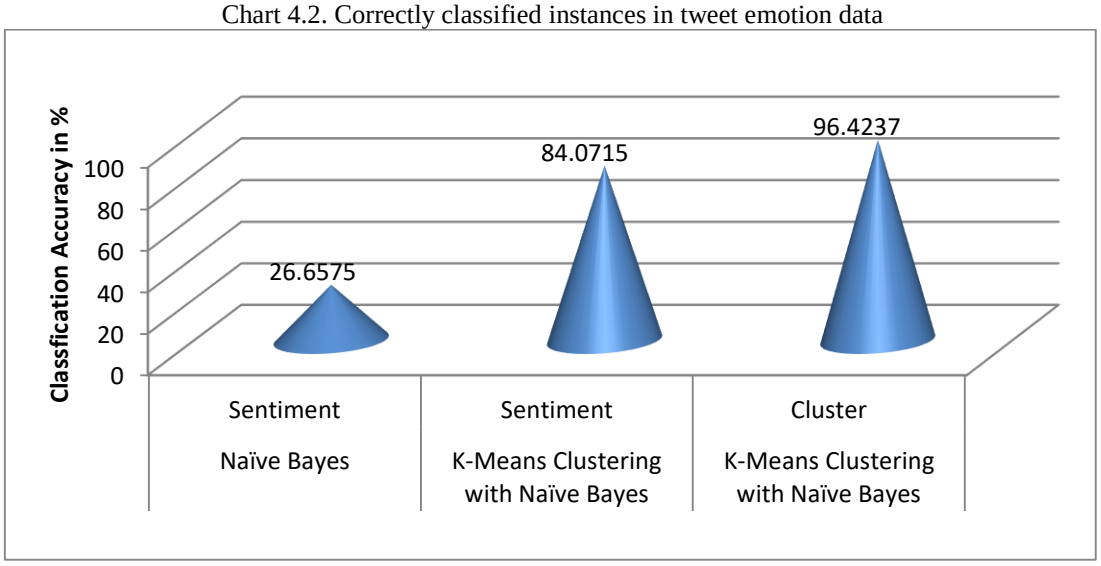
Table 4.2. Correctly classified instances in tweet_emotion data

Method	Nominal Variable	Correctly classified Instances	Percentage
Naïve Bayes	Sentiment	10659	26.6575
K-Means Clustering with Naïve Bayes	Sentiment	33616	84.0715
K-Means Clustering with Naïve Bayes	Cluster	38555	96.4237

The above table reveals that the out of 39985 instances, 10659 instances are correctly classified by the Naïve Bayes on tweet_emotion data, 33616 instances are correctly classified by the Naïve Bayes on clustered tweet_emtion data for sentiment as nominal variable and 38555 instances are correctly classified for cluster as nominal variable.

Naïve Bayes in clustered tweet-emotion data using cluster as nominal class variable produced highest accuracy (96.42%) and produce 84.97 % accuracy for sentiment as class variable. Naïve Bayes Classifier produced 26.66% accuracy in the classification of tweet_emotion data set. It shows that the hybrid classification produces highest accuracy.

Chart 4.2. Correctly classified instances in tweet emotion data



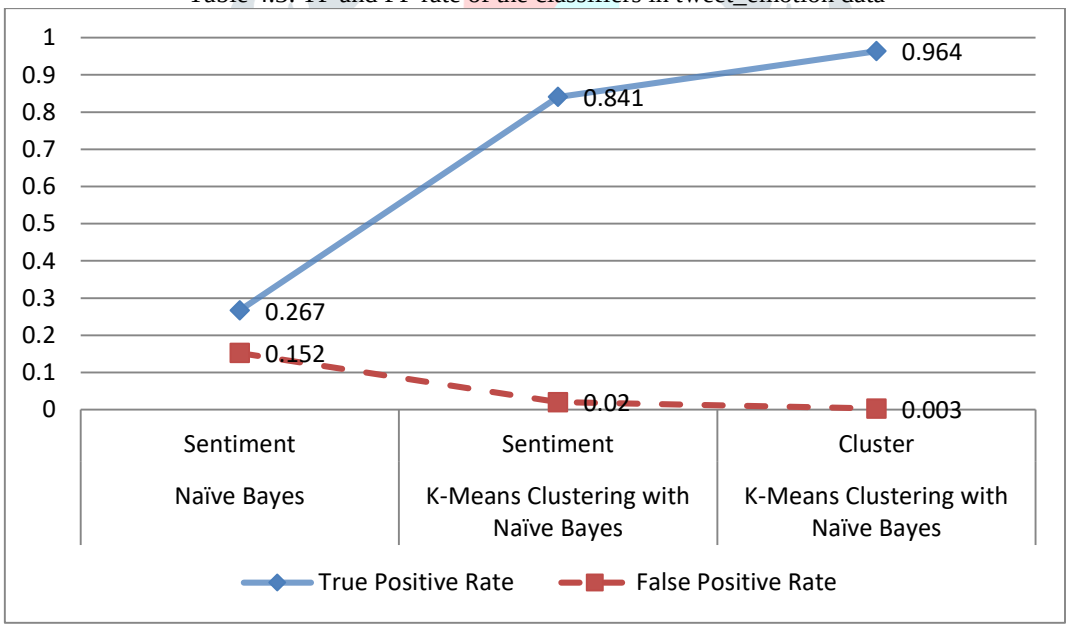
4.3. COMPARISION OF TP AND FP RATE

Table 4.3. TP and FP rate of the classifiers in tweet_emotion data

Method	Nominal Variable	True Positive Rate	False Positive Rate
Naïve Bayes	Sentiment	0.267	0.152
K-Means Clustering with Naïve Bayes	Sentiment	0.841	0.020
K-Means Clustering with Naïve Bayes	Cluster	0.964	0.003

The above table shows that the K-Means with Naïve Bayes classifier on cluster class in the tweet emotion data set have highest value (0.964) and it produces very low FP rate (0.003). and the same classifier on sentiment class produces 0.841 as TP Rate and 0.020 as FP Rate. The naïve bayes classifier in the tweet emotion data produces very low TP rate (0.267) and high FP rate(0.152).

Table 4.3. TP and FP rate of the classifiers in tweet_emotion data



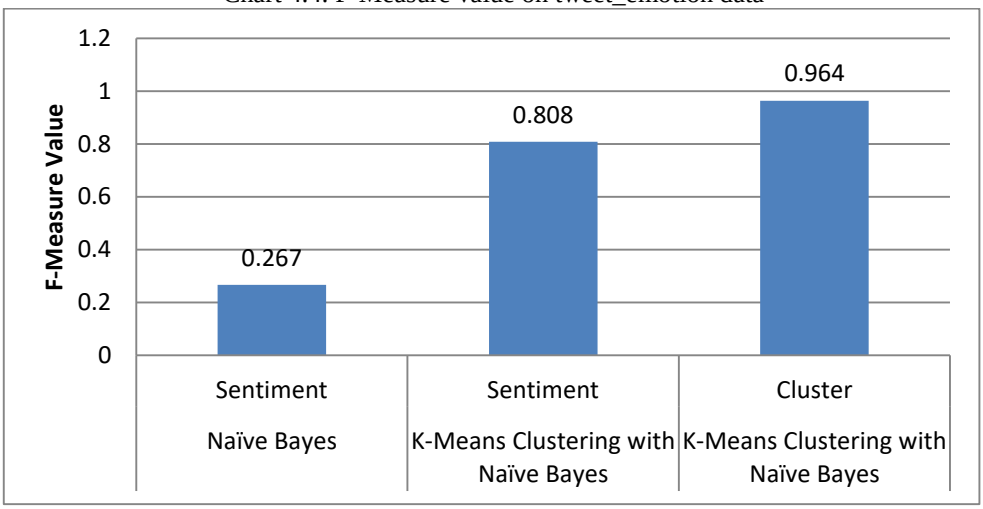
4.4. COMPARISION OF F-MEASURE VALUES

Table 4.4. F-Measure value on tweet_emotion data

Method	Nominal Variable	F-Measure
Naïve Bayes	Sentiment	0.267
K-Means Clustering with Naïve Bayes	Sentiment	0.808
K-Means Clustering with Naïve Bayes	Cluster	0.964

The above table reveals that the class variable cluster based classification using Naïve Bayes with K-Means clustering produces highest f-measure value as 0.964 and followed by sentiment class variable as 0.808. The naïve bayes classifier produces very least f-measure value (0.267).

Chart 4.4. F-Measure value on tweet_emotion data



V. CONCLUSION AND FUTURE ENHANCEMENTS

In the text mining process, classification for sentiment analysis plays important role. Many researchers have been applying different data mining classification and clustering techniques in the tweets to analyse opinion, emotion, popularity and trend analysis.

Naive Bayes is one of the successful data mining classification technique used in the social media data analysis to classify the emotions in the tweet contents, posted messages in the Facebook wall, etc.,

This research work investigates integrating K-means clustering with naive bayes in the tweet emotion data classification. This work investigates the performance of the naive bayes classification technique with and also without k-means clustering in the tweet emotion data set.

The result shows that the Naive Bayes classification in the pre-processed using K-Means tweet emotion data set produces higher accuracy than the classification of tweet emotion data set without clustering.

The best accuracy achieved is Naive Bayes classify the clustered tweet emotion data set using cluster as class variable than sentiment as class variable. This work is to be extended with classification of other social media contents and opinion mining for product reviews.

REFERENCES

1. M. Vedanayaki, “A Study of Data Mining and Social Network Analysis”, Indian Journal of Science and Technology, Vol 7(S7), 185–187, November 2014.
2. Adedoyin-Olowe, M., Gaber, M. M., & Stahl, F. (2013). A survey of data mining techniques for social media analysis. arXiv preprint arXiv:1312.4617.
3. Batista, G., & Monard, M.C., (2003), An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, vol. 17, pp.519-533.
4. Mariam Adedoyin-Olowe, Mohamed Medhat Gaber and Frederic Stahl, “A Survey of Data Mining Techniques for Social Network Analysis”, Journal of Data Mining & Digital Humanities, 2014.
5. Bogdan Batrinca and Philip C Treleaven, "Social media analytics: a survey of techniques, tools and platforms," AI \& SOCIETY, vol. 30, no. 1, pp. 89-116, 2015.
6. Karthikeyan, M., & Vyas, R. (2014).Cloud Computing Infrastructure Development for Chemoinformatics.In Practical Chemoinformatics (pp. 501-528).Springer India.
7. J. Bonneau, J. Anderson, and G. Danezis. Prying data out of a social network. pages 249 –254, july 2009.
8. D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. V. Alstyne. Computational social science. Science, 323:721–723, 2009.
9. Ritu,Ajit Singh,Akash Srivastava, “Opinion Mining Techniques on Social Media Data”, International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 6, May 2015.
10. Al-Daihani, S. M., & Abrahams, A. (2016). A Text Mining Analysis of Academic Libraries' Tweets. The Journal of Academic Librarianship, 42(2), 135-143.