Mining Tweets for Public Opinion

Dr.P.Perumal[1], R.Roopshree[2], R.D.Tarikasri[3], V.Vaishnavi[4]
[1, 2, 3, 4] Department of Computer Science and Engineering
[1]Professor, Sri Ramakrishna Engineering College, Coimbatore
[2, 3, 4] Students, Sri Ramakrishna Engineering College, Coimbatore

Abstract- *Social networks have revolutionized the manner within which folks communicate. Data Available from social networks is useful for analysis of user opinion, for instance measure the feedback on a recently free separation through this information is tedious and probably high-priced. Sentiment analysis could be a comparatively new space that deals with extracting user opinion mechanically. Associate in nursing example of a positive sentiment is, "natural language process is fun" instead, a negative sentiment is "it's an ugly day, I'm not going outside". Objective texts square measure deemed to not be expressing any sentiment, like news headlines, for instance "company shelves wind sector plans". There square measure some ways within which social network information are often leveraged to provide a far better understanding of user opinion such issues square measure at the guts of natural Language process (NLP) and data processing analysis. During this paper we have a tendency to gift a tool for sentiment analysis that is in a position to analysis Twitter information. We have a tendency to show a way to mechanically collect a corpus for sentiment analysis and opinion mining functions. Victimization the corpus we have a tendency to build a sentiment classifier that's able to verify positive, negative and objective sentiments for a document.*

Keywords- Item reputation, Reviews, Rating prediction, Recommender system, User sentiment

I. INTRODUCTION

Opinion mining is a type of natural language processing for tracking the opinion of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and categorize opinions about a product. Recently, designed a comparatively straightforward code in R to investigate the content of Twitter posts by victimizing the classes known as positive, negative and neutral. The concept of process tweets is predicated on a presentation. The formula evaluates tweets supported the quantity of positive and negative words within the tweet. The words within the tweet correspond with the words in dictionaries that you simply will notice on the web, however you'll produce an inventory yourself. It's additionally attainable to edit this list or wordbook. Great work however discovered some issue. There square measure some limitations within the API of Twitter. It depends on the entire variety of tweets you access via API, however sometimes you'll get tweets for the last 7-8 days (not longer, and it is 1-2 days only). The seven to eight day deadline to access a tweet creates a limitation in understanding what activities or events influenced a tweets or analyzing historical trends. An additive file is been created to bypass this limit and accumulate historical knowledge. If you access tweets often, then you'll analyze the dynamics of the interactions via chart like this one. To do this, they collected five hundred million Tweets made by quite 2 million individuals. They found fascinating daily and weekly trends in attitudes. It's a good example of the kind of attention-grabbing things social scientists will do with on-line social network knowledge. additional usually, the expansion of what laptop scientists decision "big data" presents social scientists with distinctive opportunities for researching previous queries, at the side of empowering US to raise new queries. Whereas a number of this huge knowledge is just numbers, abundant of it additionally consists of text. Sociologists have long had tools to help US in writing and analyzing dozens or perhaps many text documents, however several of those tools square measure less helpful once the quantity of documents is within the tens of thousands or millions. Each social science prof, grad student and college boy within the U. S. operating along couldn't code even the half daily sample of tweets that twitter provides unengaged to researchers. Luckily, laptop scientists are operating for quiet and whereas on specifically this knowledge problem–how will we collect, reason and perceive large text databases.

II. LITERATURE SURVEY

H. Feng, and X. Qian [1] proposed that online social network information guarantees to extend recommendation accuracy on the far side the capabilities of strictly rating/feedback-driven recommender systems (RS). On higher serve users' activities across totally different domains, several on-line social networks currently support a replacement feature of "Friends Circles" that refines the domain-oblivious "Friends" thought. RS ought to additionally have the benefit of domain-specific "Trust Circles". Intuitively, a user might trust {different totally different|completely different} subsets of friends concerning different domains sadly, in most existing multi-category rating datasets, a user's social connections from all classes area unit mixed along. This paper presents an attempt to develop circle-based RS. We tend to concentrate on inferring category-specific social trust circles from accessible rating knowledge combined with social network knowledge. We tend to define many variants of coefficient friends among circles supported their inferred experience levels. Through experiments on publically accessible knowledge, we tend to demonstrate that the planned circle-based recommendation models will higher utilize user's social trust info, leading to raised recommendation accuracy.

W. Luo, F. Zhuang [2] suggested that the boom of social media, it's a really fashionable trend for individuals to share what they're doing with friends across numerous social networking platforms. Nowadays, we've got a colossal quantity of descriptions, comments, and ratings for native services. The data is effective for brand new users to guage whether or not the services meet their necessities before partaking. During this paper, we have a tendency to propose a user-service rating prediction approach by exploring social users' rating behaviors. So as to predict user-service ratings, we have a tendency to concentrate on users' rating behaviors. In our opinion, the rating behavior in recommender system may be embodied in these aspects: 1) once user rated the item, what the rating is, 2) what the item is, 3) what the user interest that we have a tendency to may dig from his/her rating records is, and 4) however the user's rating behavior diffuses among his/her social friends. Therefore, we have a tendency to propose an inspiration of the rating schedule to represent users' daily rating behaviors. Additionally, we have a tendency to propose the issue of social rating behavior diffusion to deep perceive users' rating behaviors. within the projected user-service rating prediction approach, we have a tendency to fuse four factors, user personal interest (related to user and therefore the item's topics), social interest similarity (related to user interest), social rating behavior similarity (related to users' rating behavior habits), and social rating behavior diffusion (related to users' behavior diffusions), into a unified matrix-factorized framework. We have a tendency to conduct a series of experiments in Yelp dataset and Doosan show dataset. Experimental results show the effectiveness of our approach.

L. Polanyi [3] implemented the explosion of net opinion knowledge has created essential the necessity for automatic tools to investigate and perceive people's sentiments toward totally different topics. However, it's acknowledged that there's no universally best sentiment lexicon since the polarity of words is sensitive to the subject domain. Even worse, within the same domain an equivalent word might indicate {different totally different completely different} polarities with regard to different aspects. For instance, in an exceedingly portable computer review, "large" is negative for the battery side whereas being positive for the screen side. During this paper, we tend to concentrate on the matter of learning a sentiment lexicon that's not solely domain specific however additionally keen about the side in context given an unlabelled self-opinionated text assortment. We tend to propose a unique optimization framework that gives unified and scrupulous thanks to mix totally different sources of data for learning such a context-dependent sentiment lexicon. Experiments on 2 knowledge sets (hotel reviews and client feedback surveys on printers) show that our approach can't solely establish new sentiment words specific to the given domain however additionally confirm the various polarities of a word looking on the side in context. In any quantitative analysis, our methodology is well-tried to be effective in constructing a top quality lexicon by scrutiny with somebody's annotated gold customary. Additionally, victimization the learned context-dependent sentiment lexicon improved the accuracy in an aspect-level sentiment classification task.

B. Pang [4] experimented the arrival and recognition of social network, a lot of and a lot of users wish to share their experiences, like ratings, reviews, and blogs. The new factors of social network like social influence and interest supported circles of friends bring opportunities and challenges for recommender system (RS) to resolve the cold begin and scantiness downside of datasets. A number of the social factors are utilized in RS, however haven't been totally thought of. During this paper, 3 social factors, personal interest, social interest similarity, and social influence, fuse into a unified customized recommendation model supported probabilistic matrix factorization. The issue of private interest will create the RS suggest things to satisfy users' individualities, particularly for knowledgeable users. Moreover, for cold begin users; the social interest similarity and social influence will enhance the intrinsic link among options within the latent house. We tend to conduct a series of experiments on 3 rating datasets: Yelp, Movie Lens, and Doosan pic. Experimental results show the planned approach outperforms the prevailing RS approaches.

III. PROPOSED METHODOLOGY

To identify the features of an object in a review either noun based approach or verb based approach, verb based approach are quite difficult to identify. Different words or phrases can be used to refer to the same feature of an object and it is difficult task to identify these words. To determine the strength of an opinion is another challenge faced in opinion mining.

There is a need for algorithms to be used in predicting the sentiments [6] that paves the way for important decision making for the users and the manufacturers in the feedback of their product. Naive Bayes works on Bayes theorem of probability to predict the class of unknown dataset. R can access millions and millions of near real time data [5] and they have higher data processing capability. Every day we are met with complex analytic function and R language is used to sort it out using CRAN (The Comprehensive R Archive Network) library file. Text mining or text analysis on real time data from twitter to produce an interactive dashboards using tableau could be performed.

It is a classification technique supported Bayes' Theorem with an assumption of independence among predictors. In easy terms, a Naive Bayes categorizer assumes that the presence of a specific feature in an exceedingly class is unrelated to the presence of the other feature. For instance, a fruit is also thought of to be an apple if it's red, round, and concerning three inches in diameter. Notwithstanding these options depend upon one another or upon the existence of the opposite options, all of those properties severally contribute to the likelihood that this fruit is an apple which is why it's referred to as 'Naive'.

Naive Bayes model is straightforward to create and notably helpful for terribly giant knowledge sets. Beside simplicity, Naive Bayes is thought to shell even extremely refined classification ways.

Bayes theorem provides how of shrewd posterior likelihood P (c | x) from P(c), P(x) and P (x | c). Verify the equation below:

$$P(c|\,x) \quad P(x|\,c)*P(c)/P(x) \qquad (1)$$

$P(c \mid x) = P(x_1 \mid c) * P(x_2 \mid c) * \ldots * P(x_n \mid c) * P(c)$

Where,

P(c|x) is the posterior probability of given predictor (x, attributes) class (c, target).

P(c) is the prior probability of given class.

P(x | c) is the probability of predictor given class.

P(x) is the prior probability of given predictor.

System Architecture

The following Fig 1 shows the architecture of opinion mining

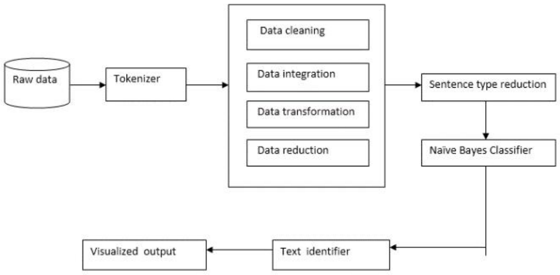


Fig 1 System Architecture

The system architecture shows how the raw data is being cleansed through tokenizer. The cleansed data is being reducted according to the sentence type using Naïve Bayes classifier.

Overall Architecture of the Proposed System

The following Fig 2 depicts the overall architecture of proposed system.

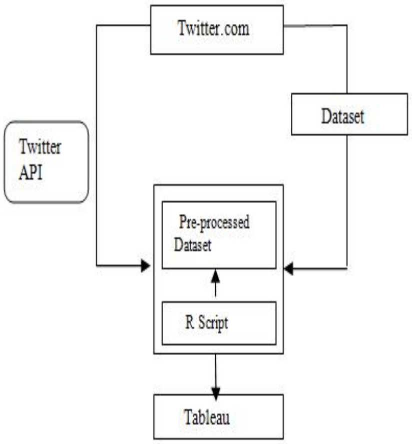


Fig 2 Architecture for Proposed System

The dataset retrieved from twitter through twitter API is being pre-processed using Naïve Bayes algorithm and is visualized through Tableau.

IV. IMPLEMENTATION

In this, the Sentence level Categorizer is employed for grouping the datasets from Twitter [7]. The datasets square measure then tokenized by tokenizer. The tokens square measure then processed by information pre-processing i.e. information cleansing, information Integrity, information Transformation Associate in Nursing Reduction is dole out to a visible format this is often then understood to an symbol for distinctive whether or not the given information square measure positive, negative, neutral, or negation. Naïve Bayes' Classifier is to classify the datasets since it's the most effective classifier for sentence level categorization. Tweets and texts square measure short (a sentence or a headline instead of a document). The language used is incredibly informal, with artistic orthography and punctuation, misspellings, slang, new words, URLs, and genre-specific word and abbreviations, such as, RT for "re-tweet" and # hash tags, that square measure a kind of tagging for Twitter messages [10].

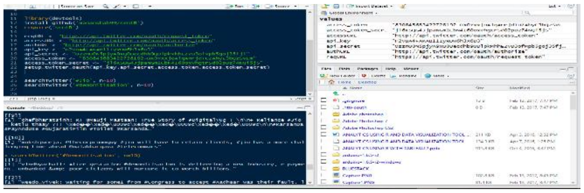
Another facet of social media information like Twitter messages is that it includes wealthy structured info regarding the people concerned within the communication. For instance, Twitter maintains info of who follows who mind re-tweets and tags within tweets give discourse info. By the Naïve mathematician Classifier algorithmic rule the classified data square measure pictured by R-platform. Tableau Public is a free service that lets anyone publish interactive data to the web. Tableau promises numerous new features for data preparation, query performance and more. Tableau products include the ability to connect to a wide variety of data sources.

V. RESULTS AND DISCUSSION

A. Data Acquisition

In data acquisition, data is been collected from twitter through twitter API's. The data with different polarity is been analyzed. The dataset is read in R or imported in R. The three dataset is appended so that it can be used for further function.

The following fig 3 shows how the data is being retrieved using four API keys.



```
library(devtools)
setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)

searchTwitter('#jio', n=10)
searchTwitter('#demonetization', n=10)
```

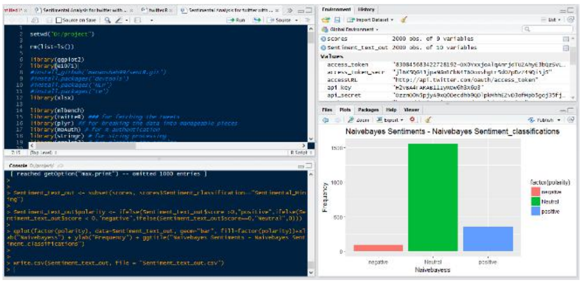
```
accessURL   "https://api.twitter.com/oauth/access_token"
authURL     "https://api.twitter.com/oauth/authorize"
reqURL      "https://api.twitter.com/oauth/request_token"
```

Fig 3 Retrieving Data

B. Analyzing Data Using R

Analyzing each data is a complex one. Data's must be pre-processed in order using Naïve Bayes Algorithm. The dataset is read in R or imported in R. The dataset is appended. Some functions is been allotted to perform its functionality. To compute these functions R is coded to perform. Every function is coded to perform. Every function is coded in single specific script and each function is stored in a specific file and in specific destination. Its output is given as an input to Tableau.

The following fig 4 shows how the opinion of the object is being mined



```
setwd("D:/project")

rm(list=ls())

library(ggplot2)
library(e1071)
#install.packages('devtools')
#install.packages('NLP')
#install.packages('tm')
library(xlsx)

library(mlbench)
library(twitteR) ### for fetching the tweets
library(plyr) ## for breaking the data into manageable pieces
library(ROAuth) # for R authentication
library(stringr) # for string processing
```

```
[ reached getOption("max.print") -- omitted 1000 entries ]
>
>
> Sentiment_text_out <- subset(scores, scores$Sentiment_classification=="Sentimental_Mining")
>
> Sentiment_text_out$polarity <- ifelse(Sentiment_text_out$score >0,"positive",ifelse(Sentiment_text_out$score < 0,"negative",ifelse(Sentiment_text_out$score==0,"Neutral",0)))
>
> qplot(factor(polarity), data=Sentiment_text_out, geom="bar", fill=factor(polarity))+xlab("Naivebayess") + ylab("Frequency") + ggtitle("Naivebayes Sentiments - Naivebayes Sentiment_classifications")
>
>
> write.csv(Sentiment_text_out, file = "Sentiment_text_out.csv")
>
```

```
scores               2000 obs. of 9 variables
Sentiment_text_out   2000 obs. of 10 variables
Values
accessURL   "http://api.twitter.com/oauth/access_token"
```

Fig 4 Naïve Bayes Representation

C. Visualization with Tableau

Tableau is the data visualization software and interactive tool which is one of the top most on demand visualization software in the world [9]. Tableau takes the input of processed files and used to create graphs in more interactive way. Tableau is used to create dashboards for end users.

The following fig 5 depicts how it is visually represented using Tableau

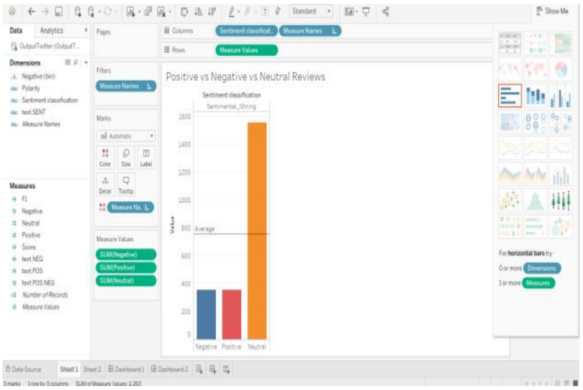


Fig 5 Visualized graph

D. Performance Evaluation

The following fig 6 shows the percentage of detection rate between Naïve Bayes and K-means algorithm.

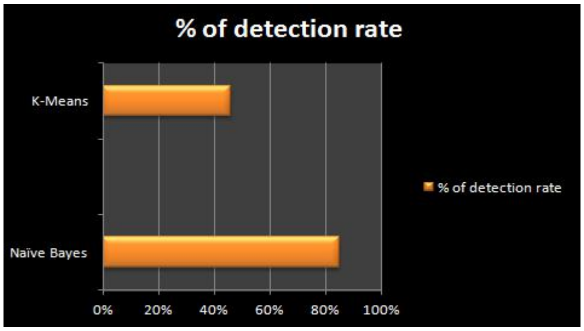


Fig 6 Detection rate

Out of 500 reviews extracted from twitter, Naïve Bayes retrieved almost 85% total no of reviews and K-means retrieved 46%.

The following fig 7 depicts the percentage of accuracy rate between Naïve Bayes and K-means algorithm.

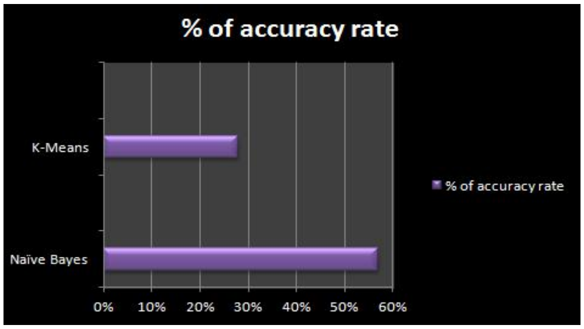


Fig 7 Accuracy rate

Out of 500 reviews extracted from twitter, Naive Bayes has acquired 57% of positive accuracy and K-means has acquired 28%

VI. FUTURE ENHANCEMENT

There remains a lot of research work in this field and there are some works which is not yet solved to give detail marks about the items and use of multiple languages at a time. In future work, topic modeling could also be explored.

VII. CONCLUSION

Twitter offers an opportunity to create and implement theories & technologies that is used to search sentiments. In this paper, it discusses the approach for mining the tweets for public opinion. For mining the sentiment, the opinion words are being extracted (a combination of the adjectives along with the verbs and adverbs) in the tweets. The corpus-based method and dictionary based method is used for orientation of the acquired sentiment of verbs and adjectives. Finally, Tableau helps in visualizing the results in statistical manner.

REFERENCES

[1] X. Yang, H. Stack, and Y. Liu, “Circle- Based Recommendation in online social networks” in Proc. 18th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, Aug, 2012, pp. 1267-1275.

[2] K. Zhang, Y. Cheng, W. Liao, A. Choudhary, “Mining millions of reviews, a technique to rank Products based on importance of reviews, ” in Proceedings of the 13th International Conference on Electronic Commerce, Aug. 2011, pp. 1-8.

[3] M. Jamal and M. Ester. “A matrix factorization technique with trust propagation for recommendation in social networks,” in Proc. ACM conf. RecSys, Barcelona, Spain, 2010, pp. 135-142.

[4] Z. Fu, X. Sun, Q. Liu, et al, “Achieving Efficient Cloud Search Services: Multi-Keyword Ranked Search over Encrypted Cloud Data Supporting Parallel Computing,” IEICE Transactions on Communications, 2015, 98(1):190-200.

[5] G. Ganu, N. Elhadad, A Marian, “Beyond the stars: Improving rating predictions using Review text content,” in 12th International workshop on the Web and Databases (WebDB 2009).pp 1-6.

[6] J. Xu, X. Zheng, W. Ding, “Personalized Recommendation based on reviews and rating alleviating the sparsity problem of collaborative filtering,” IEEE International Conference on e-business Engineering.2012, pp. 9-16.

[7] X. Qian, H. Feng, G. Zhao, and T. Mei, “Personalized recommendation combining user interest and social circle,” IEEE Trans Knowledge and data engineering. 2014, pp. 1763-1777.

[8] D.M. Biel, A.Y. Ng, and M. L Jordan, “Latent Dirichlet Allocation,” Journal of machine learning research 3. 2003, pp. 993 – 1022.

[9] W. Zang, G. Ding, L. Chen, C. Li and C. Zhang ,”Generating virtual ratings from reviews to augment online recommendations,”ACM TIST, vol. 4, no.1. 2013, pp. 1-17.

[10] Y. Ren, J. Wang, J. Han and S. Lee, Mutual Verifiable Provable Data Auditing in Cloud Storage, “Journal of Internet Technology, vol. 16, no. 2, 2015, pp. 317-323.