Journal of Engineering Sciences

Vol 9, Issue 2, 2018
ISSN NO:0377-9254

TWITTER SENTIMENT ANALYSIS USING MACHINE LEARNING

Rohan Raut[1], Sakil Ansari[2]
Associate Software Engineer[1] (rahack99@gmail.com[1]), Associate Software Engineer[2] (sakilansari4@gmail.com[2])
Accenture Solutions Pvt.Ltd[1], Larsen and Toubro Technology Services[2]

Abstract- Online Microblogging on social networks have been used for indicating opinions about certain entity in very short messages. Existing some popular microblogs like Twitter, Facebook etc, in which twitter attains maximum amount of attention in the field of research areas related to product, movie reviews, stock exchange etc. The research on sentiment analysis has been going for a long time. Sentiment analysis in present days becomes the major issue in field of research and technology. Due to day by day increase in the number of users on the social networking websites, huge amount of data produces in the form of text, audio, video and images. There is need to do sentiment analysis as texts in form of messages or posts to find whether the sentiment is negative, positive or neutral. We had extracted data from twitter i.e. movie reviews for sentiment prediction using machine-learning algorithms. We applied supervised machine-learning algorithms like support vector machines (SVM), maximum entropy and Naïve Bayes to classify data using unigram, bigram and hybrid i.e. unigram + bigram features. Result shows that SVM surpassed other classifiers with remarkable accuracy of 84% for movie reviews.

I. INTRODUCTION

Social networking is the grouping of individuals into specific groups. It could be an apolitical or religious group or a group of college students, teenagers, all together sharing information of their interests, mostly online. Twitter, MySpace or Facebook are some of the social networking sites that are free of charge and easy to access. This interaction is likely to include friendship, families, group relation and romantic ones. Social networking helps people to make new friends and develop some personal relationships and stay in touch with family very easily. Due to vast number of people connects to networking sites, number of relationships gradually increases. Social networking features combined in one website are: user groups, the latest info about music groups, places for videos and photos, blogs, personal profile, and much more. Social networking sites also helps people for maintaining and developing business contacts contact with them. LinkedIn is the best example for this, as it can be suitable place to talk about business and meet with professionals. It is easier and faster to be involving with new business clients. Internet is foremost and first communication technology with the capability to change social interaction of the people. Since early 1990s, there has been a proliferation of internet. For example,by 2003 63% of American had used the internet. In 1990s, Information technology experts expected the internet to be consigned to the trash heap of history. Internet has become an essential part of our lives; many websites have facility ways for people to keep in touch in the form of social networking. Social networking sites are the way for interact with new people and to make connections as well as share photos, videos, and activities with each other.

According to Amanda Lenhart and Mary Madden, 55% of online teenagers have created a personal profile online, and 55% have used social networking sites like MySpace and Facebook. A social networking site includes both the exchange of information among individuals and groups online. Expression also represents a view perspective, reflection, or quality of the individual or groups.

Figure 1: Abstract view of social networking

In 1997, first social site was launched named sixdegree.com. The intention of this site was to make online dating smoother, and for the first time users were able to create their personal profile and then post it online and even surf the network. After sixdegree.com, other social sites were launched, and served for a while, and failed to become a sustainable business entity. Since then Ryze.com (2002), match.com, were launched and used by many users. Uncontrolled use of microblogging separates the users from the real world life and creates shortage of attention. Use of social networking site has disadvantages and advantages. It is up to the interest and knowledge of the user to know how to use them, when and which site to use.

Social networking has some of the advantages like the meeting places but virtually where people can share thoughts with whoever they want and meet them in the first place. Some people used social networking sites to make new friends, establish relationships and even marry. Some of them used this websites, in order to find their lost friends in their life and meet them. Use of social networking makes life faster and easier to get the latest news across the world at any moment and being updated. Nowadays, colleges and universities are getting fond of social networking, which makes it easier for faculty and students to find information freely and easily. Corporate and technical sectors also started recruiting employees with the help of social networking by going through their profile and background.

Since social networking sites are operating worldwide; it breaks some of the cultural

Different people from different parts of the world can be able to connect with their loved one and families barriers around the different parts of the world.

Different people from different parts of the world can be able to connect with their loved one and families easily and without any cost. Social networking brings the world together and modifies communication. Everything in the world has both advantages and disadvantages, even for social networking, the disadvantages are as follows; personal identity theft is the most popular one.

Social networking requires user's personal details in order to gain full access to the site as sign up. Recent information and news disclosed that some of the social networking websites misuses the personal information of users. Advertisers evade users' privacy. Sex offenders and criminals often visit the sites to find new victims. Some people mostly young ones for the sake of revenge and hate post embarrassing information or photos. These types of crimes can be categorized as cyber bullying in social networking makes this much faster and easier, unfortunately sometimes even led to death of teens. The developers made the social networking sites for better communication but people are rather addicted to those sites. This will hinder the ability of young people to develop real social life, face to face meeting of people, which is very important in developing conversation and speech. The traditional face to face socializing is becoming obsolete.

Online Microblogging

Online microblogging is broadcast medium that exists similar to blogging. Microblogging is different from blogging as its content normally smaller in both total and actual file size. Microblogs allow users to share small chunks of content such as video link, individual images or short messages, which may be the major reason for their popularity.

Advantages of Microblogging over traditional blogging:-

Why would anyone want to start posting on a microblogging site? If you've been hesitant to jump on on a site like Twitter or Tumblr, here are a few reasons to consider trying them

Developing content takes less time: The traditional blogs are quite lengthy so that it takes time to complete our intent. Microblogging gives you the benefit of posting the most recently happened incident to aware your loved ones in a short time and message.

- **Individual parts of the content consumed in less time :** Hence, microblogging is such a popular and interactive form of information consumption and social media on mobile devices, because as the gist of the content to the people increases, therefore it is best way where the news comes in short and precise way as compare to long ones that takes time.
- **Increases chances of frequent posts:** Microblogging involves the more frequent posts and shorter ones whereas traditional blogging involves exactly opposite less frequent post and longer. Since you're saving so much time by focusing on just posting short pieces, you `can afford to post more frequently.
- **Share time sensitive or urgent information in an easier way:** Huge number of microblogging platforms have been made to be fast and easy to use. With a Vine video, Tumblr post, Instagram photo or a simple tweet, you can easily share to everyone on what's happening in your life or any news at this very moment.
- **Communication with followers becomes easy and direct :** In addition to communicate easily with greater short and frequent posts, microblogging platforms can be used easily to encourage and facilitate better interaction through liking, reblogging , tweeting , commenting and more.
- **Convenient using with mobile and tabs:** Microblogging gains too much of attention in present days and the main cause behind this is increasing trends of mobile browsing. It is difficult to consume, interact and write long and lengthy blog post in a tab or smartphone that's why microblogging comes into play and provide small, easy and faster posts.

Twitter

Twitter is an online microblogging service that allows users to read and write short sentences of length 140 characters called tweets. Twitter Inc. is located at San Francisco. Users should register first to post any message, whereas unregistered users can only read them. Users can access Twitter with the website interface, mobile application or SMS. Twitter was created by Noah Glass, Biz Stone, Evan Williams and Jack Dorsey in March 2006 and launched in July 2006. Twitter has 310 Million monthly active users, 1 billion unique visits monthly to sites with embedded tweets, 83% of active users access through mobile application, consists of 3500 employees around the world, more than 35 offices across the world, 79% accounts are from outside the U.S. , supports more than 40 languages and 40% employees of Twitter are from technical background. All numbers approximate as of March 31, 2016.

The company experienced rapid initial growth. In year 2007 around 4,00,000 tweets were posted per quarter. In 2008 this extends to 10 million tweets a quarter. 50 millions tweets were posted per day in February 2010. 70000 applications were registered by company as March 2010.

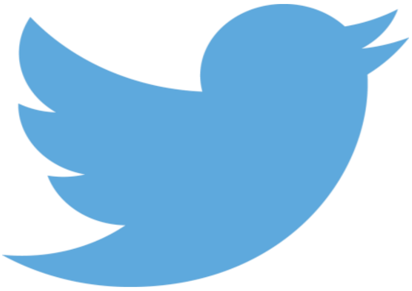


Figure 2: Twitter Logo

According to Twitter, 750 tweets are posted each second which equals to each day around 65 million tweets were posted as of June 2010. On daily basis around 140 tweets were posted by March 2011. In January 2009, since it gained a lot of popularity, Twitter becomes third-highest online microblogging site, given by Compete.com.

The reason we are using microblogging and Twitter data are the following:

- The scope of microblogging tends to grow bigger and bigger day by day. Easy to use and people can share and give opinions on a certain topic, thus it makes essential source.
- Twitter generates vast number of messages that is increasing exponentially. The extracted data can be enormously large.
- Twitter users varies from person to person as user can be politicians, film stars, celebrities, sportsmen and many leaders across the country. So, it contains all the messages of different caste, religion and sex.
- Twitter users are all over the world so it contains data for different language.

Sentiment Analysis

Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feels about it. Users share the things about their ongoing life, discuss current issues and variety of topics. Independent to write in any format without following rules that makes this more popular than older blogging sites. Movies and product reviews easily available now a days or thoughts on religious and political issues, so it becomes essential sources of user sentiment and opinion. Data that are we using in our experiment are from Twitter, it contains a vast number of messages by large number of users created by themselves. Messages can vary from public opinion to personal thought. These microblogging sites are huge source of information and it is quite easy to say that there is a need for automating the sentiment analysis process as there is too much work involved in processing this information manually. Various approaches are practiced for the automation of this process like machine learning and natural language processing. Users are increasing day by day as the population and trend of using microblogging sites are increasing, so the data can be used in research purpose of sentiment analysis and opinion mining.

In the time of election every news channel show the exit polls of every political party. So, every political party willing to know how many are in favor can do so with the help of microblogging sites where people will give their opinions about likes and dislikes of theirs party. These opinions will help parties to increase their voters.

The data we are using in this experiment are movie reviews. We have collected about 17000 movie reviews from Twitter. The movie reviews contains

categorized in three ways:

reviews of different movies. Reviews can be categorized in three ways:

1. Positive reviews: Messages in which people liked the movie.
2. Negative reviews: Messages in which people not liked the movie.
3. Neutral reviews: Messages in which people do not have emotions or based on mere fact.

We have extracted 5000 each positive, negative and neutral reviews for training set and 2000 reviews will be used in test set. We show comparison between the different machine learning classifiers and empirically conclude which will give the best result among these.

II. RESEARCH PROBLEM

Problem Statement

Microblogging is type of blogging which consists of a limited number of words. Limitation of words determined by respective microblogging sites. It gives right to share his/her thoughts, opinions and sentiments in less number of words. It is one of the revolutionary things happening in the world of technology. People these days depend upon microblogging sites such as Twitter, Facebook, Tumblr, etc. to communicate with both relatives and rest of world. Here sentiments come into the play which will be shared by anyone in the time they feel and wanted it to be shared. Sentiments are nothing but feelings with respect to event or situation. Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feels about it.

The research on sentiment analysis has been going for a long time. Sentiment analysis in present day is a prominent issue in the field of research and technology. Due to day by day increase in the number of users on social networking websites, a huge amount of data produced in the form of text, audio, video and images. There is need to do sentiment analysis as texts in form of messages or posts to find whether the sentiment is negative, positive or neutral.

Gap Analysis

A lot of research has been done in the area of sentiment analysis. Many researchers used Part-Of-Speech and polarity based feature using supervised learning techniques for classifying.

Many automatic classifiers are proposed for classifying the texts in the given expressions but with the restricted domains, but there will be new informal words that are added to the present world which means something in the common social network, so there is a need to include all the common referred terms that are used in the social networking world.

Objectives

The objectives of the paper has been discussed in the following points :-

1. To explore, analyze and study the existing sentiment analysis detection techniques in the online microblogging network.
2. To study how the tweets can be generated from the Twitter with the help of Java API.
3. To implement and analyze the results achieved after applying the supervised
4. Learning classifiers to the dataset.

III. RESEARCH METHODOLOGY

The main aim of this paper is to compare the results that are implemented with the help of supervised classifier.

The paper follows the following methodology :

1. We have collected a corpus of positive, negative and neutral tweets with the help of Twitter4j java API from Twitter. The size of our corpus is large.
2. We then remove the stop words from the collected corpus to make the content free from commas, full stops etc.
3. Lastly, we apply machine learning algorithms to our training set first and then test set and compare the results.

With the help of results we evaluate which machine learning algorithm is best for classification of data.

Preprocessing

Collection of data: We collected data from Twitter API named as Twitter4j using netbeans. given by using Hashtag(#) followed by the movie name like #FAN, #BajarangiBhaijaan, #TheJungleBook etc. Approx 17000 tweets have been collected from the various movie tweets.

Reviews can also be searched by #Hash tags followed by respective movie stars, directors, production's house and music record companies. In Twitter, hash tags become the necessary symbol to find about something and it gives user limit of 140 words to express their views and attitude.

Normalization

We have found that to get desired results from the classifier we have to make sure that the tweets can be processed properly. As tweets can be in user language, so we have to clean the data which is irrelevant to the classification. The things which can be irrelevant are as follows :-

- URL: URLs in the message will not make any sense as it simply distracts the result of classifier.
- Username: Removal of username can be necessary for cleaning purposes as it can effect falsely to our results.
- Repeated characters: If the character is repeated more than two times then it can comprise new word but the meaning is the same, so we have to eliminate that word and make the word genuine. For example 'good' can be written as 'gooooood'.

Repeated words If the message contains word which has appeared more than two times continuously then it has to be changed into two times. For example, 'great great great great movie' can be converted to 'great great movie'.

Removal of stop words

Stop words are words like "a", "is", "the", "etc" etc; These words have nothing to do with the emotion , so has to be discarded from the message. Now, the next step is to train the data using supervised classifier.

Machine Learning Techniques

We employed classification methods which is polarity based using set of positive, negative and neutral tweets provided by Twitter4j API. Polarity is given by ratio of probability of a word appeared in set of positive or negative statement which makes the word positive or negative. The classifiers we are using are based on the concept of polarity.

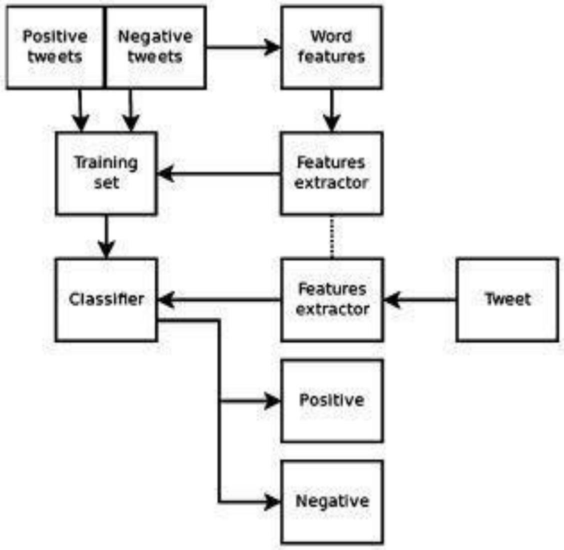
If the feature is independent and based only on Standard English Dictionary then only this technique works. This method fails when we tried to record the sentiment shown with respect to comparison. Further, the polarity based technique also fails to record query related sentiment. In order to fulfill the requirement of classification we involved machine learning techniques.

The machine techniques comprised of following supervised classifier that are given below:-

- Naïve Bayes
- Support Vector Machines (SVM)
- Maximum Entropy (MaxEnt)

Supervised Classifiers

Naïve Bayes*:* The Naïve Bayes classifier is one of the simplest probabilistic model which works positively on text categorization and employed on Bayes rule with self-supporting feature collection



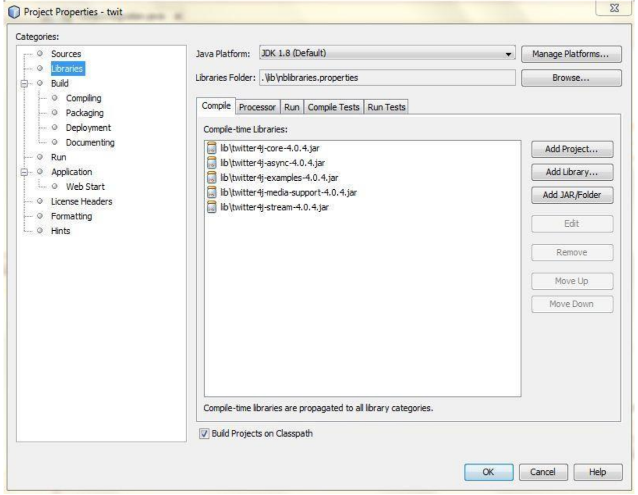
[3].

Figure 3: Flow Diagram of Supervised Classifiers

IV. IMPLEMENTATION

Implementation

Data Extracting: We are extracting tweets from the Twitter with the help of the Java API called Twitter4j. It consists various number of libraries that are used in the extraction. At first we have added this library into our java project. Then with the help of twitter app we have obtained Consumer Token Key and Access Token Key. Further, extraction of tweets will be start only after when we generate Access Key. Generation of Access Key is needed every time for the extraction of the tweets.



Java Platform: JDK 1.8 (Default)

Libraries Folder: .\lib\nblibraries.properties

Compile-time Libraries:

```
lib\twitter4j-core-4.0.4.jar
lib\twitter4j-async-4.0.4.jar
lib\twitter4j-examples-4.0.4.jar
lib\twitter4j-media-support-4.0.4.jar
lib\twitter4j-stream-4.0.4.jar
```

Compile-time libraries are propagated to all library categories.

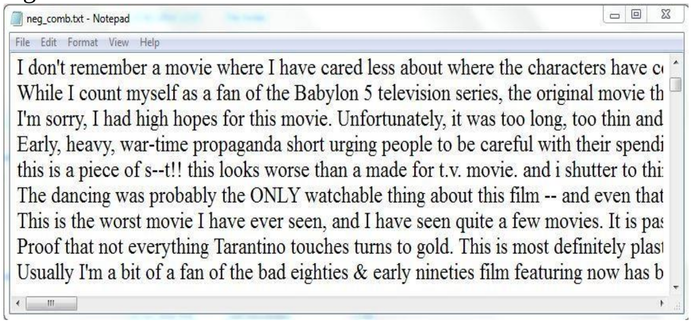
☑ Build Projects on Classpath

Figure 4: Twitter4j libraries

Consumer Token Key will be provided by the Twitter app. There is a unique key for every app and that key is known as Consumer Token Key. In order to obtain tweets we have to apply consumer token key and access token key into the java code.

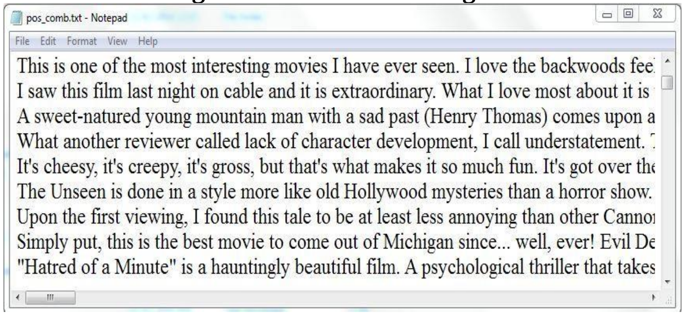
Preprocessing using R

In this step collected data is pre-processed. We have used R language for the pre -processing. Stop words, user references, urls etc are removed from the data. Regular expressions are used to remove url. Collected tweets are then manually labeled and stored in files as test dataset. We have two data sets: positive and negative.



I don't remember a movie where I have cared less about where the characters have c
While I count myself as a fan of the Babylon 5 television series, the original movie th
I'm sorry, I had high hopes for this movie. Unfortunately, it was too long, too thin and
Early, heavy, war-time propaganda short urging people to be careful with their spendi
this is a piece of s--t!! this looks worse than a made for t.v. movie. and i shutter to thi
The dancing was probably the ONLY watchable thing about this film -- and even that
This is the worst movie I have ever seen, and I have seen quite a few movies. It is pa
Proof that not everything Tarantino touches turns to gold. This is most definitely plast
Usually I'm a bit of a fan of the bad eighties & early nineties film featuring now has b

Figure 5: Positive Training Set



This is one of the most interesting movies I have ever seen. I love the backwoods fee
I saw this film last night on cable and it is extraordinary. What I love most about it is
A sweet-natured young mountain man with a sad past (Henry Thomas) comes upon a
What another reviewer called lack of character development, I call understatement.
It's cheesy, it's creepy, it's gross, but that's what makes it so much fun. It's got over the
The Unseen is done in a style more like old Hollywood mysteries than a horror show.
Upon the first viewing, I found this tale to be at least less annoying than other Canno
Simply put, this is the best movie to come out of Michigan since... well, ever! Evil De
"Hatred of a Minute" is a hauntingly beautiful film. A psychological thriller that takes

Figure 6: Negative Training Set

V. RESULTS

We are using R language for implementation. R language offers maximum support when it comes machine learning techniques. Machine learning techniques can be easily implemented in R language. Packages that we are using are “RTextTool”,”Rweka”and “e1071”. RTextTools have most of the machine learning algorithms but not have Naïve Bayes, which is included in e1071 package and Rweka package is used for n-gram feature.

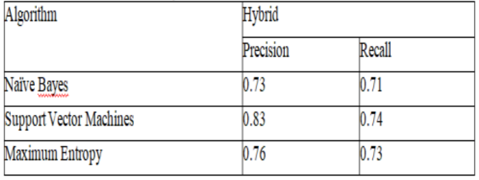
Table 1: Precision and recall for Unigram feature

Algorithm	Unigram	
	Precision	Recall
Naïve Bayes	0.75	0.71
Support Vector Machines	0.82	0.76
Maximum Entropy	0.74	0.70

Table 2: Precision and recall for Unigram feature

Algorithm	Bigram	
	Precision	Recall
Naïve Bayes	0.72	0.70
Support Vector Machines	0.76	0.71
Maximum Entropy	0.73	0.70

Table 3: Precision and recall for Hybrid feature



Algorithm	Hybrid	
	Precision	Recall
Naïve Bayes	0.73	0.71
Support Vector Machines	0.83	0.74
Maximum Entropy	0.76	0.73

The reason we are using R language because when the dataset is large, it is fast and efficient in terms of performing. The packages in the R tool are updated regularly and have greater number of probabilistic and statistical functions.

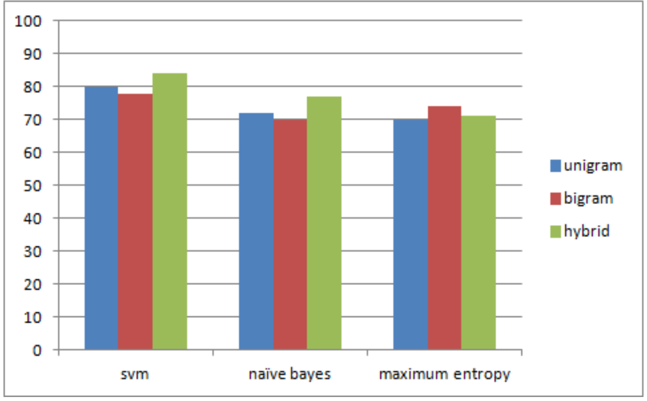


Figure 7: Results of machine learning algorithms

We have obtained the result as hybrid feature with SVM classifier gives the best results for prediction of sentiment of Twitter data. We obtained 84% accuracy using hybrid feature on SVM classified data. 70% is least we have obtained in bigram with Naïve Bayes classifier. Max Ent outcomes Naïve Bayes in bigram feature and thus obtained 74% accuracy. The results can be shown in Figure 7.

VI. CONCLUSION AND FUTURE SCOPE

Conclusion

In this project, we have done comparative analysis on supervised classifiers like Naïve Bayes, Support vector machines and Maximum entropy using unigram, bigram and hybrid (unigram + bigram) feature . There is need to do sentiment analysis as texts in form of messages or posts to find whether the sentiment is negative, positive or neutral. We had extracted data from Twitter i.e. movie reviews for sentiment prediction using machine-learning algorithms. First, we extracted the data from Twitter using Twitter API. Then in pre-processing, we cleaned the data and made the data available to train using classifiers. We have collected 15000 tweets for training set and 2000 tweets for testing set. SVM using hybrid feature outperforms all other classifiers and selection feature with accuracy of 84% .Max Ent surpass Naïve Bayes with bigram feature. Max Ent, on some data sets gives better results than Naïve Bayes. It is concluded that SVM gives better results than other classifiers.

Future Scope

In the future, we are planning to make automatic sentiment classifier for more than one languages starting from the Hindi language because multilingual messages are posted on twitter, so that we will be able to predict the sentiment for any language.

REFERENCES

[1] M. Anjaria and R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning", Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on, pp. 1--8, 2014.

[2] A. Barhan and A. Shakhomirov, "Methods for Sentiment Analysis of Twitter Messages", 12th Conference of FRUCT Association, 2012.

[3] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.", LREc, vol. 10, pp. 1320--1326, 2010.

[4] A. Andrew, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods20016Nello Christianini and John Shawe-Taylor.

[5] Dang, Y.; Zhang, Y.; Chen, H. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. IEEE Intell. Syst. **2010**, 25, 46–53.

[6] Jagdale, O.; Harmalkar, V.; Chavan, S.; Sharma, N. Twitter mining using R. Int. J. Eng. Res. Adv. Tech. **2017**, 3,252–256.

[7] Dubey, G.; Chawla, S.; Kaur, K. Social media opinion analysis for indian political diplomats. In Proceedings of the 2017 7th International Conference on Cloud Computing, Data Science & Engineering, Noida, India,12–13 January 2017; pp. 681–686.