TWITTER SENTIMENT ANALYSIS OF MOVIE REVIEWS

Shalaka Ghanekar, Aishwarya Surve, Prachi Abhyankar
Student, Student, Professor/Guide
Information Technology,
Finolex Academy of Management and Technology, Ratnagiri, India

Abstract : Now a days, online social networks usage are pervasive. Mining text present in online social networks will be useful for predictive analytic. Predicting information from unstructured data present in social networks is a challenging research problem. Extracting, identifying or characterizing the sentiment content of the text unit using statistics and machine learning methods are referred as sentiment analysis or text analysis. In this work sentiment analysis using algorithms such as Decision trees and Support vector machines, which are machine learning algorithms will be demonstrated using tools like Anaconda, WEKA etc. Sentiment analysis using Support Vector Machines (SVM) showed high accuracy when compared to Naïve Bayes.

***IndexTerms* - Twitter sentiment; ensemble learners; twitter sentiment analysis.**

I. INTRODUCTION

Now a day's online reviews has become an increasingly trendy approach for public to share their sentiments and opinions towards the product bought and services received. Online reviews become one of the most important part of any business today. Online reviews presents wealth of information on products and services, if the reviews are properly utilized then it is valuable for vendors in network and social intelligence so that business will be improved. A recent study focused on cost-effective values of online reviews and provides deep understanding between product reviews and their sales performance. People tend to read online reviews understanding the opinions and sentiments and trust them as much as they are recommended by their friends or families. Twitter, a social networking service plays significant role in social networking research. Tweets give rich information about movie, product, or service. Sentiments perform very important role for predicting future sales performance, a mix of good and even bad reviews will create a positive effect on the sales performance and sales prediction. Generally, a binary opposition in opinions is assumed. For/against, like/dislike, good/bad, etc. [4][5][6]
Some sentiment analysis jargon are Semantic orientation and Polarity.

Twitter sentiment analysis is commonly treated as a (supervised) classification problem, where tweets are commonly categorized into four classes--positive, negative, neutral and irrelevant—based on the opinions expressed in a tweet. In practice, the irrelevant class is often consolidated with neutral, thus converting this to a ternary-classification problem. To this end, a wide collection of classification algorithms such as Support Vector Machine (SVM) and Random Forests has been applied. Recently, one sees the emergence of constructing ensemble learners for Twitter sentiment classification. Coupled with feature engineering, ensemble learners have shown to outperform individual learners. [1]

II. CHARACTERISTICS

Characteristics of tweets:

•Length: The determined probable length of a Tweet is 140 letters. Our training data is having a mean length of 82 characters.
•Data availability: Using twitter API we can amass lakhs of tweets for analysis purpose.
•Linguistic model: Users of twitter media are allowed to post their tweets in their native languages. The frequency of misspells is higher in twitter when compared with other domains.
•Domain: Twitter users can post their tweets from several media including PCs, Cell phones etc. Tweets are generated about a variety of things (actually no constraint) which differs from specific domains like movie reviews. We perform sentiment analysis which is the process of finding out whether a piece of statement is neutral, positive or negative and is also referred to as opinion mining that finds out the attitude or opinion of a speaker. To construct a sentiment analyser, the first thing is to get equipped with the correct approach. One such approach is machine learning where we can develop several approaches to classify opinions. We use supervised machine learning algorithms which allows automated aggregated feedback without the need of manual intervention. We use Support Vector Machines (SVM) and Naive Bayes (NB) classifiers to achieve high accuracy. Both of these approaches require training data.

III. RELATED WORK

A comprehensive survey in note that Naïve Bayes and SVM have been the two most commonly used algorithms for twitter sentiment analysis. As for ensemble learners, authors in describe a collection of ensemble classifiers consisting of SVM, Random Forests (RF), Logistic Regression and Multinomial Naïve Bayes (MNB). One of the best ensemble learners includes SVM, RF and MNB using features such as words and other lexical features. In, ensemble methods built on top of Naïve Bayes, Maximum Entropy, Decision Tree, k-Nearest Neighbors and SVM are discussed. The authors however did not evaluate the ensemble learners on Twitter data, but on user reviews. Note that both studies involve intensive feature engineering towards constructing competitive ensemble learners. Comparing our work with the above ensemble methods, this works considers off-the-shelf implementations of four relatively different classifiers, namely, MaxEnt and Naïve Bayes in the Mallet machine learning toolkit, SentiStrength and Pattern. No additional feature engineering is involved. This ensures simplicity and reproducibility of our algorithms. This work also studies how well these classifiers and our ensemble learners adapt to different training datasets. This will help address the lack of labeled data when performing supervised Twitter sentiment analysis. [1].

There are numerous research papers and studies that focus on sentiment classification for Twitter. These studies describe some interesting methodologies of detecting and identifying sentiment from twitter data. Supervised machine learning approach requires an annotated training dataset. The annotation was done based on the presence of either positive smiley emoticons such as ":)" or negative frowning emoticons such as ":(".The process of training the system on the negative and positive sentiment messages alone ignores the important effect of neutral sentiment tweets. This paper includes neutral and irrelevant tweets. Go’s work strips emoticons from the training data, eliminating a rich source of sentiment information. This work includes emoticons for feature extraction. Go’s work also considers the usefulness of different feature sets, including unigram, bigram, unigram + bigram, and parts of speech tags. The results of this study show that the unigram features work very well, with small improvement when bi-grams are also used. We also used unigram as a feature in this paper. Hu and Liu’s system for customer review opinion mining uses a sentiment analysis component based on word lists generated for both positive and negative sentiment orientation. This paper also incorporated these sentiment word lists. According to Liu’s work, word lists of sentiment orientation are useful but insufficient on their own to determine overall text sentiment. We are also not relying just on opinion lexicon and have used additional attributes as features for classification.[3]

Abbreviations and Acronyms

WEKA – Waikato Environment for Knowledge Analysis
SVM- Support Vector Machines

Equations [3]

Equation I - The Precision (P) is calculated as the ratio of instances that were predicted correctly with respect to the predicted size of the instances. It is calculated using the following equation:

$$Precision = \frac{TP}{TP + FP} \qquad (3.0)$$

Equation II - The Recall (R), also known as the True Positive Rate (TPR) is the ratio of correctly predicted instances against the actual number of instances. It is represented using the following equation:

$$Recall = \frac{TP}{TP + FN} \qquad (3.1)$$

Equation III - The accuracy (AC) is the calculated ratio of the total number of correct predictions against the total number of instances in the data set size. It is determined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3.2)$$

Equation IV- F-score (F) is used to express the balance between the recall and precision of a classification run. The efficiency of the accuracy of a test is measured using its F-score. The F-score is calculated as the harmonic mean of recall and precision. It is expressed through following equation:

$$F1 - Score = \frac{Precision + Recall}{2} \qquad (3.3)$$

IV. ALGORITHMS

SVM-

A. Pre-processing

The tweets are extracted for pre-processing. For Sentiment classification, numbers, stop words are not needed. They create noise in data. The pre-processing steps are removing numbers, punctuation, stop words and stemming. Stemming is the process of transforming the word into root word. [2][9][10]

B. POS Tagging

The tweets are POS tagged to extract the frequent adjectives, nouns and adverbs. These are the valuable adaptive sentiment words. The PMI-IR values are calculated for the sentiment words and are assigned to the tweets to get the feature vector. [2]

C. Learning with Unlabelled Data

The feature vectors with the labelled data are given to adaptive training algorithm. The classifier is trained on different topics. The unlabelled data U are chosen such that it has confidence threshold. These unlabelled data are predicted and augmented to L. The new features are updated to train the SVM for next iteration. [2]

D. Calculation of Evaluation of Metrics

The evaluation of the query results is calculated by using equation (I), (II), (III), (IV). The proposed model is evaluated by testing the model by providing the test data. The metrics are calculated for the model and results are found. [2]

E. Performance Analysis

Step length is the maximum number of most confident unlabelled data that can be chosen from the test dataset.. The graph shows that even the step length increases the performance of the algorithm better. As the training data size increases the new feature values are updated for accurate classification. [2][7][8]

There are four main advantages:

- It has a regularisation parameter, which makes the user think about avoiding over-fitting.
- It uses the kernel trick, so that you can build in expert knowledge about the problem via engineering the kernel.
- An SVM is defined by a convex optimisation problem for which there are efficient methods.
- It is an approximation to a bound on the test error rate, and there is a substantial body of theory behind it which suggests it should be a good idea.

The disadvantages are:

The theory only really covers the determination of parameters for a given value of the regularisation and kernel parameters and choice of kernel. In a way the SVM moves problem of over-fitting from optimising the parameters to model selection. [2]

V. ARCHITECTURE DIAGRAM

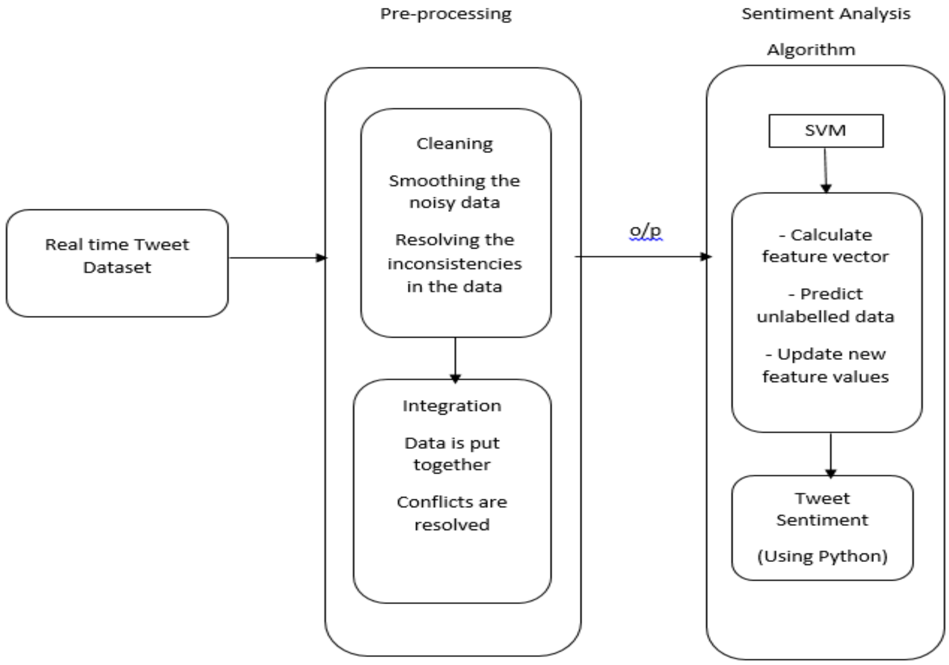


Fig 5.1. Architecture diagram

VI. IMPLEMENTATION

Extraction:

Twitter Archiver lets you easily save tweets for any search keyword or hashtag in a Google Spreadsheet. Enter a search query, or hashtag, and all matching tweets are automatically saved in the Google Sheet. Use boolean search or include advanced Twitter search operators to create more complex queries.

Twitter Archiver can be used for saving tweets around trending hashtags, conference tweets, your brand mentions, geo-tagged tweets, and more. It polls twitter every hour and pulls all the matching tweets into the Google Spreadsheet.

Pre-processing:

1. Load the raw text:
   The text is small and will load quickly and easily fit into memory. This will not always be the case and you may need to write code to memory map the file. Tools like NLTK will make working with large files much easier.
2. Split into tokens:
   Split the document by white space, including " ", new lines, tabs and more. We can do this in Python with the split () function on the loaded string.
3. Convert to lowercase:
   Converting tweets to lowercase will ease the work of pre-processing.
4. Remove punctuation from each token:
   Python provides a constant called string.punctuation that provides a great list of punctuation characters.
5. Filter out remaining tokens that are not alphabetic

6. Filter out tokens that are stop words:

Stop words do not contribute to the deeper meaning of the phrase. They are the most common words such as: "*the*", "*a*", and "*is*".
After applying SVM algorithm we will get the final output i.e. Classification of tweets.
SVM:

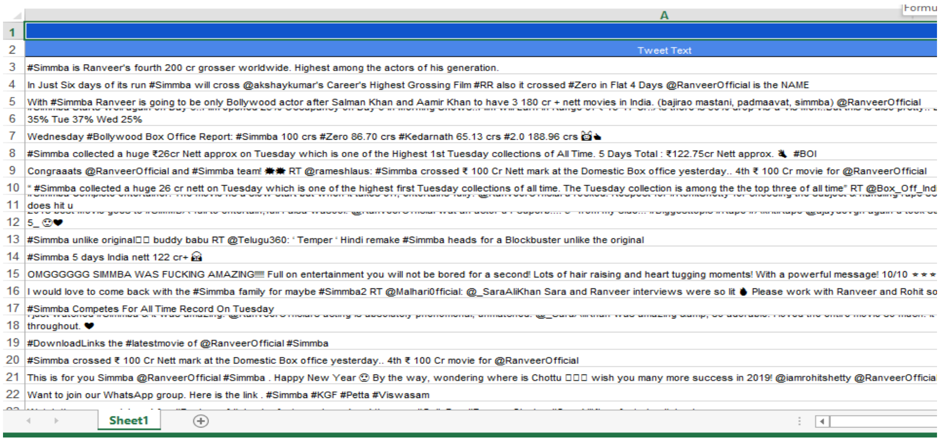
1. Install scikit_learn, nltk, scipy, numpy.
2. Import sys, time, re, nltk, joblib.
3. Text feature set X and feature values;
4. Non-text feature set X1 and feature values;
5. L: labeled tweets containing K sentiment classes on mixing topics;
6. U: unlabeled tweets on topic
7. t: classification confidence threshold;
8. l: the maximum number of unlabeled tweets selected in one iteration;
9. M: the maximum number of co-training iterations
10. Loop {
    a. Repeat
    b. [Adapting to unlabeled data]
    c. Calculate confidence scores
    d. Select the l most confident unlabeled tweets
    e. Move them with predicted class labels from U to L
    f. [Adapting to features]
    g. Calculate the current feature vector X
    h. Select the c most significant and topic adaptive sentiment words for each class
    i. Update the feature values }
11. End loop
12. Train the multiclass SVM C on the features consists of X and X1 using augmented labeled data L.
13. Return X X1 C [3]
14. After execution, the positive tweets are labelled 1 and negative tweets are labelled 0.

AbbreviationsandAcronyms

SVM- Support Vector Machine

IV. RESULTS AND DISCUSSION

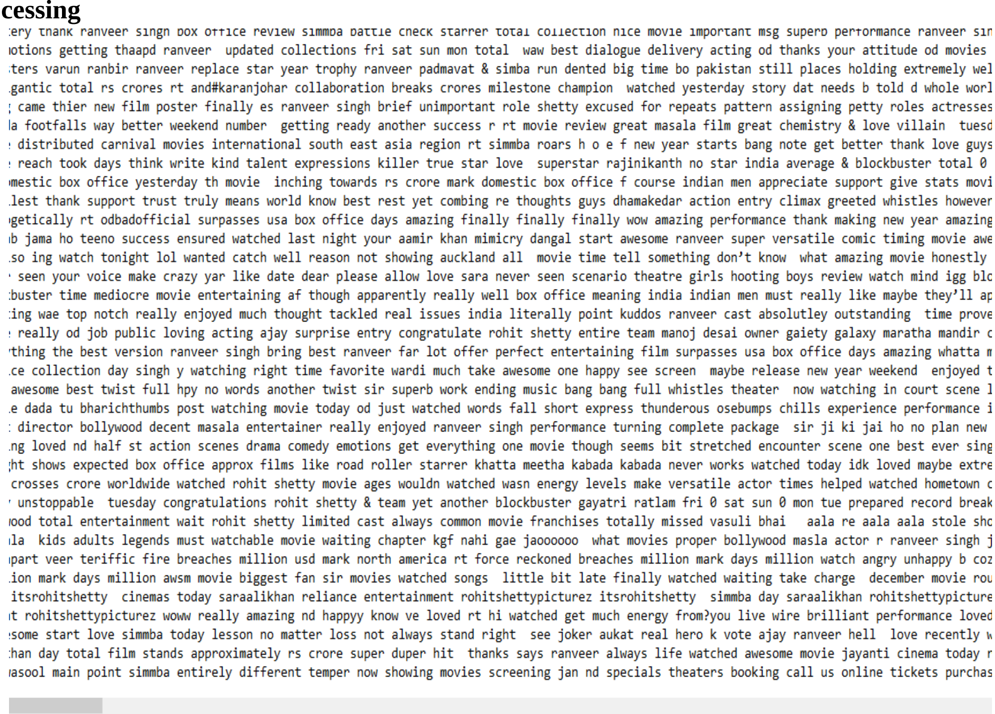
4.1 After Extraction



	A
1	
2	Tweet Text
3	#Simmba is Ranveer's fourth 200 cr grosser worldwide. Highest among the actors of his generation.
4	In Just Six days of its run #Simmba will cross @akshaykumar's Career's Highest Grossing Film #RR also it crossed #Zero in Flat 4 Days @RanveerOfficial is the NAME
5	With #Simmba Ranveer is going to be only Bollywood actor after Salman Khan and Aamir Khan to have 3 180 cr + nett movies in India. (bajirao mastani, padmaavat, simmba) @RanveerOfficial
6	35% Tue 37% Wed 25%
7	Wednesday #Bollywood Box Office Report: #Simmba 100 crs #Zero 86.70 crs #Kedarnath 65.13 crs #2.0 188.96 crs
8	#Simmba collected a huge ₹26cr Nett approx on Tuesday which is one of the Highest 1st Tuesday collections of All Time. 5 Days Total : ₹122.75cr Nett approx. #BOI
9	Congraaats @RanveerOfficial and #Simmba team! RT @rameshlaus: #Simmba crossed ₹ 100 Cr Nett mark at the Domestic Box office yesterday.. 4th ₹ 100 Cr movie for @RanveerOfficial
10	" #Simmba collected a huge 26 cr nett on Tuesday which is one of the highest first Tuesday collections of all time. The Tuesday collection is among the the top three of all time" RT @Box_Off_Ind
11	does hit u
12	5_ 😊❤
13	#Simmba unlike original□□ buddy babu RT @Telugu360: ' Temper ' Hindi remake #Simmba heads for a Blockbuster unlike the original
14	#Simmba 5 days India nett 122 cr+
15	OMGGGGGG SIMMBA WAS FUCKING AMAZING!!!!! Full on entertainment you will not be bored for a second! Lots of hair raising and heart tugging moments! With a powerful message! 10/10 ★★★
16	I would love to come back with the #Simmba family for maybe #Simmba2 RT @Malhari0fficial: @_SaraAliKhan Sara and Ranveer interviews were so lit 🔥 Please work with Ranveer and Rohit so
17	#Simmba Competes For All Time Record On Tuesday
18	throughout. ❤
19	#DownloadLinks the #latestmovie of @RanveerOfficial #Simmba
20	#Simmba crossed ₹ 100 Cr Nett mark at the Domestic Box office yesterday.. 4th ₹ 100 Cr movie for @RanveerOfficial
21	This is for you Simmba @RanveerOfficial #Simmba . Happy New Year 😊 By the way, wondering where is Chottu □□□ wish you many more success in 2019! @iamrohitshetty @RanveerOfficia
22	Want to join our WhatsApp group. Here is the link . #Simmba #KGF #Petta #Viswasam

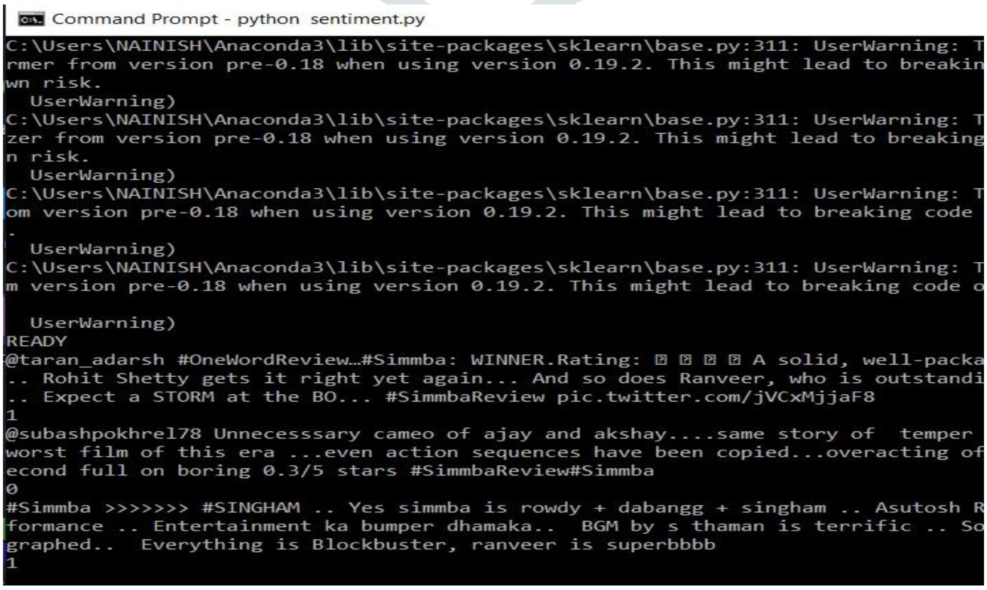
Fig. 4.1: Extracted tweets

4.1 After Pre-processing



```
:ery thank ranveer singh box office review simmba battle check starrer total collection nice movie important msg superb performance ranveer sin
iotions getting thaapd ranveer  updated collections fri sat sun mon total  waw best dialogue delivery acting od thanks your attitude od movies
:ters varun ranbir ranveer replace star year trophy ranveer padmavat & simba run dented big time bo pakistan still places holding extremely wel
.gantic total rs crores rt and#karanjohar collaboration breaks crores milestone champion  watched yesterday story dat needs b told d whole worl
; came thier new film poster finally es ranveer singh brief unimportant role shetty excused for repeats pattern assigning petty roles actresses
la footfalls way better weekend number  getting ready another success r rt movie review great masala film great chemistry & love villain  tuesd
: distributed carnival movies international south east asia region rt simmba roars h o e f new year starts bang note get better thank love guys
: reach took days think write kind talent expressions killer true star love  superstar rajinikanth no star india average & blockbuster total 0
imestic box office yesterday th movie  inching towards rs crore mark domestic box office f course indian men appreciate support give stats movi
.lest thank support trust truly means world know best rest yet combing re thoughts guys dhamakedar action entry climax greeted whistles however
igetically rt odbadofficial surpasses usa box office days amazing finally finally finally wow amazing performance thank making new year amazing
ib jama ho teeno success ensured watched last night your aamir khan mimicry dangal start awesome ranveer super versatile comic timing movie awe
.so ing watch tonight lol wanted catch well reason not showing auckland all  movie time tell something don't know  what amazing movie honestly
' seen your voice make crazy yar like date dear please allow love sara never seen scenario theatre girls hooting boys review watch mind igg blc
:buster time mediocre movie entertaining af though apparently really well box office meaning india indian men must really like maybe they'll ap
:ing wae top notch really enjoyed much thought tackled real issues india literally point kuddos ranveer cast absolutley outstanding  time prove
: really od job public loving acting ajay surprise entry congratulate rohit shetty entire team manoj desai owner gaiety galaxy maratha mandir c
'thing the best version ranveer singh bring best ranveer far lot offer perfect entertaining film surpasses usa box office days amazing whatta m
.ce collection day singh y watching right time favorite wardi much take awesome one happy see screen  maybe release new year weekend  enjoyed t
 awesome best twist full hpy no words another twist sir superb work ending music bang bang full whistles theater  now watching in court scene l
.e dada tu bharichthumbs post watching movie today od just watched words fall short express thunderous osebumps chills experience performance i
: director bollywood decent masala entertainer really enjoyed ranveer singh performance turning complete package  sir ji ki jai ho no plan new
.ng loved nd half st action scenes drama comedy emotions get everything one movie though seems bit stretched encounter scene one best ever sing
;ht shows expected box office approx films like road roller starrer khatta meetha kabada kabada never works watched today idk loved maybe extre
 crosses crore worldwide watched rohit shetty movie ages wouldn watched wasn energy levels make versatile actor times helped watched hometown c
' unstoppable  tuesday congratulations rohit shetty & team yet another blockbuster gayatri ratlam fri 0 sat sun 0 mon tue prepared record break
iood total entertainment wait rohit shetty limited cast always common movie franchises totally missed vasuli bhai   aala re aala aala stole shc
ila  kids adults legends must watchable movie waiting chapter kgf nahi gae jaoooooo  what movies proper bollywood masla actor r ranveer singh j
ipart veer teriffic fire breaches million usd mark north america rt force reckoned breaches million mark days million watch angry unhappy b coz
.ion mark days million awsm movie biggest fan sir movies watched songs  little bit late finally watched waiting take charge  december movie rou
 itsrohitshetty  cinemas today saraalikhan reliance entertainment rohitshettypicturez itsrohitshetty  simmba day saraalikhan rohitshettypicture
it rohitshettypicturez woww really amazing nd happyy know ve loved rt hi watched get much energy from?you live wire brilliant performance loved
:some start love simmba today lesson no matter loss not always stand right  see joker aukat real hero k vote ajay ranveer hell  love recently w
:han day total film stands approximately rs crore super duper hit  thanks says ranveer always life watched awesome movie jayanti cinema today r
iasool main point simmba entirely different temper now showing movies screening jan nd specials theaters booking call us online tickets purchas
```

4.2 After sentiment analysis



```
C:\Users\NAINISH\Anaconda3\lib\site-packages\sklearn\base.py:311: UserWarning: T
rmer from version pre-0.18 when using version 0.19.2. This might lead to breakin
wn risk.
  UserWarning)
C:\Users\NAINISH\Anaconda3\lib\site-packages\sklearn\base.py:311: UserWarning: T
zer from version pre-0.18 when using version 0.19.2. This might lead to breaking
n risk.
  UserWarning)
C:\Users\NAINISH\Anaconda3\lib\site-packages\sklearn\base.py:311: UserWarning: T
om version pre-0.18 when using version 0.19.2. This might lead to breaking code
.
  UserWarning)
C:\Users\NAINISH\Anaconda3\lib\site-packages\sklearn\base.py:311: UserWarning: T
m version pre-0.18 when using version 0.19.2. This might lead to breaking code o

  UserWarning)
READY
@taran_adarsh #OneWordReview…#Simmba: WINNER.Rating: ? ? ? ? A solid, well-packa
.. Rohit Shetty gets it right yet again... And so does Ranveer, who is outstandi
.. Expect a STORM at the BO... #SimmbaReview pic.twitter.com/jVCxMjjaF8
1
@subashpokhrel78 Unnecesssary cameo of ajay and akshay....same story of  temper
worst film of this era ....even action sequences have been copied...overacting of
econd full on boring 0.3/5 stars #SimmbaReview#Simmba
0
#Simmba >>>>>>> #SINGHAM .. Yes simmba is rowdy + dabangg + singham .. Asutosh R
formance .. Entertainment ka bumper dhamaka..  BGM by s thaman is terrific .. So
graphed..  Everything is Blockbuster, ranveer is superbbbb
1
```

Fig. 4.2 Analysed tweets

VII. Acknowledgment

We would like to express our special thanks of gratitude to our project guide as well as our HOD Dr. Vinayak Bharadi who gave us the golden opportunity to do this wonderful project on the topic Twitter Sentiment Analysis of Movie Reviews. Secondly we would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited time frame.

References

[1] Yeqing Yan, Hui Yang, Hui-ming wang, “Two Simple and Effective Ensemble Classifiers for Twitter Sentiment Analysis”, in Computing Conference 2017, London, UK.

[2] K Lavanya, C Deisy “Twitter Sentiment Analysis Using Multi-Class SVM” in International Conference on Intelligent Computing and Control 2017, (I2C2’17)

[3] Ajay Deshwal, Sudhir Kumar Sharma, “Twitter Sentiment Analysis using Various Classification Algorithms”, Department ofComputer Science and Engineering KIIT college of Engineering, Gurgaon , India 2016.

[4] D. Gayo-Avello, "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" -- A Balanced Survey on Election Prediction using Twitter Data,http://arxiv.org/abs/1204.6441,April 2012

[5] A. Lamb, M. J. Paul, and M. Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. In Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 789–795, June 2013.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In Int’l Conference on World wide web (WWW), pages 851–860, 2010

[7] Kang Liu, Liheng Xu, and Jun Zhao, Co-Extracting Opinion Targets and Opinion Wordsfrom Online Reviews Based on the Word Alignment Model IEEE Transactions on Knowledgeand Data Engineering, Vol. 27, No. 3, March 2015.

[8] Duyu Tang, Bing Qin, Furu Wei, Li Dong, Ting Liu, and Ming Zhou, A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification, IEEE/ACMTransactions on Audio, Speech, and Language Processing, Vol. 23, No. 11, November 2015.

[9] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, Dual Sentiment Analysis: Considering Two Sides of One Review, IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 8, August 2015

[10] Eduard C. Dragut, Hong Wang, Prasad Sistla, Clement Yu, and Weiyi Meng, Polarity Consistency Checking for Domain Independent Sentiment Dictionaries, IEEE Transactionson Knowledge and Data Engineering, Vol. 27, No. 3, March 2015.