International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 1524-1529 www.ijaem.net ISSN: 2395-5252

Twitter Analyzer: Twitter Trend Detection and Visualization

Shaikh Abdul Majid, Shaikh Shoaib, Dr. Shabina Sayed
Information Technology Department MH Saboo Siddik,Mumbai
AssistantProf. Information Technology Department MH Saboo Siddik,Mumbai

Submitted: 15-05-2021 Revised: 26-05-2021 Accepted: 28-05-2021

ABSTRACT: Twitter is most popular social media that allows its user tospread and share information.It Monitors their user postingsanddetectmostdiscussedtopicofthemovement.Theypublish these topics on the list called "Trending Topics". Itshowwhatishappeningintheworldandwhatpeople's opinions are about it. For that it uses top 10 trending topic list.Some topic will trend at some point in the future and otherswillnot. We wishtopredictwhichtopics willtrend. Andapply algorithm to find out what public opinion about thattopic which use to predict mood. In this paper, we proposemodelwhichusemachinelearningalgorithmandclassifysentimentoftwittermessage.Forthatwecollecttweet,preprocessthattweet,findtrendingtopicandapplymulticlassifier algorithm which predict public mood. We are goingto use different measure such as precision, recall, F-measure.Wewill goingtoachieve betteraccuracy.

GeneralTerms Machine learning algorithm, information retrieval,classification.

Keywords: Socialmedia,Twitter,TwitterTrendingTopic,TopicDetection,Textmining,Polaritydetection.

I. INTRODUCTION

Social media is a rich resource of information about actualworld action of all type twitter is one of them. It is mostpopular micro blogging site which allow their user to shareinformation and short message which is called tweet. Wheremillions of people tweet every day. Twitter exchange wide variety of local and real-world event. Twitterhavin gtwofeatures[2]:

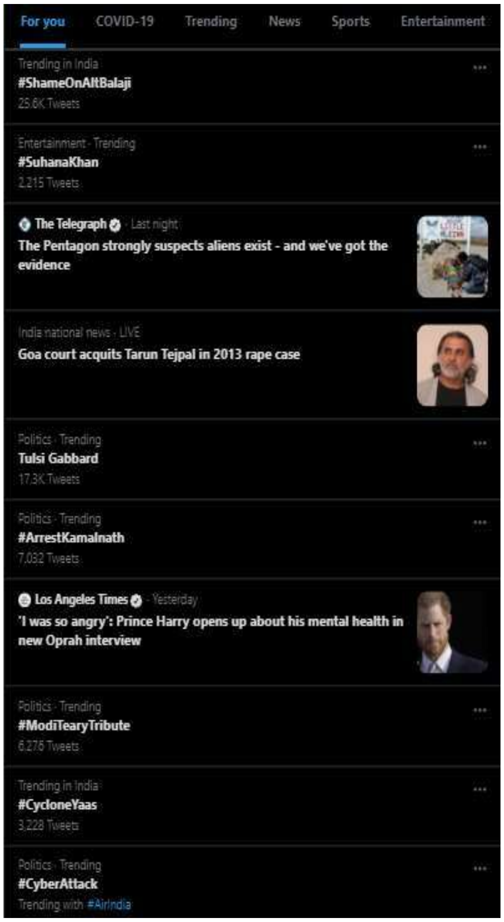
- The shortness of tweets, which cannot go beyond140 characters, it facilitates Creation and sharing ofmessagesinafew seconds
- Easiness of spreading message to a large number ofuserwithinlittle time.

Twitterhasstandardsyntax whichlisted follow[3]:

- UserMentions:whenausermentionsanotheruserintheirtweet,Place@-signbeforethecorrespondingusername.Like@Username
- Retweets: Re-share of a tweet which is posted byanotherusercalledretweet.Bycopingoriginaltweetuserconsiderthatmessageofinteresttoother.
- Replies: when a user wants to reply an earlier tweet,they place the @username mention at the beginningof the tweet, e.g., @username I have question onwhatyousay.
- Hashtags:Hashtagsincludedinatweettendtogroup tweets in conversations or represent the mainterms of the tweet, it usually referred to topics orcommoninterestsofacommunity.Itisdifferentiated from the rest of the terms in the tweetin that ithasa leadinghash,e.g.,#hashtag.

Twitter gives list of most discussedtopic at the movement which iscalled"Trendingtopic".Itshowswhatpeoplediscussing what is goingontheirmind.

Followingimage showshowtrend showson twitter:-



Trending in India
#ShameOnAltBalaji
25.6K Tweets

Entertainment · Trending
#SuhanaKhan
2,215 Tweets

The Telegraph · Last night
The Pentagon strongly suspects aliens exist - and we've got the evidence

India national news · LIVE
Goa court acquits Tarun Tejpal in 2013 rape case

Politics · Trending
Tulsi Gabbard
17.3K Tweets

Politics · Trending
#ArrestKamalnath
7,032 Tweets

Los Angeles Times · Yesterday
'I was so angry': Prince Harry opens up about his mental health in new Oprah interview

Politics · Trending
#ModiTearyTribute
6,276 Tweets

Trending in India
#CycloneYaas
3,228 Tweets

Politics · Trending
#CyberAttack
Trending with #AirIndia

Fig1:TwitterTopTrendlist

In this paper we propose model which is use to predict publicopinionwhattheytalkingabout.Wecanpredictpolarityaboutdifferentevents,sports,Economy,politics etc.Wecollecttweetsaboutparticulareventandpredict publicopinionabout thateventforthatfirstwe have todo pre-processingoftweetsthenapplyfeatureextractionand find out polarity by applying machine learning algorithm. For polaritydetectionwecanusetwotypeofclassification. Binaryclassification andmulticlassclassification. In binary classification we have to predict public opinion intwo category like positive or negative. Where is multiclassclassification, we can use more than two category like positive,negative,neutral.

II. LITERATURESURVEY

Trend analysis and based on that predicting public opinions. Itplays important role, many researchers working on automatictechnique of extraction and analysis of huge amount of twitterdata. In [1] author compare six trend detection method andfindthatstandardnaturallanguageprocessingtechniqueperformwellforsocialstreamsonparticulartopic. Theyconclude that n-gram give best performance other than state-of-art techniques. In [4], the authors have used three differentmachine learning algorithms Naïve Bayes, Decision Trees andSupport Vector Machine for sentiment classification of Arabicdataset which was obtained from twitter. This research hasfollowedaframeworkforArabictweetsclassificationinwhichtwospecialsub-taskswereperformedinpre-processing,TermFrequency-InverseDocumentFrequency(TF-IDF) and Arabic stemming. They have used one datasetwith three algorithms and performance has been evaluated onthebasisthreedifferentinformationretrievalmetrics precision,recall,andf-measure.In[6]authorproposedsupervisedlearningtechniquestoclassifytwittertrendingtopic for that they use text based and network based classifierandconcludeC5.0gavebestperformance.In[19]authorpropose model which predict public opinion on political eventbyApplingdifferentclassifierwhichpredictthat whethermood is positive or negative. In [26], the authors proposed away to get the pre labeled data from twitter which can be usedto train SVM classifier. They used the twitter hash tags tojudgethepolarityoftweet.Toanalyzetheaccuracyofproposedtechnique,ateststudyontheclassifierwasconductedwhich showed theresultwith theaccuracyof85%.

The authors in [27] introduced a new technique to classify thesentimentoftweetsaspositiveornegative.Theypresentedand discussed the results of machine learning algorithms fortwittersentimentanalysisbyusingdistantsupervision.Trainingdata,theauthorsusedconsistedoftweetswithemotionswhichwereusedasnoisylabels.Accordingtoauthors,themachinelearningalgorithmssuchasNaiveBayes,MaximumEntropyandSVMwhentrainedwithemotion tweets can have accuracy more than 80%. The studyalsohighlightedthestepsusedinpreprocessingstageofclassification for high accuracy. In [28] sentiment analysisisperformusingSVMinthattwopreclassifieddatasetsoftweetsareusedthendocomparativeanalysis,theyusemeasuresPrecision,Recall andF-Measure.

III. TOOLSANDTECHNOLOGY

In proposed model coding is a Web Page

that displays the real time tweet and calculates the sentiment using Machine Learning Algorithms.
python for we have to installpython, anaconda.
Python is an interpreted, object-oriented,high-levelprogramming languagewithdynamicsemantics.Pythonsupportsmodulesandpackages,whichencouragesprogram modularity and code reuse. Anaconda is a freemiumopen source distribution of the Pythonlanguageforlarge-scaledataprocessing, predictive analytics, andscientificcomputing,thataimstosimplifypackage managementanddeployment.Inanacondaweusejupyternotebook.
TheJupyterNotebookisanopen-sourcewebapplication that allows you to create and share documents thatcontain live code, equations, visualizations and narrative text.Usesinclude:datacleaningandtransformation,numericalsimulation, statistical modelling, data visualization, machinelearning,andmuchmore.

IV. PROPOSEDMODEL

Descriptionaboutmodelwhichweareproposedasgivenbelow.
The modelhavingfollowingsteps:

- Datacollectionoftweets
- Pre-processtweet
- FeatureExtraction
- TrendDetection

CalculatemoodTendency(Positive,Negative,andNeutral).Followingfigureshowsproposedmodel:-

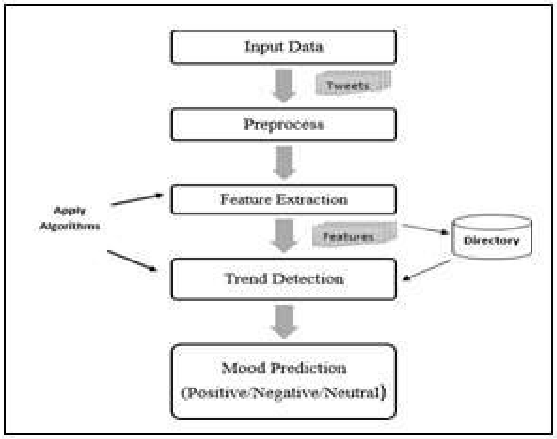


Fig2:ProposedmodelforTrenddetectionandpolaritydetection

1) Dataset:
CollecttweetdatathroughtwitterstreamingAPI.Whichdownload tweets in JSON format. We can apply keyword,hashtag,usernameto downloadtweetsrelated to them.
2) Pre-processing:
Tweetpre-processingmodulehavingseveralstages.Afterdownloading tweets we have to extract text data form that anddiscard video, audio, image etc .store English text which isretrieveformtweet.Thenremove@,#,urlandotherpunctuation form tweets and apply stop word remove, wordtokenize.
1) FeatureExtraction:
After pre-processing stage next module is Feature extractionwhich is done in two way through Term frequency calculationand pos tagging
2) TrendDetectionand MoodPrediction
We can determine trend by using TF-IDF calculation. Andpredict positive, negative, neutral mood tendency by applyingmachinelearningalgorithms.Applysentimentclassification.

3) FeatureExtraction:
After pre-processing stage next module is Feature extractionwhich is done in two way through Term frequency calculationand pos tagging
4) TrendDetectionand MoodPrediction
We can determine trend by using TF-IDF calculation. Andpredict positive, negative, neutral mood tendency by applyingmachinelearningalgorithms.Applysentimentclassification.

V. CLASSIFICATIONTECHNIQUES

There are different types of classifiers that are generally usedfortextclassificationwhichcanbealsousedfortwit tersentimentclassification.

A. SVMClassifier[24]

ThemaingoalofSupportVectorMachineis tomaximizemargin. SVM separates the tweets using a hyper plane. SVMusesadiscriminativefunctiondefinedas

$g(\mathrm{X})=w^{T}\text{Ø}(\mathrm{X})+b$ (1)

'X' is the feature vector, 'w' is the weights vector and 'b' isthe bias vector. 'w' and 'b' are learned automatically on thetrainingset.

SVM having hard margin and Soft margin. There are linearlyseparablemethodandNon-linearseparablemethod.Forlinearlyseparablemethod wehavefollowingequation[22]:

$f(x)=\sum \alpha_i y_i \mathrm{X}^T\mathrm{X}$ (2)

i

Where $\alpha_i$is Lagrange multiplier, $y_i$is class and $x_i$is input.This is Equation for Hard margin and for soft margin we useslackvariable.

For non- linearly separable method we use different kerneltrickslikelinear, polynomial,radial basisfunctionetc.

B. NaveBayesClassifier[24]

Nave Bayes is probabilistic model [7]. This Classifier makesuse of all the features in the feature vector and analyzes themindividually asthey are equally independentof ~~ea~~chother.Thecon~~di~~tionalprobabilityforNaiveBayes can bedefinedas

Inlogisticregression,thedependentvariableisbinaryor dichotomous, i.e. it only contains data coded as 1 (TRUE,success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant,etc.).

The goal of logistic regression is to find the best fitting (yetbiologicallyreasonable)modeltodescribetherelationshipbetween the dichotomous characteristic of interest (dependentvariable=responseoroutcomevariable)andasetofindependent(predictororexplanatory)variables.Logisticregression generates the coefficients (and its standard errorsandsignificancelevels)ofaformulatopredictalogittransformationoftheprobabilityofpresenceofthecharacteristicofinterest:

$logit(p)=b_0+b_1\mathrm{X}_1+b_2\mathrm{X}_2+\cdots+b_k\mathrm{X}_k$(4)

Where p is the probability of presence of the characteristic ofinterest.Thelogittransformationisdefinedasthelog gedodds:

odds=p/(1-p)=(Probabilityofpresenceofcharacteristic)/(Probabilityofabsenceofcharacteristic)

And

$logit(p)=$
$ln\ \underline{p}$
( ) (5)D. Decision
$1-p$
Tree

Decision tree [24] builds classification models in the form of atreestructure.Itbreaksdownadatasetintosmallerandsmaller subsets while at the same time an associated decisiontree is incrementally developed. The final result is a tree withdecisionnodesandleafnodes.Adecisionnode(e.g.,Outlook) has two or more branches (e.g., Sunny, Overcast andRainy). Leaf node (e.g., Play) represents a classification ordecision.Thetopmostdecisionnodeinatreewhichcorresponds to the best predictor called root node. Decisiontrees can handle both categorical and numerical data. C4.5 isan algorithmusedtogeneratea decisiontree.

E)KNNclassifier

K nearest neighbors [24] is a simple algorithm that stores allavailablecasesandpredictthenumericaltargetbased ona similaritymeasure(e.g.,distancefunctions).

X m x_i

$P(_y)=$G $P(_y)$

(3)

The algorithm assumes that it is possible to classify

i i=1 j

'X' is the feature vector definedas X= $\{x_1, x_2 \ldots x_m\}$andyj is the class label. Here, in our work there are differentindependent features like emoticons, emotional Keyword,countofpositiveandnegativekeywords,andcountofpositive and negative hash tags which are effectively utilizedby Naïve Bayes classifier for classification. Nave Bayes doesnot consider the relationships between features. So it cannotutilize the relationships between part of speech tag, emotionalkeywordandnegation.

C. LogisticClassifier

Logistic regression [25] is a statistical method for analyzing adataset in which there are one or more independent variablesthat determine an outcome. The outcome is measured with adichotomous variable (in which there are only two possibleoutcomes).

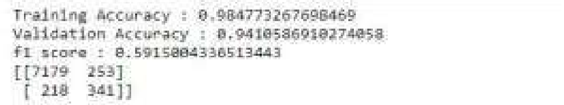
documentsintheEuclideanspaceaspoints.Euclideandistanceisthedistancebetweentwopointsin Euclideans pace.Thedistancebetweentwopointsinthe planewithcoordinatesp=(x, y)andq=(a, b)canbecalculated

$$d(p,a)=\sqrt{}(x-a)^2+(y-b)^2 \quad (6)$$

VI. IMPLEMENTATION AND RESULTS

Dataset having 40000 tweets after pre-processing we have 38000tweets.Thenapplydifferentclassifier which generate results. Results having informationretrieval measure like Precision, Recall, F-measure, accuracy,Rootmeansquarederroretc. Resultsareshown asbelow:

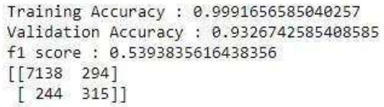
Logistic Classifier Results:



```
Training Accuracy : 0.984773267698469
Validation Accuracy : 0.9410586910274058
f1 score : 0.5915004336513443
[[7179  253]
 [ 218  341]]
```

Fig3:Logistic Classifier Results

Informationretrievalmeasure:Thisfieldhavingdifferentmeasureslikeprecision,recall,F-measure,accuracywecompare them and analysis their results based on the graphwhich areshownas below:



```
Training Accuracy : 0.9991656585040257
Validation Accuracy : 0.9326742585408585
f1 score : 0.5393835616438356
[[7138  294]
 [ 244  315]]
```

Fig4:Support Vector Machine Classifier Results

VII. CONCLUSION

Tweet having short message we use that for predicting publicopinionsonsports,Economy,ongoingeventsetc.Wearefinding keyword in tweet andpredict whether it is havingweightage positive or negative by applying machine leaningalgorithms. We can apply multi classification algorithms likeSVM,NaïveBayes,Logisticclassification,KNNandDecision tree. We observe that Information retrieval measureslike precision, recallandF-measure. We get results sobyobservingtheresultswecansaySVMhavinglessmeansquare error so it is good classifier for this type of dataset. Infuturewecantestthiswithpythoncodingandfindbest classifier.

REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin,

DavidCorney,SymeonPapadopoulos,RyanSkraba,AyseGöker, Ioannis Kompatsiaris, Senior Member “SensingTrending Topics in Twitter” IEEE, and Alejandro JaimesIEEETransactionsOnMultimedia,Vol. 15,No.6,October2013.

[2] Soyeon Caren Han, Hyunsuk Chung, Do Hyeong Kim,Sungyoung Lee, and Byeong Ho Kang “Twitter TrendingTopics Meaning Disambiguation” Springer InternationalPublishingSwitzerland2014.

[3] ArkaitzZubiaga,DamianoSpina,RaquelMart´ınez,V´ıctorFresno“Real-TimeClassificationofTwitterTrends” Journal of the American Society for InformationScienceandTechnologycopyright @2013.Sentiment Analysis” International Journal on AdvancedScience, Engineering and Information Technology, 6(6),1067-1073.

[4] Altawaier,M.M.,&Tiun,S.(2016)“ComparisonofMachineLearningApproachesonArabicTwitter

[5] Rong Lu and Qing Yang, “Trend Analysis of News TopicsonTwitter”,International Journal of Machine Learningand ComputingVol.2,No.3, June2012

[6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md.Mostofa Ali Patwary, Ankit Agrawal, Alok Choudhary,“Twitter Trending Topic Classification” 2011 11th IEEEInternationalConferenceon DataMining.

[7] ErwinB.Setiawan,DwiH.Widyantoro,KridantoSurendro,“FeatureExpansionusing WordEmbeddingforTweet Topic Classification”IEEE,2016.

[8] http://www.socialmediatoday.com/social-networks/heres- why-twitter-so-important-everyone

[9] http://www.newsmedialive.com/wpcontent/uploads/2015/1 0/TWITTER.jpg

[10] http://www.twitter.com

[11] YubaoZhang,StudentMember,IEEE,XinRuan,Student Member, IEEE, Haining Wang, Senior Member,IEEE,HuiWang,andSuHe“TwitterTrendsManipulation:AFirstLookInsidetheSecurityofTwitterTrending”IEEEtransactionsoninformationforensicsand security, vol. 12,no.1,january2017.

[12] Amina Madani, Omar Boussaid,Djamel

Eddine Zegour"Real-timetrendingtopicsdetectionanddescriptionfromTwitter content" Springer-2015.

[13] ArkaitzZubiaga,DamianoSpina,RaquelMartinez,VictorFresno,"Real-TimeClassificationofTwitterTrends" American Society for Information Science andTechnology2013.

[14] Arkaitz Zubiaga, Damiano Spina, Víctor Fresno, RaquelMartínez "Classifying Trending Topics: A Typology ofConversation TriggersonTwitter"ACM 2011.

[15] María del Pilar Salas-Zárate, José Medina-Moreira, PaulJavier Álvarez-Sagubay "Sentiment Analysis and TrendDetection inTwitter"Springer 2011.

[16] https://statinfer.com/204-6-8-svm-advantages-disadvantages-applications/?c=361cde8465e4

[17] https://www.slideshare.net/ashrafmath/naive-bayes- 15644818

[18] http://www2.cs.man.ac.uk/~raym8/comp37212/main/node 264.html

[19] A. Hernandez-Suarez, G. Sanchez-Perez, V. Martinez-Hernandez,H.Perez-Meana,K.Toscano-Medina,M.NakanoandV.Sanchez"Predicting PoliticalMoodTendenciesbasedonTwitterData"

[20] http://www.kdnuggets.com/2017/06/which-machine- learning-algorithm.html