Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

*Special Issue for ICPCDECT 2016,Volume 3 Issue 1*

# Sentiment Analysis of Tweets using Apache Flume andSpark

Mayur More[1],MonikaDarekar[2], KomalDeshmukh[3]

*[1] Computer Engineering, AISSMS-IOIT,mmore1315@gmail.com[2]Computer Engineering,AISSMS-IOIT,darekarmonika24@gmail.com[3]ComputerEngineering,AISSMS-IOIT,komaldeshmukh1995@gmail.com*

**Abstract**–

*Throughsocialmediatheuserscansharetheirthoughtswithfriends,family,andcolleagues,anditalsogivestheuseraplatform totalkandcommunicateontheirfavoritetopics.This"unstructured"conversationcangive businesses valuable insight into how consumers perceive their brand, and allow them to actively makebusinessdecisions to maintain their image. With a rapid increasing of data of sentiments in social media on web has leadtheresearchers into increased interests regarding opinion mining and sentiment analysis. However, SentimentAnalysisis now considered as a Big Data task due to the large amount of social media available on theweb.*

*To find a technique such that it can efficiently perform sentiment analysis on big data sets was the main focusofthis research. In this paper, Hadoop Apache ecosystem's data ingestion tool was used to perform SentimentAnalysison the large sets of data consisting tweets and stream processing with Spark. Using this technique theexperimentalresult shows very good efficiency in handling big data sets ofsentiment.*

***Keywords***- *Sentiment Analysis, Opinion Mining, Apache Spark, Apache Flume, MachineLearning*

## I. INTRODUCTION

Sentiment analysis is the field which deals with computational treatment of sentiment, opinionandsubjectivityintext.Sentimentscanbedescribedasideas,opinions,judgmentsoremotionspromptedbyemotions. World Wide Web has completely changed the way of expressing people's views. Now a day's people areexpressingtheirviewsandthoughtsthroughonlineblogs,discussionformsandalsosomeOnlineapplicationslikeFaceboo k,Twitter,etc.IfwetakeTwitterasourexamplenearly1TBoftextdataisGeneratingwithinaweekintheformoftweets. So, by this it is understand clearly how this Internet is changing the way of living and style ofpeople.Amongthesetweetscanbecategorizedbythehashvaluetagsforwhichtheyarecommentingandpostingtheirtweets. So,nowmanycompaniesandalsothesurveycompaniesareusingthisfordoingsomeanalyticssuchthatthey can predict the success rate of theirproduct.

Tocalculatetheirviewsisverydifficultinanormalwaybytakingtheseheavydatathataregoingtogenerate day by day. If we consider getting the data from Twitter one should use any one programming languagetocrawl the data from their database or from their web pages. Coming to this problem here we are collecting thisdatabyusingApacheEcosystem'sDataingestionToolknownasFlumeandthenprocessingsamedatausingSpark.

**Levels of sentimentanalysis**

**Word level**: Determines whether an expression is neutral or polar and then disambiguated the polarity ofthepolarexpressions.

**Sentence level:** Categorization attempts to classify positive and negative sentiments for each or whether asentenceis subjective orobjective.

**Document level:** There are two kinds ofapproaches

1. Term -counting approachesand
2. Machine learningapproaches.

Term-countingapproachesusuallyinvolvederivingasentimentmeasurebycalculatingthetotalnumberofnegativeand positiveterms.

Machine learning approaches recast the sentiment classification problem as a statistical classificationtask.

 

## II. PROBLEMDEFINITION

### A. EXISTINGSYSTEM

As per the older ways of crawling data and performing the sentiment analysis on those data was done with thehelpofJavaandRDBMS.Peoplewereusingsomecodingtechniquesforcrawlingthedatafromthetwitterwheretheycould extract the data from the Twitter web pages by using some code that might be written either in JAVA,Pythonetc.Forthosetheywereusinglibrariesthatwereprovidedbythetwitterorganization.Byusingthistheywerecra wling the data that they wantedparticularly.

After getting it the raw data was filtered by using some old techniques and also they foundoutthepositive,negativeandmoderatewordsfromthelistofcollectedwordsinatextfile.Allthesewordswouldbecollect edtofilteroutordosomesentimentanalysisonthefiltereddata.Thesewordscanbecalledasadictionarysetbywhichtheyperfor medsentimentanalysis.Also,afterperformingallthesethingstheywantedtostoretheseresultsinadatabaseandcomingtohere theyusedRDBMSwheretheywerehavinglimitationsincreatingtablesand also accessing the tableseffectively.

### B. APPROACHES

**Lexicon** based approach can be used with an assumption work that the collective polarity of a document orsentenceis the sum of polarities of the individual words orphrases.

**Keyword spotting** approach can be used to classify the text by affect categories based on the presenceofunambiguous affect words such as sad, bored, happy,afraid.

**Bagging**, In this approach, relationships between the individual words are not considered and a documentisrepresented as a mere collection of words. To determine the overall sentiment, sentiments of every wordisdetermined and those values are combined with some aggregationfunctions.

**Statisticalclassification**approachisusedonelementsfrommachinelearningsuchassupportvectormachine,latent semantic analysis, ”bags of words” and Semantic Orientation-point wise mutualinformation.

**Lexicalaffinity**notonlydetectsobviousaffectwords,Italsoassignsarbitrarywordsaprobable“affinity”toparticularemotio ns.

 

 