International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

Real Time Data Analysis using Apache Spark

Prof. Balaji Bodkhe[1], Priyanka Shinde[2], Shruti Jagtap[3], Shraddha Kamble[4], Saroj Sonawane[5]

[1] Assistance Professor, Modern Education Society's College of Engineering, Pune, Maharastra, India

[2-5] UG Students, Modern Education Society's College of Engineering, Pune, Maharastra, India

ABSTRACT: Social media websites have emerged as one of the platforms to raise users' opinions and influence the way any business is commercialized. Opinion of people matters a lot to analyse how the propagation of information impacts the lives in a large-scale network like Twitter. Sentiment analysis of the tweets determines the polarity and inclination of vast population towards specific topic, item or entity. Industries are using Hadoop extensively to analyse their data sets. The reason is that Hadoop[1] framework is based on a simple programming model (MapReduce) and it enables a computing solution that is scalable, flexible, fault-tolerant and cost effective. Here, the main concern is to maintain speed in processing large datasets in terms of waiting time between queries and waiting time to run the program. Spark is a framework for performing general data analytics on distributed computing cluster like Hadoop. It provides in memory computations for increase speed and data process over mapreduce. It runs on top of existing hadoop cluster and access hadoop data store (HDFS), can also process structured data in Hive and Streaming data from HDFS, Flume, Twitter.In this paper, we will see the brief descriptions of sentiment analysis using Spark, its features and working with Spark using Hadoop.

KEYWORDS: Apache Spark, Sentiment Analysis, Twitter, Opinion mining, Natural Language Processing

I.INTRODUCTION

Spark was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process. As against a common belief, Spark is not a modified version of Hadoop and is not, really, dependent on Hadoop because it has its own cluster management. Hadoop is just one of the ways to implement Spark. Spark uses Hadoop in two ways – one is storage and second is processing. Since Spark has its own cluster management computation, it uses Hadoop for storage purpose only. The main feature of Spark is its inmemory cluster computing that increases the processing speed of an application. Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

Every day massive amount of data is generated by social media users which can be used to analyze their opinion about any event, movie, product or politics. Conventional tools like Apache Storm analyze stream in micro-batch whereas novel tools like Apache Spark process data in real time making analyzing and processing of real time data possible.

1. In large scale data processing application it is very importantly Mapreduse can be achieved far better in apache spark
2. Spark can run independently. Apart from that it can run on hadoop 2's YARN cluster manager, and can read any existing hadoop data.
3. If you see Apache spark ecosystem, It provide rich set of tools to work with real time data. Browse for Spark ecosystem and its capabilities.
4. Importantly Cluster management is a tedious task in the industries , Apache mesos does a great job minimal effort in handling cluster.

As whole Apache spark has a good future and many industries has already embraced this open source technology

II.LITERATURE SURVEY

A tweet-level semi-supervised spam detection (S3D) system. The proposed architecture consists of two main modules: the spam detection module that works in real-time mode and the batch mode configuration update module. The spam detection module consists of four lightweight detectors: 1) blacklisted domain detector for labelling tweets containing blacklisted URLs; 2) near-duplicate labeling tweets that are almost duplicates of confidently pre-labeled tweets; 3) accurate labeling ham detector tweets that are posted by trustworthy users and do not include spammy words; and 4) multi-classifier-based labeling detector. The details that the detection needs. The detectors are built on the basis of our observations from a set of 14 million tweets, and the detectors are computationally efficient and suitable for detection

in real time. More specifically, our detectors use two-level, tweet and cluster-level classification techniques. A cluster here is a group of similarly characterized tweets[1].
Twitter is one of the largest networking site for microblogging, it has more than half a billion tweets shared by millions of users on Twitter every day on average. Twitter is easily intruded with malicious activities by such versatility and widespread use. Malicious activities include intrusion of malware, distribution of spam, social attacks, etc. Spammers use the attack strategy of social engineering to send spam tweets, spam URLs, etc. This has made twitter a perfect place for anomalous spam accounts to proliferate. The impact encourages researchers to develop a model that analyzes, detects and recovers from twitter's defamatory actions. Twitter network is inundated with tens of millions of fake spam profiles which may jeopardize the normal user's security and privacy. The word social network online has originated from numerous interdisciplinary fields. Social fields, psychology, sociology, statistics, and graph theory represent a structure of social networks consisting of a set of individuals or organizations with different interactions or relationships between them. Online social networks are considered to be the most sought after social tool used by the world's masses for sharing common interest and interacting with each other[3].

A technique that uses Twitter data sentiment analysis and tests post feelings in or-der to determine the sources of malicious content. This was done by also taking into account the public influence of the posts. Our work's prominent feature is the methodology used for feature-oriented study of feelings. This involves an an algorithm parsing a done tweet and building each sentence's Dependence tree to effectively identify the tweet's feeling. Sentiment Analysis is the method aimed at defining data's emotion or subjective significance in terms of their thinking. It is an application of Natural Language Processing, Computational Linguistics and Text Analytics fields that have been widely explored[4].
An effective approach to the exploration of information from imbalance databases designed specifically for opinion mining. The suggested solution to Under Sampled Imbalance Data Learning (USIDL) uses the special methodology for sampling majority subset instances. The experimental results show that on seven performance metrics, the new solution performs better than the current C4.5 algorithm. The opinion mining data can be found on the websites of social networking, where users express their opinions on a product or topic. The challenging task of opinion mining is about the data available with different formats or types for sentiment analysis. The suggested solution to Under Sampled Imbalance Data Learning (USIDL) uses the special methodology for sampling majority subset instances.[5]

Structures an improved CNN and LSTM model for the Arabic text data resourcefulness feature on freely available benchmark datasets, with word2with representation model for each corpus. The model is predicted in highlight for the study of Arabic opinion (ASA). Compared to previous research, the new architecture achieved better results on three out of five datasets. Twitter for Ar-Twitter dataset formation with manual sentiment class labeling. The corpus includes for each 1000 tweets for a fair representation of positive and negative classes[6].

Real-time Information Visualization and Analysis System–RIVA to collect social network data, such as Facebook, through the use of the Spark cloud computing framework to analyze popular topics around the world. We also carry out the sentiment analysis to get the feeling of people for each problem and to display the distributions of positive and negative feelings. In addition, we are gathering web news titles to test whether they are compatible with the findings of data analysis on the social network. Our experimental results show that RIVA is able to process data in real time, acquire and automatically visualize information from heterogeneous sources[7].

Big data mining for limited-resource researchers. Our genuine theory was tested experimentally and is outlined here. We used the available Brexit data and the sentiment analysis of the subsequent tweets and the connection with stock exchange data as a case study for our process. We selected the Apache Spark1 as an industry standard for fast handling of big data on large computer clusters to accomplish this. We deployed it as a pre-processing component of the bigdata analytical stack in a single-node cluster modeon as a standard computer. Because Apache Spark attempts to conduct as many fast main memory (RAM) operations as possible, in some cases we have had to deal with limited resources[8].

Analysis of feelings approach. This work proposes a text analysis system for the use of Apache spark for twitter data and is therefore more robust, fast and scalable. In the proposed method, Naïve Bayes and Decision trees machine learning algorithms are used to test sentiment. A tweet is a text-based post with only 140 characters, about the length of a standard newspaper headline and subheading.Machine learning based approach (ML) uses many machine learning algorithms (supervised or unmonitored algorithms) to classify data. Lexicon-based approach uses a dictionary that contains positive and negative words to determine polarity of sentiment. Hybrid-based approach uses a classification approach mixing ML and lexicon-based approach[9].

Using the Apache Spark system, an open source distributed data processing application that uses distributed memory abstraction, sentiment analysis is considered. The aim of using the Machine Learning Library (MLIB) of Apache

Spark is to work effectively with an enormous amount of data. We suggest preprocessing and mastering the machine. Sentiment Twitter data analysis approach relies on the Apache Spark framework that uses supervised learning techniques to identify tweets.[10]

III.RESEARCH METHODOLOGY

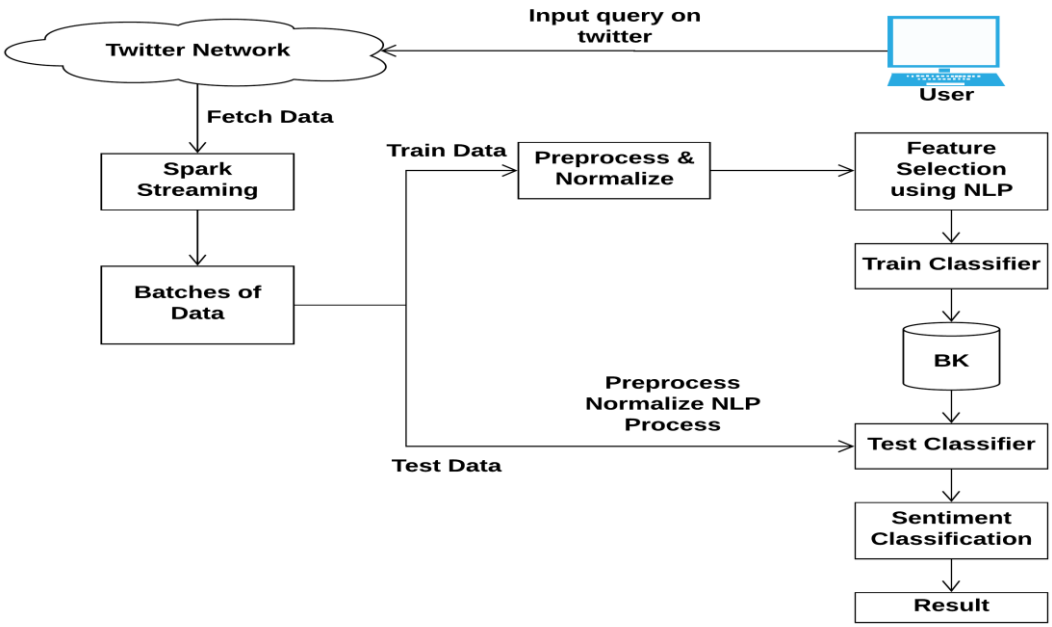

Figure 1 : Proposed system Design

Pre-processing: Then we will apply various pre-processing steps such as lexical analysis, stop word removal, stemming (Porters algorithm), index term selection and data cleaning in order to make our dataset proper.

Lexical analysis: Lexical analysis separates the input alphabet into,Word characters (e.g. the letters a-z) andWord separators (e.g space, newline, tab).

Stop word removal: Stop word removal refers to the removal of words that occur most frequently in documents. The stop words includes,
1) Articles (a, an, the,....)
2) Prepositions (in, on, of,...)
3) Conjunctions (and, or, but, if,....)
4) Pronouns (I, you, them, it....)
5) Possibly some verbs, nouns, adverbs, adjectives (make ,thing, similar....)

Stemming: Stemming replaces all the variants of a word with a single stem word. Variants include plurals, gerund forms (ing forms), third person suffixes, past tense suffixes, etc.).Example: connect: connects, connected, connecting, connection and so on. Here we will use Porter's algorithm for stemming.

Feature Extraction

The appropriate set of features from the given document canbe extracted such that it can improve the overall performance. In feature extraction, based on some counter measure the feature can be extracted.

Training of the classifier:

After choosing proposed text classification algorithms Naive Bayes and feed the training corpus to the classifier to get a training model.

Classification:

After we get the training model, we can feed the testing data into it and get the prediction of classification. The testing stage includes pre-processing of testing text, vectorization and classification of the testing text.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

Algorithm Design
1 : Stop word Removal Approach
Input: Stop words list L[], String Data D for remove the stop words.
Output: Verified data D with removal all stop words.
Step 1: Initialize the data string S[].
Step 2: initialize a=0,k=0
Step 3: for each(read a to L)
If(a.equals(L[i]))
Then Remove S[k]
End for
Step 4: add S to D.
Step 5: End Procedure

2 Stemming Algorithm.
Input : Word w
Output : w with removing past participles as well.
Step 1: Initialize w
Step 2: Intialize all steps of Porter stemmer
Step 3: for each (Char ch from w)
If(ch.count==w.length()) && (ch.equals(e))
Remove chfrom(w)
Step 4:if(ch.endswith(ed))
Remove 'ed' from(w)
Step 5: k=w.length()
If(k (char) to k-3 .equals(tion))
Replace w with te.
Step 6: end procedure

3 TF-IDF
Input : Each word from vector as Term T, All vectors V[i…n]
Output : TF-IDF weight for each T
Step 1 : Vector = {c1, c2, c3….cn}
Step 2 : Aspects available in each comment
Step 3 : D = {cmt1, cmt2, cmt3, cmtn}
and comments available in each document
Calculate the Tf score as
Step 4 :tf (t,d) = (t,d)
t=specific term
d= specific document in a term is to be found.
Step 5 :idf = t → sum(d)
Step 6: Return tf *idf

IV.RESULTS AND DISCUSSIONS

The proposed system performance evaluation, we calculate matrices for accuracy. We implement the system on java architecture framework with INTEL 3.0 GHz i5 processor and 8 GB RAM. Some user comments are positive or negative, and the data contains around 80,000 user's comments. The system finally classifies all the comments as positive, negative as well as neutral. Negation handling also works at the time of aspect classification. Here table 1 shows the estimated system performance with different existing systems. So, proposed results are around on satisfactory level.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

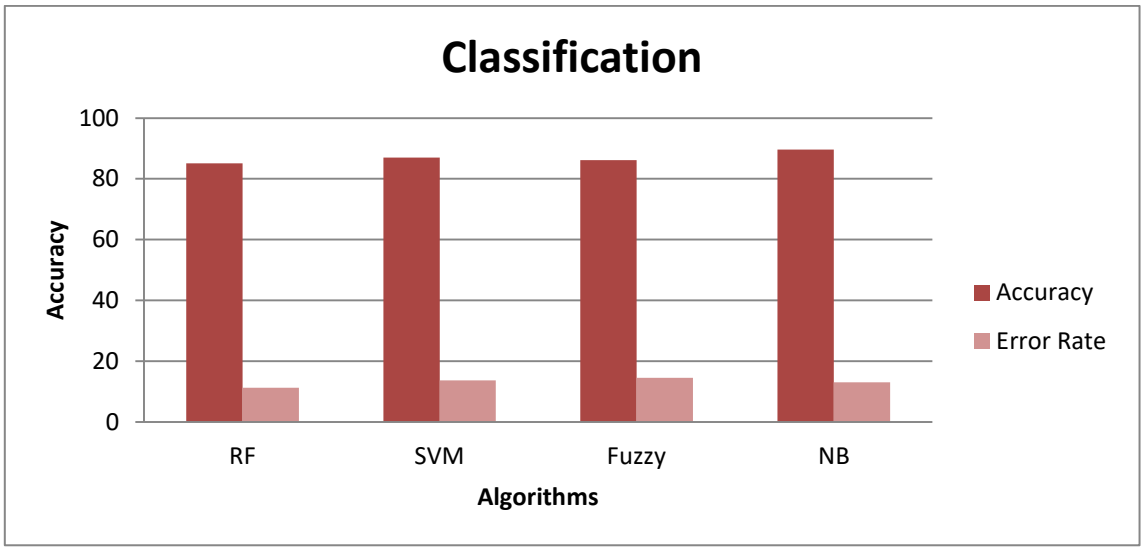

Figure 2: Average accuracy with 5 Fold cross validation using Naïve Bayes algorithm on twitter dataset

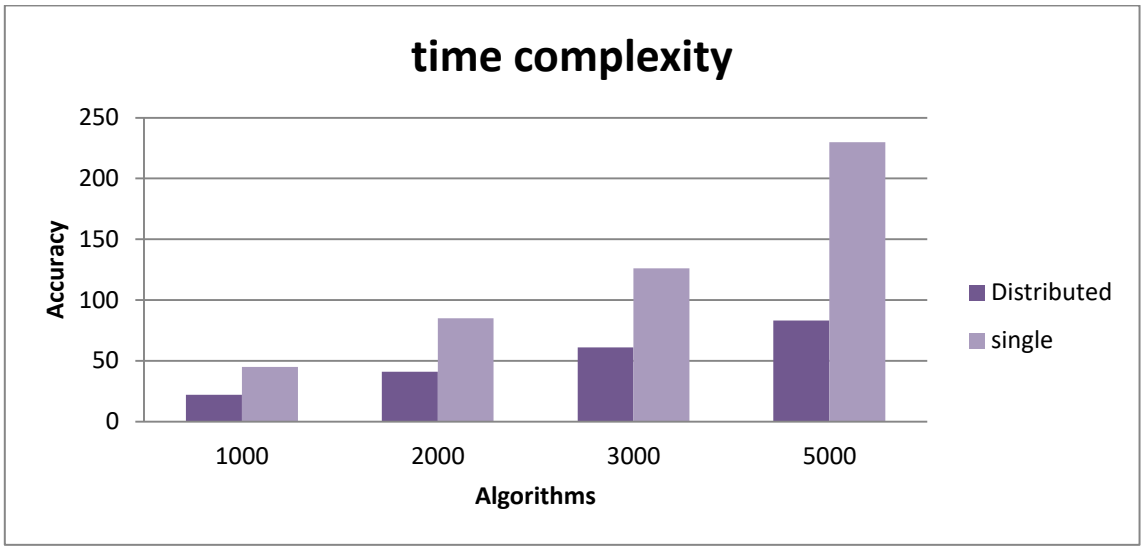

Figure 3: Processing time with apache spark with number of instances using single system and distributed system

V.CONCLUSION

Apache Spark is a powerful open sourceprocessing engine built around speed, ease of use, andsophisticated analytics. Since its release, ApacheSpark has seen rapid adoption by enterprises across awide range of industries. Internet powerhouses such asYahoo, Baidu, Airbnb, eBay and Tencent, have eagerlydeployed Spark at massive scale. The modular nature of the API (based on passingdistributed collections of objects) makes it easy tofactor work into reusable libraries and test it locally.Conclude that there is an inverse proportional relationship between runtime and number of machines in the Spark Cluster, higher performance capacity will be obtained if additional nodes are added to the cluster. Our system can be defined as efficient and scalable from the earlier tests.

REFERENCES

[1] Sedhai, Surendra, and Aixin Sun. "Semi-supervised spam detection in Twitter stream." IEEE Transactions on Computational Social Systems 5.1 (2017): 169-175.
[2] Savyan, P. V., and S. Mary SairaBhanu. "Behaviour Profiling of Reactions in Facebook Posts for Anomaly Detection."2017 Ninth International Conference on Advanced Computing (ICoAC). IEEE, 2017.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 5, May 2020||

[3] Gheewala, Shivangi, and Rakesh Patel. "Machine Learning Based Twitter Spam Account Detection: A Review." 2018 Second International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2018.
[4] Shilpa, P., and SD Madhu Kumar. "Feature oriented sentiment analysis in social networking sites to track malicious campaigners." 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS).IEEE, 2015.
[5] Adinarayana, Salina, and E. Ilavarasan. "An efficient approach for opinion mining from skewed twitter corpus using under sampling approach."2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET).IEEE, 2017.
[6] Al Omari, Marwan, et al. "Hybrid CNNs-LSTM Deep Analyzer for Arabic Opinion Mining."2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE, 2019.
[7] Wu, Yong-Ting, et al. "RIVA: A Real-Time Information Visualization and analysis platform for social media sentiment trend." 2017 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT).IEEE, 2017.
[8] Andrešić, David, PetrŠaloun, and IoannisAnagnostopoulos. "Efficient big data analysis on a single machine using apache spark and self-organizing map libraries."2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP).IEEE, 2017.
[9] Jain, Anuja P., and Padma Dandannavar. "Application of machine learning techniques to sentiment analysis."2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT).IEEE, 2016.
[10] Elzayady, Hossam, Khaled M. Badran, and Gouda I. Salama. "Sentiment Analysis on Twitter Data using Apache Spark Framework."2018 13th International Conference on Computer Engineering and Systems (ICCES).IEEE, 2018.