International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488|

||Volume 8, Issue 5, May 2020||

Real Time Data Analysis using Apache Spark: An Overview

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

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

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

ABSTRACT: Social media websites have emerged as one of the platforms for raising the opinions of users and influencing the marketing of any business. Due to the nature, variety and quantity of the data, sentiment analysis on Twitter Data is a challenging topic.Implement the Apache Spark method, an open-source Big Data programming platform. Together with the Natural Language Processing techniques, the sentiment analysis tool is based on Machine Learning methodologies and uses the Machine Learning library of Apache Spark, MLlib. We implement several pre-processing steps to achieve better outcomes in Sentiment Analysis to tackle the complexity of Big Data.We carry out corpus-based Sentimental Analysis of social networking data and check the total time taken for data processing by both the systems and their sub-components. Results for the implementation of meta-action generation methods show faster computational time for Spark system compared to Hadoop Map Reduce. Results identify the view of the user into positive and negative through tweets. Second, we discuss in detail various techniques for carrying out an examination of feelings on twitter info.

I.INTRODUCTION

As the Web expands, the horizons are expanding. Social media and micro blogging sites such as Facebook, Twitter,Tumbler dominate fast-paced dissemination of encapsulated news and trending topics around the globe. A subject becomes a phenomenon when more and more people share their views and observations, making it a reliable source of understanding online. Twitter is an online tweet-driven networking site that is limited to 140 character tweets. The character limit therefore enforces the use of hashtags to classify text. Roughly 6,500 tweets are currently being published per second, resulting in around 561.6 million tweets a day. Such tweet streams are typically noisy representing various subjects, shifting details about attitudes in unfiltered and unstructured format. Analysis of feelings on Twitter means using natural language processing to isolate, classify and describe the meaning of feelings. Sentiment Analysis is often conducted at two stages (1) coarse, and (2) fine. At the gross level, the analysis of whole documents is carried out while at the fine level, the analysis of attributes is carried out. In terms of tonality, polarity, lexicon and tweet grammar, there are many challenges involved. We tend to be highly unstructured and unstructured. This project uses the rapid processing capabilities of Apache Spark to analyze feelings from such high-speed tweets in real-time.A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These topics generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents.

II.LITERATURE REVIEW

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 labeling 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].

An unattended clustering tool that analyzes user reactions called smileys. The reactions are profiled and are further categorized through the application of similarity measures and unsupervised clustering techniques. This approach shows the actions of users ' immediate emotional reactions to the various Facebook posts. The study of these reactions provides important information to find anomalous activity in Facebook accounts because reactions are instant. OSNs have become an integral part of modern society and are used in all spheres of life. Facebook is now the most commonly used OSN, and every year the number of active Facebook users is growing[2]

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]

ID	Title of paper	Base classification technique	Other techniques tested	Dataset	Performance parameter
1.	Using Machine Learning to Detect Fake Identities : Bots vs. Humans [11]	Random Forest	SVM Linear, Adaboost	Twitter dataset	F1 score, PR-AUC, Accuracy
2.	Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram [12]	Multi-Layer Perceptron	Logistic Regression, Random Forest,AdaBoost, XGBoost	Instagram dataset	Precision, Recall, AUC
3.	Semi-Supervised Spam Detection in Twitter Stream [13]	S3D (Naïve Bayes, Logistic Regression and Random Forest.)	Naïve Bayes , Logistic Regression, Random Forest	Twitter dataset	F1 score, Precision, Recall
4.	Detecting Clusters of Fake Accounts in Online Social Networks [14]	Random Forest	Logistic Regression, SVM	LinkedIn dataset	Recall, AUC
5.	Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms [15]	SVM	Decision Tree, AdaBoost, KNN, Random Forest, Naïve Bayes	Facebook dataset	Precision , Recall, F1 Score
6.	Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots? [16]	AdaBoost, Gradient boosting	Gaussian naive Bayes, SVM, Random Forest, Extremely Randomized Trees	Twitter (Indian Election Dataset)	AUROC
7.	Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms [17]	K- Nearest Neighbor (KNN),	Decision Tree, Random Forest (RF)	System taken data from UCI machine learning repository	Accuracy, semi structured dataset.
8	Mining Anonymity: Identifying Sensitive Accounts on Twitter [18]	Random Forests &	_	Twitter dataset	Precision, Recall

		Binary classifiers			
9.	On Profiling Bots in Social Media [19]	Logistic Regression	SVM, Naïve Bayes, Random Forests	Twitter dataset (Singapore)	F1 Score, Precision, Recall,
10.	Towards Detecting Anomalous User Behaviour in Online Social Networks [20]	KNN	_	Facebook dataset	AUROC, TP rate, FP rate

III.PROPOSED SYSTEM ARCHITECTURE

A common characteristic of communication on online social networks is that it happens via short messages, often using nonstandard language variations. These characteristics make this type of text a challenging text genre for natural language processing. It would therefore be useful if user profiles can be checked on the basis of text analysis, and false profiles flagged for monitoring. This research work presents an exploratory study in which system apply an age group categorization approach base on the text features.

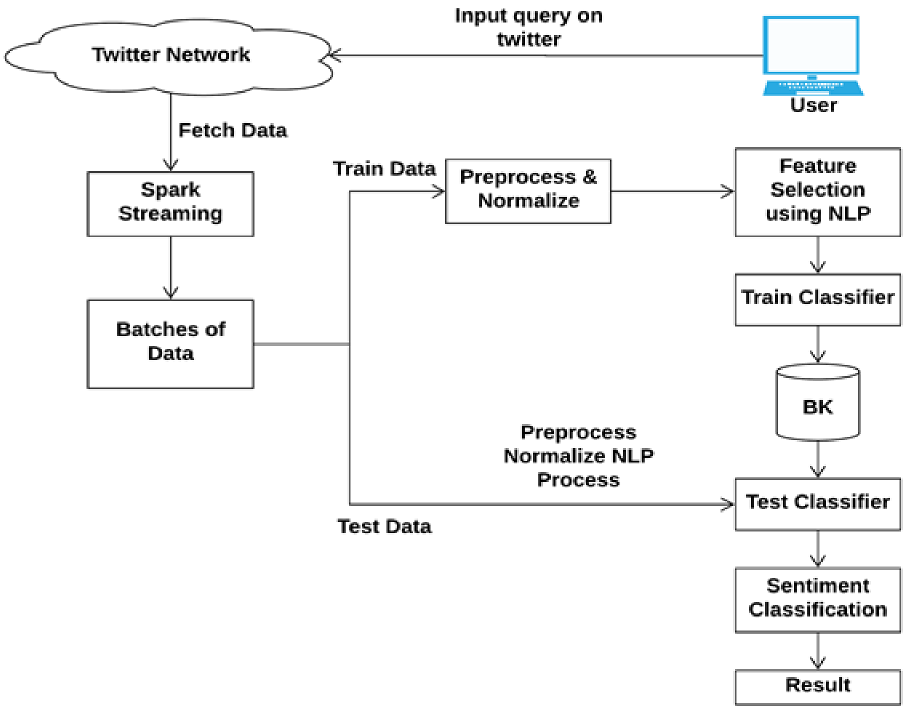

Fig. 1: System Architecture

IV.DISCUSSION

Preprocessing: Then we will apply various preprocessing 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,

1)Word characters (e.g. the letters a-z) and
2)Word 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 preprocessing of testing text, vectorization and classification of the testing text.

V.CONCLUSION

In this work, we propose an efficient technique of prediction of sentiment, using the Machine Learning library of Apache Spark to execute various algorithms of classification.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.

VI.FUTURE SCOPE

Furthermore, accuracy can be increased in future by enhancing features set and testing for other classification techniques such as deep learning with different activation functions. For the near future, we plan to analyze the effect on the input vector by adding other different features and using larger datasets. In addition, our goal is to create an online service that benefits from Spark Streaming, which is considered the Apache Spark library for the management of data streams, providing users with real-time predictions and analysis of the feelings of the subjects they need.

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.
[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.

[11] Estee Van Der Walt and Jan Eloff, “Using Machine Learning to Detect Fake Identities: Bots vs Humans,” IEEE, 2018.

[12] Indira Senet. al. “Worth its Weight in Likes: Towards Detecting Fake Likes on Instagram,” ACM, 2018.

[13] SurendraSedhai and Aixin Sun, “Semi-Supervised Spam Detection in Twitter Stream,” IEEE , 2018.

[14] Cao Xiao, David Freeman and Theodore Hwa , “Detecting Clusters of Fake Accounts in Online Social Networks,” ACM , 2015.

[15] Ikramet. al., “Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms,” ARXIV, 2016.

[16] J. Dickerson, V. Kagan and V. Subhramanian “Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?” IEEE , 2014.

[17] Ikramet. al., “Combating Fraud in Online Social Networks: Detecting Stealthy Facebook Like Farms,” ARXIV, 2016.

[18] S. Peddinti, K. Ross and J. Cappos “Mining Anonymity: Identifying Sensitive Accounts on Twitter,” ARXIV ,2016.

[19] R. Oentaryoet. al. “On Profiling Bots in Social Media,” ARXIV, 2016.

[20] B. Viswanathet. al. “Towards Detecting Anomalous User Behaviour in Online Social Networks,” USENIX, 2014.