JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Scalable Platform to Collect and Analyse Big Data in Twitter using Random forest Algorithm

Mr. N.Sendhil Kumar[1], Mr. S.Karthick[2], Ms.P.Haritha Reddy[3], Ms. V.Pavani[4]

Associate Professor & HOD, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[1]

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[2]

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[3]

MCA Student, Department of MCA, Sri Venkateswara College of Engineering and Technology, Chittoor, India[4]

Abstract: **Scalability is a key feature for big data analysis and machine learning frameworks and for applications that need to analyze very large and real-time data available from data repositories, social media, sensor networks, smart phones, and the Web. Scalable big data analysis today can be achieved by parallel implementations that are able to exploit the computing and storage facilities of high performance computing (HPC) systems and clouds, whereas in the near future Exascale systems will be used to implement extreme-scale data analysis. n. The platform supports the collection of social media data and applies many filters for cleaning and further use for machine learning (ML) which is random forest algorithm based systems. Our focus has been primarily on healthcare-related research, which shows the strength of the presented platform. However, the platform itself is malleable to any topic of interest. Data collected and processed are suitable for further ML analysis. We present our platform using a specific healthcare search topic to emphasize the power of our system for future research endeavors in the healthcare field**.

Index Terms **– Twitter, Machine Learning, Random Forest classifier *etc.***

1. INTRODUCTION

Solving problems in science and engineering was the first motivation for inventing computers. After a long time since then, computer science is still the main area in which innovative solutions and technologies are being developed and applied. Also due to the extraordinary advancement of computer technology, nowadays data are generated as never before. In fact, the amount of structured and unstructured digital datasets is going to increase beyond any estimate. Databases, file systems, data streams, social media and data repositories are increasingly pervasive and decentralized. As the data scale increases, we must address new challenges and attack ever-larger problems. New discoveries will be achieved and more accurate investigations can be carried out due to the increasingly widespread availability of large amounts of data. Scientific sectors that fail to make full use of the huge amounts of digital data available today risk losing out on the significant opportunities that big data can offer. To benefit from the big data availability, specialists and researchers need advanced data analysis tools and applications running on scalable architectures allowing for the extraction of useful knowledge from such huge data sources. High performance computing (HPC) systems and cloud computing systems today are capable platforms for addressing both the computational and data storage needs of big data mining and parallel knowledge discovery applications. These computing architectures are needed to run data analysis because complex data mining tasks involve data- and compute-intensive algorithms that require large, reliable and effective storage facilities together with high performance processors to get results in appropriate times. Now that data sources became very big and pervasive, reliable and effective programming tools and applications for data analysis are needed to extract value and find useful insights in them. New ways to correctly and proficiently compose different distributed models and paradigms are required and interaction between hardware resources and programming levels must be addressed. Users, professionals and scientists working in the area of big data need advanced data analysis programming models and tools coupled with scalable architectures to support the extraction of useful information from such massive repositories. Initially, a team can create a Twitter stream by specifying their search keywords and the target number of tweets. An existing Twitter-specific challenge is the limit on the number of tweets that can be collected concurrently. Therefore, we offer study teams the convenience of entering our queue, where we schedule limited resources to a server with as many data collection requests as possible. Upon completion of data collection, a study team can identify and apply the appropriate preprocessing options, ranging from removal operations (e.g., of repeated tweets) to transformations (e.g., of abbreviations, acronyms, and emoticons into fully formed words). Furthermore, as options need to be applied in a specific order, our platform provides a comprehensive description of options to support study teams in understanding the choices and automatically handles the ordering between the selected options. Finally, our platform provides an innovative technical solution to the problem of annotating large sets of tweets. Specifically, we combine social network data collection with crowdsourcing by using Amazon Mechanical Turk to label Twitter data. If a study team seeks to annotate a large set of tweets, our platform connects with Amazon Mechanical Turk to create a study that recruits participants and provides them with an online survey to perform the annotation. In addition, the

survey is designed automatically and considers redundancy (i.e., how many participants need to annotate each tweet) and survey length (i.e., a number of tweets that each participant must annotate).

This paper is organized in five sections. After this introduction, in Section II, literature survey discussed of the paper, section III about the Existing system, Section IV about Proposed System, as well as the novel feature of the proposed method. Finally, Sections V and VI provide the simulation results and the conclusions and Future work, respectively.

2. LITARATURE SURVEY

In this section, we will outline works related to the collection, processing, and sentiment analysis of social data.

Cambria et al. [1] offered a contemporary overview of computational approaches of big social data analysis and identified important real-world applications in health research and education.

Alotaibi et al. [2] presented a systematic review applying big data concepts to the supply chains specifically for healthcare. They reviewed important concepts, including big data (analytics), specifically healthcare big data. Furthermore, they also examined supply chain management in relation to healthcare.

Wasilewski et al. [3] used Twitter to recruit participants for health research. By tracking the use of Twitter users online, the authors created a set of parameters that can be used to determine whether Twitter users would be suitable for certain healthcare research projects based on their tweet activity.

Sequeira et al. [4] performed a large-scale study on the data of 0.42 million Twitter users to investigate and characterize the spread of prescription drug abuse (DA) information through online social networks. They collected Twitter data around a set of keywords containing generic and brand names of common drugs of abuse and used various machine learning (ML) and deep learning (DL) techniques for classification of the data into DA or nonabuse (NA) classes. This allowed for further investigation of the cascades of tweets promoting DA throughout the platform.

Adrover et al. [5] worked on the sentiment analysis of people with HIV on Twitter trying to detect them and see if their drug treatments created the positive sentiment on social media. A data set of around 50 million tweets was used, where they used both computational and manual approaches, including keyword filtering, crowdsourcing, computational algorithms, and ML to filter noise.

Angiani et al. [6] provided an in-depth analysis of commonly practiced pre-processing techniques used in sentiment analysis of microblogging data and emphasized the importance of these techniques in improving system accuracy. Highlighted are the relevant cleaning techniques for Twitter data that take into consideration use of URLs, mentions, hashtags, emoticons, and other colloquially used Internet typography. This is important and concise work on employing preprocessing for sentiment analysis.

Ginn et al. [7] presented a collection of 11 000 tweets that deal with drug reactions. These tweets were mined from the Twitter application programming interface (API) and manually annotated by experts with medical and biological science backgrounds. This study clearly shows the need for a more user-friendly approach to Twitter mining as presented in this article. Not only was the data set extracted from Twitter's API small, but also most of the analysis of those tweets was done manually.

Emadzadeh et al. [8] provided a classification method to deal with adverse drug reactions (ADRs) for healthcare data using advanced NLP techniques. Three data sets were developed to identify ADRs from data on the Internet posted by users.

Gokulakrishnan et al. [9] compared the accuracy of the most common classification models and algorithms for sentiment analysis on Twitter data streams. It is found that Discriminative Multinomial Naïve Bayes (DMNBText) classifiers preform most accurately when compared with other classifiers, notably sequential minimal optimization (SMO), support vector machines (SVMs), and basic Naïve Bayes algorithms.

However, as mentioned by Hutto and Gilbert [10], there are drawbacks to using classifiers such as extensive training, which can be time consuming and computationally expensive.

Behera and Eluri [11] gave a methodology for sentiment analysis for disease spread monitoring on Twitter that used location as well as time. The goal was to measure the degree of concern in tweets regarding three specific diseases, malaria, swine flu, and cancer. The tweets were taken through a two-step sentiment classification process to identify negative personal tweets.

Signorini et al. [12] made headway in the study of the H1N1 flu, where they specifically looked at Twitter data and its ability to track the public sentiment of the actual disease. The study uses the keywords to filter the Twitter API and obtain a time-stamped and geolocated data set using the tweet author's ability to enable these features on their specific Twitter account.

Coletta et al. [13] studied the public sentiment classification of tweets using a combination of SVM and cluster ensemble techniques. This algorithm, C3E-SL, is capable of combining classifiers with cluster ensembles to refine tweet classifications from additional information provided by the clusters.

Ji et al. [14] used Twitter to track the spread of public concern regarding epidemics. Their methods included separating tweets into personal and news (nonpersonal) categories to focus on the public concern. The personal tweets were further classified into personal negative and personal nonnegative, depending on the sentiment detected.

Priya et al. [15] proposed a method of Twitter information retrieval to assess infrastructure damage during emergency situations. The method implements spit query with topic aligned query expansion (TAQE) to identify relevant tweets in a stream that are used to estimate the infrastructure damage severity for locations mentioned in the tweets. A higher damage score for a given location is characterized by a higher number of related tweets containing negative sentiment. With the TAQE technique, they offered an accurate approach to damage assessment through the analysis of Twitter data.

Wang et al. [16] presented a system for real-time sentiment analysis on Twitter for the 2012 U.S. Presidential Election. The system is based on a Naïve Bayes classifier that is trained via a baseline sentiment model of annotated tweets collected through Amazon Mechanical Turks. The trained classifier can then determine the sentiment of the tweets related to the election in real time

3. EXISTING SYSTEM

In existing system, sentiment analysis and textual analytics is presented, as well as that on ML methods, twitter and NLP. Significant data challenges are evolving and need to be addressed, and strategic information characteristics restructuring data, as well as the ML techniques. The analysis of past epidemics, crisis situational analysis and tracking, has also involved tracking Twitter data. A better understanding of the US's geographical spread concerning the valances of both healthy and unhealthy sentiment. The people in rural areas tweet less than those in urban areas and suburbs, using the spatial distribution of analyzed tweets. This work also notes that food tweets per capita were less in small urban areas than in larger towns and cities. It was revealed using linear regression that in low-income areas, tweets related more to unhealthy sentiment analytics has also had avenues for the use of Twitter data.

4. PROPOSED SYSTEM

The proposed model is introduced to overcome all the disadvantages that arise in the existing system. This system will increase the accuracy of the classification results by classifying the data based on the tweets. As raw tweets are often short, unstructured, informal, and noisy, the first step of sentiment analysis is to pre-process the data. The Sentiment timeline helps to understand trends in positive and negative sentiment over time. Multinomial Naïve Bayes algorithm predicts the tag of a text such as a piece of email or article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. The multinomial distribution normally requires integer feature counts. In the Random Forest algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Support Vectors are simply the co-ordinates of individual observation. The Random Forest classifier is a frontier which best segregates the two classes (hyper-plane/ line). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable where the two values are labeled "0" and "1".

A. System Architecture

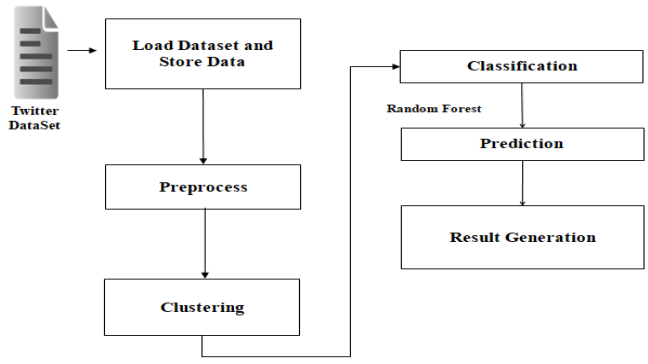


Figure 1: System Architecure

As shown in Fig.1 We are proposing the scalable data to collect from twitter using Random Forest algorithm which mainly implies the detection of abnormal packets using past experience of the system. Here the incoming packets are analysed and categorised according to values of the attributes to produce dataset. Using this data set the next arriving packets are detected as normal or abnormal packets. If abnormal packets are detected reporting can be done. Scalability is the ability of a system, network, or process to manage the growing workload or to expand its potential to accommodate this growth. For example, when adding resources to a system, a system can be considered as scalable when it increasingly produces overall output. Our system is horizontally scalable; by adding more servers to the cluster, we are able to increase the output of our system exponentially.

B. Implementation

Modules:

- Select and Load Dataset
- Data Pre-processing
- Feature Selection
- Classification
- Prediction
- Result Generation

C. Modules Description:

Select and Load Data Set:

- The input data was collected from dataset repository.
- In our process, the Twitter COVID_NLP dataset is used.
- Data selection is the process of selecting the appropriate data set for processing.
- Each of the record consists of 7 features and one marked as attack.
- TheTwitter COVID_NLP Dataset is used for detecting the Sentiment.
- All the data's are selected and loaded into the database for detecting the sentiment analysing.

Data Pre-processing:

- Data pre-processing is the process of removing the unwanted data from the dataset.
- Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.
- This step also includes cleaning the dataset by removing irrelevant or corrupted data that can affect the accuracy of the dataset, which makes it more efficient.
- Missing data removal
- Encoding Categorical data
- Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.
- Missing and duplicate values were removed and data was cleaned of any abnormalities.
- Encoding Categorical data: That categorical data is defined as variables with a finite set of label values.
- That most machine learning algorithms require numerical input and output variables.

Feature Selection:

- Here our dataset will be Clustered into two categories like,
  (i)Positive(ii).Negative (iii)Extremely Positive (iv)Extremely negative (v)Neutral
- Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs.
- The meaningful data's are selected from the extracted features of Twitter COVID_NLPDataset

Classification:

- Classification is a data mining function that assigns items in a collection to target categories or classes.
- A classification model could be used to identify the normal and attacks from the Twitter COVID_NLP Dataset.
- The goal of classification is to accurately predict the target class for each case in the data.

Prediction:

- The goal of classification is to accurately predict the target class for each case in the data.
- The purpose of this module is to predict the attack from the Twitter COVID_NLPDataset.
- Prediction in data mining is to identify data points purely on the description of another related data value. It is not necessarily related to future events but the used variables are unknown. It identify the closely related value. The attacks are predicted from the dataset. It increase the accuracy of the prediction result.

Result Generation

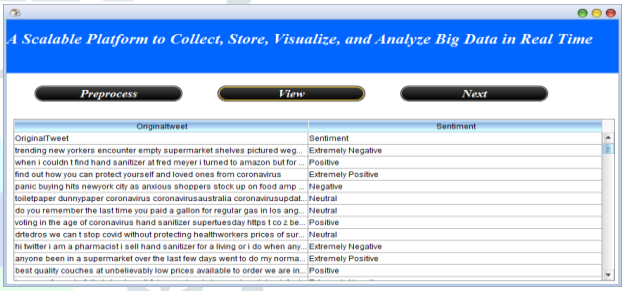
- The Final Result will get generated based on the overall classification and prediction.
- The performance of this proposed approach is evaluated using some measures like,
- The overall classification report is generated based on the normal and attack that is presented in the Twitter COVID_NLPDataset.
- The report which contain dynamically distributed reviews about the products.
- The level that describe the attacks and normal data.

5. SIMULATION RESULTS

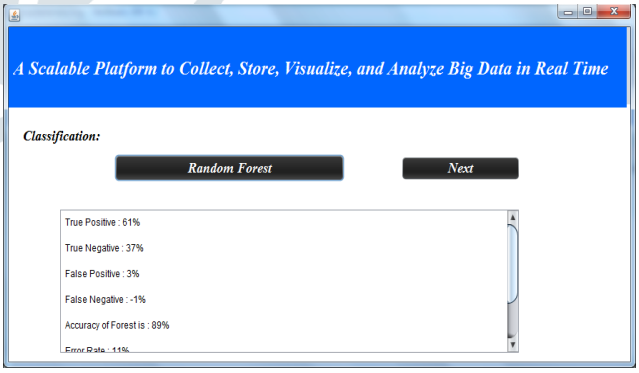
Data selection

Saving the Data

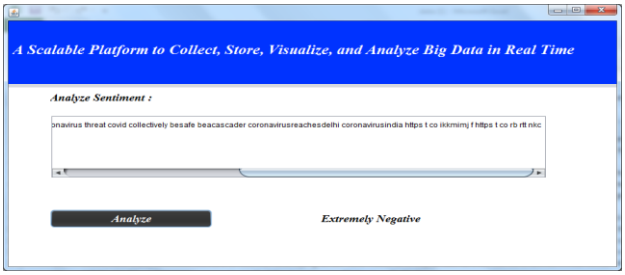
Pre-processing:



Classification:



Result Generation:



6. CONCLUSION AND FUTURE SCOPE

In this article, Twitter has become a standard place to begin research projects where there is a need for large amounts of user created or user-driven data. However, due to the value of big data in today's economic world, Twitter's complimentary API only releases small portions of the entire data set that may be available to a paying customer of the API using Machine learning Algorithms such as Random Forest. In the proposed model, the accuracy of the detection capabilities, and demonstrates the efficiency of parallel computation through the evaluation and comparison of model training implementations

Future Scope

In future we can moving forward, the platform could benefit from additional functionality to allow for dynamic filter creation by the user to further expand the list of available filters to the researchers. This feature would hold great value as there is a constant need to design new filters; for instance, a researcher may only be interested in analyzing the use of emojis or emoticons to understand the state of mental health of the general public. Having an ability to add user-defined filters would further strengthen the platform. In addition, providing a webhook API could enable seamless passing of the collected and cleaned tweets directly to their endpoint to be used in further analysis. These types of endeavors and upgrades to our current platform are left as future work. Through word of mouth, our platform has increased user ship, and our plan is to continue this trend into the coming years.

REFERENCES

[1] E. Cambria, N. Howard, Y. Xia, and T.-S. Chua, "Computational intelligence for big social data analysis," IEEE Comput. Intell. Mag., vol. 11, no. 3, pp. 8–9, Aug. 2016.

[2] S. Alotaibi, R. Mehmood, and I. Katib, "The role of big data and Twitter data analytics in healthcare supply chain," in Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies. Berlin, Germany: Springer, 2019, pp. 267–279.

[3] M. B. Wasilewski, J. N. Stinson, F. Webster, and J. I. Cameron, "Using Twitter to recruit participants for health research: An example from a caregiving study," Health Inform. J., vol. 25, no. 4, pp. 1485–1497, Dec. 2019.

[4] R. Sequeira, A. Gayen, N. Ganguly, S. K. Dandapat, and J. Chandra, "A large-scale study of the Twitter follower network to characterize the spread of prescription drug abuse tweets," IEEE Trans. Comput. Social Syst., vol. 6, no. 6, pp. 1232–1244, Dec. 2019.

[5] C. Adrover, T. Bodnar, Z. Huang, A. Telenti, and M. Salathé, "Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter," JMIR Public Health Surveill., vol. 1, no. 2, p. e7, Jul. 2015.

[6] G. Angiani et al., "A comparison between preprocessing techniques for sentiment analysis in Twitter," in Proc. KDWeb, 2016, pp. 1–15.

[7] R. Ginn et al., "Mining Twitter for adverse drug reaction mentions: A corpus and classification benchmark," in Proc. 4th Workshop Building Evaluating Resour. Health Biomed. Text Process., 2014, pp. 1–8.

[8] E. Emadzadeh, A. Sarker, A. Nikfarjam, and G. Gonzalez, "Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology," in Proc. AMIA Annu. Symp. Bethesda, MD, USA: American Medical Informatics Association, 2017, p. 679.

[9] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, "Opinion mining and sentiment analysis on a Twitter data stream," in Proc. Int. Conf. Adv. ICT Emerg. Regions (ICTer), Dec. 2012, pp. 182–188.

[10] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Proc. 8th Int. AAAI Conf. Weblogs Social Media, 2014, pp. 1–12.

[11] P. Naresh Behera, S. Eluri, and J. University Kakinada, "Analysis of public health concerns using two-step sentiment classification," Int. J. Eng. Res., vol. V4, no. 09, pp. 606–610, Sep. 2015.

[12] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. During the influenza a H1N1 pandemic," PLoS ONE, vol. 6, no. 5, May 2011, Art. no. e19467.

[13] L. F. S. Coletta, N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Combining classification and clustering for tweet sentiment analysis," in Proc. Brazilian Conf. Intell. Syst., Oct. 2014, pp. 210–215.

[14] X. Ji, S. A. Chun, Z. Wei, and J. Geller, "Twitter sentiment classification for measuring public health concerns," Social Netw. Anal. Mining, vol. 5, no. 1, p. 13, Dec. 2015.

[15] S. Priya, M. Bhanu, S. K. Dandapat, K. Ghosh, and J. Chandra, "TAQE: Tweet retrieval-based infrastructure damage assessment during disasters," IEEE Trans. Comput. Social Syst., vol. 7, no. 2, pp. 389–403, Apr. 2020.

[16] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time Twitter sentiment analysis of 2012 US presidential election cycle," in Proc. ACL Syst. Demonstrations. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 115–120.