Work with Streaming Data using Twitter API to Build a Job Portal

Jayakumar Sadhasivam, Senthil Jayavel, Arpit Rathore, Akshay Pratap Singh, Avi Singh, J Cynthia

Abstract—In this digital era that we are living in everyone is obsessed with mobiles, computers and is on the internet, so job portals play an important role in aiding job seekers in their job hunt. It will make them aware of the various job openings, and they will not miss an opportunity. The job is essential for the students who do their graduation, post-graduation, etc. So, we know that there are lots of websites where can we can find the lots of vacancy in various firms but the website is not updated the job vacancy minutely or many of them not updated the job vacancy hourly so we cannot get the minutely updated about the jobs. The vacancy appears quickly on social networking websites like Twitter, LinkedIn, Facebook, etc. as compared to appears on the job websites like indeed.com, monster.com, etc. so many of the people are not on these social media or maybe not regular on the social media, so there is a chance miss the opportunity. Our objective with this paper is to come up with a portal that will provide the user's details about various job openings in respective domains. The portal will stream data from Twitter API to find out the recently published jobs. Classification of relevant and irrelevant tweets is accomplished using the machine-learning algorithm, i.e., Logistic Regression. Using the algorithm, we have measured the 97% accuracy.

Keywords—Twitter, API, Job Portal, Stream Data, Machine-Learning, Logistic Regression

1. INTRODUCTION

In the century, everyone is surrounded by the digital world from waking up from the bed to the sleeping again in it. People around here are now more inclined towards the digital means of information. Even today, big giants to small firms are also using the digital mode for spreading the news for mass communication. To fulfill these necessities, platforms like Twitter, Facebook, LinkedIn etc. providing an excellent way to distribute in the mass and they are also a good example of mass communication. Twitter is a very useful stage for mass communication and many organization is also using this a platform to give a notification regarding the vacancies, new opportunities in an organization. To utilize this information by incorporating with development tools we are implementing a job portal which is using the Twitter API.

This job portal will provide real-time knowledge about any of the tweets which contain any information about any new job vacancies and new opportunities. This will useful for getting fast updates in respect to the getting the search results for the particular posting of the opportunities.

We are using Twitter for the following reasons like:-

1. Twitter is one of the top social media platforms.
2. The number of users connected to this platform as a user and an overall number of active users per month is quite splendid.
3. The diversity of the users ranging from the global leaders to the big names in the job sector to the small-scale industries/ organizations.
4. The job openings related tweets are increasing day by day and moreover quite genuine.

For such reasons, we chose Twitter for our Application Programming Interface (API).

2. LITERATURE SURVEY

Yu and Yang [1] have identified the different perspectives and possibilities that big data drives to us. Conventional marketing analysis and big data analysis have fixed the relationship for different chances. A big data analysis model has been created to analyze the distinguished operation to the users for strengthening the challenges that may be faced by the current product.

This involves a lot of statistics from various associates. As there is a rapid increase in technology, regulation, competition, and inputs of these elements become unfavorable to a specific commodity. The work that has been done in this research in the future time will be presenting a more optimized algorithm with the effects so that the data that have been worked on give the more accurate result whether it contains the large or enormous amount of data.

Zhang and his team members [2] suggested that the main offering of our research is categorized into two folds. First, we will set-up a structure of the big data analytics in which major work is to grasp the distributed computing and streaming to systematic process that is contained in the data streams of big social media. Second, we create the catalog for the suggested architecture that is acquired by a parallel evolution genetic algorithm to flexibly observe disorder reviews with respect to different reviews that are been collected from the various sources. The research that has been taken clearly recommend that the architecture can systematically and easily admit disorganized reviews and comments from a data stream of big social media and that will work better than other non-distributed big data analysis response.

Revised Manuscript Received on August 19, 2019.

Jayakumar Sadhasivam, School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, Tamil Nadu, India.(E-Mail:jayakumarsvit@gmail.com)

Senthil Jayavel, Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India (E-mail: senthil.j.vit@gmail.com)

Arpit Rathore, School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, Tamil Nadu, India

Akshay Pratap Singh, School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, Tamil Nadu

Avi Singh, School of Information Technology and Engineering (SITE), Vellore Institute of Technology, Vellore, Tamil Nadu, India

J Cynthia, Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India

As a conclusion, the effectiveness of the data that has been extracted from the online social media for business intelligence has been upgraded. The crucial work that has been performed is that business intelligence is made available so through which organization administration practice and marketing strategies can be upgraded and it is been created in the more decisive way.

Divya and her team [3] introduced through using the data in the real time they have analysis and integrated both types of data which are been utilized in big data technologies and in the methods of analyzing. As this technology has been applied it can make them help out to determine the patters which are hidden in a way that various organization of the customers can participate and that will help the customers for making their decisions about the purchasing. As a conclusion, the huge and abundant data can be particularly acquired in the different forms whether in the form of tweets or comments.

We have admitted that a combination of data mining into the marketing knowledge structure can be performed. By using the modern information technology the grand and hefty amount of data can be easily processed, analyzed and manage this data to extract that particular information that is essential to the customer.

Castano and his colleagues [4] introduced a current parallel system for sentiment analysis on multilingual tweets that are established on Big Data technologies is introduced. On the one hand, the sentiment classifier executes as well as other state-of-the-art classifiers taking into consideration tweets written in various languages. On the other hand, our system is accomplished of processing millions of tweets in short times taking preference of Big Data processing and parallel architectures, displaying a good scalability in all the considered scenarios.

Twitter can be appraised as a large source of short texts (tweets) comprising of a user point of view. Making sentiment analysis on tweets is difficult from the natural language processing perspective, but also in terms of performance when huge amounts of tweets should be processed.

Sulthana and her team [5] analyzed the tweets that have been done on the twitter are been observed on the basis of time and location. Analyzing the tweets are primarily dependent on the customer description the language that has been used for doing the tweet and the source. Audio file and video file and all other types of files can easily depict the information that has been concentrated in the tweet. Major and crucial work that is conducted is to up-bring the heavy concerning tweets. The accuracy rate that can be classified of the correct tweets can be ranged up to 80%.

Linear regression technique has been adapted here to conclude the correctness. As compared to the support vector machine and Naïve Bayes this approach is much better in this condition or factors. Using then other data analytical techniques linear regression is much adaptable.

Egendy and Elragal [6] researched that the extensive quantity of data that have acquired for the guidance for decision makers. The big data is the data that contain the variety of data and the information that is coming through the online source will be of high velocity, as it is not only that it is big but there are many other things related to it and to handle that data through old techniques and approach may be difficult. There is much need to observe the data that has been provided in order to clinch and inspect so that important data can be easily extracted as there is quick enlargement of the information. On inspecting the genuine case of big data, which has been newly acquired lots of influence and improvement. Today, where the various divisions of the high speed of data are processed every day and there is such hidden information is present which is needed to be examined and then be applied in an excellent manner.

Cuzzocrea and his team [7] have examined that we have done the analytical part over the big data and the most vital thing that can be said through this paper is that analytics is been not done alone on the single dimensional data it has been performed on the multidimensional data. The massive amounts of data have been captured whether in a disorganized manner and for collecting the data there are many sources, as the analytics on the big data depository has been in the trend, the most major task is to extract the important and useful information from the massive and huge depositories. Performing analysis on the multidimensional data authorize us to strengthen the power and the abilities of the inspection and allow us to examine actual experience in the context of big data analysis and the main motive of the paper is to emerge the analysis that has been implemented on multidimensional data.

Jadav and Vaghela [8] classifies the reviews that are been provide by the people on the social media, as we know that social media is been among one of the dominant areas where we can get the reviews from various categories, the major job that has to be done is to extract the data and convert that data into organized way. Here we have taken optimized support vector machines into consideration for classifying the reviews so that we get the result with better accuracy. The essential motive of this paper is to inspect the efficient feature which provides us with better results and effective selection method, there can be a various feature selection method that can be applied. During the activity when data is been pre-processed they have eliminated the unclear data and the blank spaces. The differentiation is been made between the optimized SVM and Naïve Bayes and optimized SVM provide with better accuracy.

Tripathy and his team [9] presented a differentiation of the results that are been acquired by implementing the Naïve Bayes and the SVM algorithm techniques , both of this technique are used to analyze a sentimental review having the positive or a negative side, the training dataset that has been taken for the examining has been made for the critical inspection. As there has been the drastic increment of the internet services, people usually convey their reviews over the internet. Machine learning technique is beneficial to analyze and predict whether the review is in a positive or negative sentiment. This model will basically inspect the movie review which is given in the matrix representation from where the machine learning algorithm is been applied to train the model.

3. METHODOLOGY & RESULTS

In this digital era that we are living in everyone is obsessed with mobiles, computers and is on the internet, so job portals play an important role in aiding job seekers in their job hunt. It will make them aware of the various job openings and they will not miss an opportunity. Our objective with this project is to come up with a portal that will provide the user's details about various job openings in respective domains. The portal will stream data from Twitter to find out the recently published jobs. The business case in this project is that the portal will provide an apply early advantage to the users looking for the jobs in their respective domains. In this project, we are using external datasets provided by Twitter. We will harvest the data using APIs. The data will consist of tweets by various companies having job openings. Twitter streaming APIs offer Public streams which can be used to access all public data based on specific keywords. This data will be stored in the HDFS for analysis. We are using Apache Flume for data acquisition from twitter.

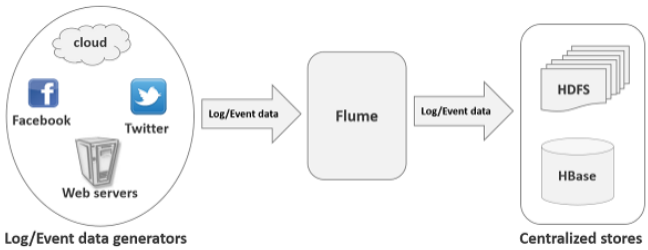

Fig. 1- Data Flow Diagram

After acquiring and filtering data we will still need to transform and normalize it into a uniform dataset which can be analyzed by our big data tool since the acquired data came from different sources with different Formats.

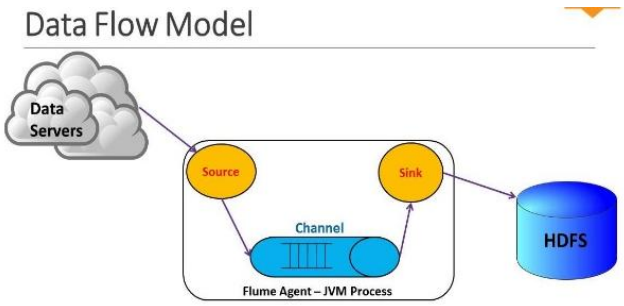

Fig. 2- Data Model

This formatting will also store only the data useful and scrap the excess data in real time. To prevent any data anomalies, we will have to validate all the data to ensure the data we gathered is readable, can be processed and doesn't lead to corrupt or bogus output. Languages, fonts, special characters, emoji all will have to consider and we will have to ensure that if not positive, it will not at least bring a negative contribution to the analysis. The even normalized dataset will be a large relational database with a lot of redundancies of the common attributes at least. Through these common attributes and unique keys will have to integrate all this data so that it could be seen through a single dimension at a time. Finally, the data will be analyzed and for one batch at a time and to maximize the efficiency of our use of multiprocessors and large memory the algorithm will be applied to the dataset to get the results. Our analysis will also be depicted in the form that would give an insight into the collected data at a glance. The reports of what trend was discovered more and in what region can be mapped using analytical graphics. For making the portal more efficient using the machine learning algorithm i.e., Logistic Regression, to classify the tweets into genuine and not useful tweets we have used the algorithm because some of the tweets don't contain relevant information. While training the algorithm on the gathered data of tweets and applying it on the remaining tweets for classification. We have achieved the 97% accuracy in the classification of actual tweets among all the collected tweets. For providing the User Interface (UI) for the users, we have used python for creating the interface in which we are displaying information from the classified tweets.

4. CONCLUSION

The tweets which contain the job openings, new opportunities, vacancies are gathered from the Twitter through streaming the tweets using its Application Programming Interface (API). Data is streamed using the Flume tool and stored inside the Hadoop cluster (HDFS). We have made a different analysis on the gathered data and the data which is streamed is in an unstructured format. We streamed a lot of tweets which containing the job openings, new opportunities, and several other vacancies. After using the machine-learning in the system made it more reliable and produced the valid results.

5. FUTURE DEVELOPMENT

There is some future development which can are planning to implement the information section based on the industry type and a profile based section to the users to produce an notification system in which user can utilize it to get notification on their selection choices so that they can avail this whenever any new tweets posted containing job opening details based on their selection

REFERENCES

1. Yu, S. and Yang, D. (2016). The Role of Big Data Analysis in New Product Development. 2016 International Conference on Network and Information Systems for Computers (ICNISC).
2. Zhang, W., Lau, R., & Li, C. (2014). Adaptive big data analytics for deceptive review detection in online social media.
3. "Product marketing improvement using big data analysis", International Journal of Science & Engineering Development Research (www.ijrti.org), ISSN: 2455-2631, Vol.2, Issue 3, page no.126 - 129, March-2017
4. Martınez-Castano, R., Pichel, J. C., & Gamallo, P. Sentiment Analysis on Multilingual Tweets using Big Data Technologies.
5. Razia Sulthana, A., Jaithunbi, A. and Sai Ramesh, L. (2018). Sentiment analysis in twitter data using data analytic techniques for predictive modelling. Journal of Physics: Conference Series, 1000, p.012130.

6. Elgendy, N. and Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. Advances in Data Mining. Applications and Theoretical Aspects, pp.214-227.
7. Cuzzocrea, A., Song, I. and Davis, K. (2011). Analytics over large-scale multidimensional data. Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP - DOLAP '11.
8. Jadav, B. M., & Vaghela, V. B. (2016). Sentiment analysis using support vector machine based on feature selection and semantic analysis. International Journal of Computer Applications, 146(13).
9. Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of sentimental reviews using machine learning techniques. Procedia Computer Science, 57, 821-829.
10. J. Sadhasivam, M. Kubendiran, P. Tomy, B. Jeyakumar, M. Sathish Kumar, and R. Anusha, “Review of Gaming and Its Evolution Over Networks,” Int. J. Civ. Eng. Technol., vol. 8, no. 11, pp. 61–68, 2017.
11. S. Nithya, M. Asha Jerlin, R. Charanya, S. Jayakumar, and R. Rathi, “Self Restorative Cluster Head Selection In Heterogeneous Network,” Glob. J. Pure Appl. Math., vol. 11, no. 3, pp. 1655–1662, 2015.
12. J. S. Ojaswa Swarnkar, Shubham Agarwal, Sanyam Jain, Nallakaruppan M.K., “E-learning through Wireless Sensor Networks using SATCOM and ZRP,” Int. J. Appl. Eng. Res., vol. 9, no. 12, pp. 2019–2025, 2014.
13. R. C, asha J. M, J. Sadhasivam, N. S, and R. Rohit, “A Case Study on Attack Models And Privacy Models In Mining Medical Datasets,” Int. J. Mech. Eng. Technol., vol. 8, no. 11, pp. 964–976, 2017.
14. J. Sadhasivam, S. Jayavel, B. Jeyakumar, and S. Merchant, “HOCS : Host Os Communication Service Layer,” Int. J. Civ. Eng. Technol., vol. 8, no. 11, pp. 35–41, 2017.
15. S. Jayakumar, S. Jayavel, and M. Senthilkumar, “Network Security – MAC Address Block,” Int. Conf. Netw. Commun. Comput., pp. 419–422, 2011.
16. S. Jayakumar, S. Jayavel, and N. S, “Automatic Campus Network Management using GPS,” Int. J. Comput. Sci. Issues, vol. 9, no. 3, pp. 468–472, 2012.
17. J. Sadhasivam, ashaJerlin M, and N. S, “Intelligent Interior Mapping using Wall Following Behaviour,” Int. J. Trend Res. Dev., vol. 3, no. 6, p. 112, 2016.
18. J. Sadhasivam, R. Charanya, S. Harish Kumar, and A. Srinivasan, “Identifying images of handwritten digits using deep learning in H2O,” IOP Conf. Ser. Mater. Sci. Eng., vol. 263, p. 042033, 2017.
19. J. Sadhasivam, M. Alamelu, R. Radhika, S. Ramya, K. Dharani, and S. Jayavel, “Enhanced way of securing automated teller machine to track the misusers using secure monitor tracking analysis,” IOP Conf. Ser. Mater. Sci. Eng., vol. 263, p. 042032, 2017.