Mesopotamian journal of Big Data
Vol. **(2021)**, 2021, **pp**. 1–5
DOI: https://doi.org/10.58496/MJBD/2021/001 ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BigData

Review Article

Big Data Sentiment Analysis of Twitter Data

Ahmed Hussein Ali,[*1], , Harish Kumar[2] , , Ping Jack Soh[3],

[1] *Alsalam University College, Iraq*

[2] *Department of Computer Science, College of Computer Science, King Khalid University, Abha 61413, Saudi Arabia*

[3] *Associate Professor, University of Oulu, Finland*

ARTICLE INFO

Article History

Received 22 Sep 2020
Accepted 13 Dec 2020
Published 15 Jan 2021

Keywords

Big data

Hadoop

Sentiment analysis

Social networks

Twitter

ABSTRACT

The term "big data" is becoming increasingly common these days. The amount of data generated is directly proportional to the amount of time spent on social media each day. The majority of users consider Twitter to be one of the most popular social networking platforms. The rise of social media has sparked an incredible amount of curiosity among those who use the internet nowadays. The information collected from these social networking sites may be put to a variety of uses, including forecasting, marketing, and the study of user sentiment. Twitter is a social media platform that is commonly used for making remarks in the form of brief status updates. A sentiment analysis may be performed on some or all of the millions of tweets that are received each year. Managing such a massive volume of unstructured data, on the other hand, is a laborious effort to do. To effectively manage large amounts of data, the analytics tools and models that are now on the market are insufficiently equipped and positioned. For this reason, it is essential to make use of a cloud storage solution for the applications of this kind. As a result, we have used Hadoop for the intelligent analysis as well as the storing of large amounts of data. In this article, we offer a system that does sentiment analysis on tweets using the Cloud.

1. INTRODUCTION

The quantity of information included inside social networking sites like Twitter[1], Facebook, and Instagram is increasing at a rate that can only be described as exponential on a daily basis[2]. This is due to the fact that there are an increasing number of users who are posting this information. These data contain a wealth of views and perspectives contributed by members of the general public. The process of gleaning information from data is becoming increasingly important in a wide range of fields, including dynamic pricing, emotional analysis, and many others. The authors of this study, Zhao et al[3]., suggest a dynamic pricing system based on the analysis of sentiment on Twitter. The method that they developed is called sentimental lexicon, and it involves collecting tweets from Twitter by using Twitter's APIs and extracting them using the phrase "electric." In order to acquire a score and a time stamp, the tweets are compared to a list of positive and negative phrases that have already been specified. This emotional tendency is utilised in order to create a pricing that is dynamic. It has been proposed by Subramaniam et al[4]. to develop a website for conducting surveys that will automatically update the findings of sentiment analysis taken from Twitter and would provide trending subjects in the order that they are now most popular. The Significance of Large Amounts of Data

1. Organizations have the capacity to enhance their performance, provide better customer service depending on the preferences of customers, and raise their profitability by utilising the Big Data that has amassed in their systems.

2. Big data is also used by researchers in the medical field to identify disease risk factors and to assist physicians in correctly diagnosing patients' illnesses and conditions.

3. Big data helps financial institutions understand the income and spending patterns of customers, which may then be utilised to make predictions and select appropriate banking products.

Big data analytics, on the other hand, is a complicated procedure that is utilised to determine client preferences, market trends, and undiscovered relationships in order to assist businesses in making educated judgments. There are several different analysis techniques for big data:

- Spark – It facilitates the processing of large amounts of data in real time and their subsequent analysis[5].

*Corresponding author. Email: ahmedhusseinali@alsalam.edu.iq

- Talend – It is utilised for the administration of data as well as the combining or integration of data[6].
- Cassandra- is a type of distributed database that is both open-source and free to use. Its primary purpose is to manage massive amounts of data[7].
- Storm- a type of distributed real-time computing system that is employed for the purpose of processing data in big volumes and at a rapid rate. Storm is a rapid database management system that can process more than one million entries per second on each node[8].
- Kafka- utilised for data streams that occur in real time, as well as the collection and processing of data. It is a platform that allows for distributed streaming[9].
- Hadoop- is a framework that is freely available and can assist with the storage and processing of data[10].

The remaining portions of the paper are structured as follows: In the next section, we will discuss both the background and the works that are linked to it. In section 3, we explore the gap that exists between the methodologies of sentiment analysis and the features of big data by demonstrating how to take each V of big data into consideration. The study is brought to a close in section 4, where we also provide some potential future research.

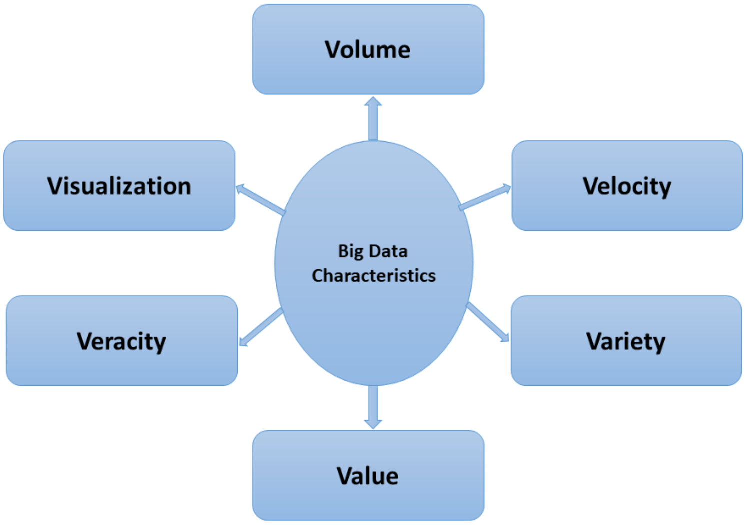

Fig.1.Big Data Characteristics

2. BACKGROUND

2.1 Literature

On the basis of a variety of metrics connected to the Indian Premier League that was played in 2015, the author of the research came to the conclusion that it is very popular. According to the findings of the research that was carried out, the Indian Premier League is not only well-known in India, but it has also attained an extraordinary level of popularity all over the world. The period of time during which cricket fans were most active on Twitter. The many different measures that have already been evaluated are quickly becoming the most discussed players. Using HADOOP, the author[11] of the research accomplished similar work in her paper by analyzing the data from Twitter, which is also known as big data. The analytics tools and methods that the author of the study utilised are insufficient to manage huge data, however they are used in the paper[12]. Because of this, there is a prerequisite of making use of cloud storage for the kind of applications being discussed here. Hadoop has been applied by the author for many intelligence processing and storing of large amounts of data. The author of this study presented a method for analysing the sentiment of tweets in the Twitter platform. Cloud environment. The conclusion of the paper was that a lot of businesses were able to enhance their client retention rates with the assistance of big data and the capabilities that it offers, which in turn helped the businesses gain speed and complexity. E-commerce businesses frequently analyse the traffic on their websites or the patterns of travel to identify likely perspectives or interests of an individual or a group as a whole based on the products that have been purchased in the past. They took all of these considerations into account before comparing the findings acquired from the various data analysis technologies.

2.2 Platforms

A. Spark

Spark[13] is a platform for performing cluster computing in real time. It is a data processing engine that was developed to provide analytics in a rapid and easy manner. It can also share data across numerous machines. Both of these qualities are essential in the fields of machine learning and Big Data[14]. Spark also makes it easier for developers to programme by providing an easy-to-use application programming interface (API) that decreases the amount of effort involved in the processing of big data and distributed computing. The AMP Lab at the University of California, Berkeley was responsible for the development of Spark. After its completion, the Apache Program Foundation gained ownership of the software.

B. Hadoop

Hadoop[15] is a Java-based open-source platform that facilitates the storage and processing of Big Data across several computer clusters using just basic code. Hadoop was developed by Apache. Hadoop makes use of a collection of algorithms known as MapReduce to partition a task into smaller pieces before allocating those portions to a collection of computers. This enables Hadoop to store and retrieve data from its nodes more quickly.

3. SENTIMENT ANALYSIS

The process of identifying the underlying feeling conveyed by a string of words is referred to as SA[16]. This analysis is utilised to gain a better understanding of the ideas and feelings that are stated in a statement that is found online. The relevance of SA is growing steadily as a result of its utilisation of big data technologies, which make it possible to handle any form of data. Not just in the realm of scientific study, but also in that of advertising and marketing, SA has recently garnered a significant amount of interest in recent times. Recent developments in social networks and the speed at which information can be sent are mostly responsible for this. In addition, large amounts of real data collected from social networks are frequently used for the purpose of precisely analysing opinions. On the other hand, analysing recent messages collected from social networks has the potential to provide users and decision-makers with a general opinion regarding a particular subject. The first step in the SA process is to determine if the polarity of the text should be considered positive, negative, or neutral. The polarity may be determined based on a number of various thresholds, and it can be interpreted as one of three distinct classes using the following three methods:

1) Lexical approaches: Using a semantic analysis of the words in a sentence, it is possible to deduce the opinion that the statement is trying to convey. Using this method, the sentence is classified based on examples of previously written sentences that have already had emotions recognised and a polarity score assigned to them. Specifically, these previously written phrases have been given a polarity score. As a result, we consult dictionaries that provide references to the annotated terms of the polarity as well as the context for which it is appropriate. The polarity score for the entire document is modified whenever a new word is discovered.

2) Learning approaches Because of its versatility and precision, machine learning has emerged as one of the most significant methodologies that is attracting the attention of scholars. The supervised learning variations of this method are the ones that are used almost exclusively in sentiment analysis. For the purpose of supervised learning, there is a collection of data that has been labelled, which may take the form of examples that have been assigned to a category by a trainer or an experienced person. The foundation for learning is provided by this collection of illustrations. The objective of supervised learning techniques is thus to construct, beginning with the learning base, classifiers or ranking functions. These are the kinds of functions that make it feasible, based on the description, to recognise a certain feature, or the class.

3) Hybrid algorithms: Researchers have been exploring the possibility of a hybrid method in response to recent developments in the field of sentiment analysis. This method would collectively present the precision of an automatic learning approach as well as the speed of a lexical one.

3.1 Living with The Data

A Twitter dataset presented as an example is comprised of two different sets. The first one, which is called Twitter 1, is a curated database that has 300 million records that are geotagged. The second collection, referred to as Twitter 2, is an archive of one billion unprocessed tweets that were posted over a period of 15 months[17]. The full dataset may be inspected in order to formulate initial hypotheses thanks to interactive visualisations of all the data. Analysts have the ability to traverse between high-level overviews and the most minute levels of information. This "living with the data" approach used by EDA makes it possible to observe emergent structures and traits.

3.2 Processing of Twitter Data

The framework that has been developed includes four primary operations. which comprise gathering data from many social networks, using a big data management system that involves the storing and preprocessing of aggregated data, extracting and categorising attitudes, and displaying the findings that have been produced. The results will be used to plot visual data on the dashboard and will also offer a rating list for the services, complete with their current scores.

A. Data Collection

Twitter, Facebook, Google My Business (GMB), Blogs, YouTube, and TripAdvisor are just some of the places online where people share their thoughts and ideas. As a result, there is a massive amount of content available online. Additionally, each platform provides its own means by which users can access the data. Those enormous quantities of information that are now accessible over the internet by way of application programming interface (API). In this research project, the Twitter and GMB application programming interfaces (APIs) were investigated in order to extract sentiment texts from tweets and google reviews that were given to individual businesses.

B. Big Data Management

At this point, big data environments like Hadoop Distributed File System (HDFS) and Apache Spark are being evaluated for storing and analysing enormous volumes of data. During the process of integration, they will help to match and categorise the services and reviews that are provided to each business by ensuring that the design and execution are carried out in an effective manner.

C. Preprocessing:

Apache Spark is used for the study because of its capacity for high performance and its ability to speed up BDP. Apache Spark's Resilient Distributed Dataset (RDD) is, in fact, the representation of a collection of data that is spread over numerous computers. It also includes application programming interfaces (APIs) that enable developers to take action on the data. RDD refers to a distributed, immutable, fault-tolerant collection of things that may be worked on in parallel.

D. Results Visualization

The objective of this work is to provide a graphical representation that concisely communicates the outcome that was reached after using the categorization method. Some of the information that has to be represented may be found below: In order to provide users with an idea of which service they should think about using. Which services are receiving the most positive comments? Which specific kind of services are receiving the highest ratings? What are the phrases that appear most frequently in negative reviews? to get reports that can serve as a foundation for further research and decision-making inside enterprises.

4. CONCLUSION AND FUTURE WORK

For social users, the above-mentioned model for the analysis of survey data from Twitter has been presented. This model may be applied by surveying tweets according to the emotions expressed in them, which is the number of users who have submitted their tweets on Twitter. It automatically updated on websites in order to know what the most recent common conversation was that was surrounded by social users. A rating method that is based on the reactions from social media was presented, and the study also introduced four primary activities. The paper analyzed the impacts of SA on SM on top of the big data environment to evaluate the services in Rwanda, and it ranked the services. These responsibilities may be thought of as separate modules that make up the programme. These modules perform operations inside the application such as collecting, managing, mining, and presenting text data that indicates a customer's opinion regarding a service. Additionally, those modules may be maintained and updated independently from one another without having an impact on the others.

Conflicts of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors would like to express their gratitude to the Iraqia University, the Informatics Institute for Postgraduate Studies, and the Al Salam University College for their moral support. Please accept my sincere gratitude for the useful recommendations and constructive remarks provided by the anonymous reviewers.

Funding

The authors receive no funding for this work.

References

[1] K. H. Manguri, R. N. Ramadhan, and P. R. M. Amin, "Twitter sentiment analysis on worldwide COVID-19 outbreaks," *Kurdistan Journal of Applied Research,* pp. 54-65, 2020.

[2] N. D. Zaki, N. Y. Hashim, Y. M. Mohialden, M. A. Mohammed, T. Sutikno, and A. H. Ali, "A real-time big data sentiment analysis for iraqi tweets using spark streaming," *Bulletin of Electrical Engineering and Informatics,* vol. 9, no. 4, pp. 1411-1419, 2020.

[3] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," *Expert Systems with Applications,* vol. 110, pp. 298-310, 2018.

[4] G. A. Ruz, P. A. Henríquez, and A. Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers," *Future Generation Computer Systems,* vol. 106, pp. 92-104, 2020.

[5] J. Wan, Y. Zhuang, Y. Huang, Y. Qian, and L. Qian, "A review of water injection application on spark-ignition engines," *Fuel Processing Technology,* vol. 221, p. 106956, 2021.

[6] C. K. Emani, N. Cullot, and C. Nicolle, "Understandable big data: a survey," *Computer science review,* vol. 17, pp. 70-81, 2015.

[7] K. Anusha, N. Rajesh, M. Kavitha, and N. Ravinder, "Comparative Study of MongoDB vs Cassandra in big data analytics," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 1831-1835: IEEE.

[8] C.-C. Lee, M. Maron, and A. Mostafavi, "Community-scale Big Data Reveals Disparate Impacts of the Texas Winter Storm of 2021 and its Managed Power Outage," *arXiv preprint arXiv:2108.06046,* 2021.

[9] A. Bandi and J. A. Hurtado, "Big data streaming architecture for edge computing using kafka and rockset," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 2021, pp. 323-329: IEEE.

[10] O. Azeroual and R. Fabre, "Processing big data with apache hadoop in the current challenging era of COVID-19," *Big Data and Cognitive Computing,* vol. 5, no. 1, p. 12, 2021.

[11] X. Zhang and Y. Wang, "Research on intelligent medical big data system based on Hadoop and blockchain," *EURASIP Journal on Wireless Communications and Networking,* vol. 2021, no. 1, pp. 1-21, 2021.

[12] K. J. Merceedi and N. A. Sabry, "A Comprehensive Survey for Hadoop Distributed File System," *Asian Journal of Research in Computer Science,* 2021.

[13] R. Guo, Y. Zhao, Q. Zou, X. Fang, and S. Peng, "Bioinformatics applications on apache spark," *GigaScience,* vol. 7, no. 8, p. giy098, 2018.

[14] A. H. Ali, "A survey on vertical and horizontal scaling platforms for big data analytics," *International Journal of Integrated Engineering,* vol. 11, no. 6, pp. 138-150, 2019.

[15] K. K. Reddi and D. Indira, "Different Technique to Transfer Big Data: survey," *IEEE Transactions on,* vol. 52, no. 8, pp. 2348-2355, 2013.

[16] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment," in *Proceedings of the international multiconference of engineers and computer scientists*, 2019, vol. 122, p. 16.

[17] H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. Zhang, "Energy big data: A survey," *IEEE Access,* vol. 4, pp. 3844-3861, 2016.