Examining Heterogeneity Structured on a Large Data Volume with Minimal Incompleteness

Nahla Aljojo

Department of Information system and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

***Abstract—Whereas* Big Data analytics can provide a variety of benefits, processing heterogeneous data comes with its own set of limitations. A transaction pattern must be studied independently while working with Bitcoin (BTC) data. Hence, this study examines twitter data related to BTC and investigate communications pattern on BTC transactional tweet. Using the hashtags #BTC or #BTC on Twitter, a vast amount of data was gathered, which was mined to uncover a pattern that everyone either (speculators, teaches, or the stakeholders) uses on Twitter to discuss BTC transactions. This aim is to determine the direction of BTC transaction tweets based on historical data. As a result, this research proposes using Big Data analytics to track BTC transaction communications in tweets in order to discover a pattern. Hadoop platform MapReduce was used. The finding indicate that in the map step of the procedure, Hadoop's tokenize the dataset and parse them to the mapper where thirteen patterns were established and reduced to three patterns using the attributes previously stored data in the Hadoop context, one of which is the Emoji data that was left out in previous research discussions, but the text is only one piece of the puzzle on BTC transaction interaction, and the key part of it is "No certainty, only possibilities" in BTC transactions.**

***Index Terms*—Heterogenouis dataset; Bitcoin transactions; Bitcoin tweets**.

I. Introduction

Different people may have varying motivations for using Bitcoin (BTC) because it can be used for a variety of purposes, including investment, transfer (both sending and receiving) from one BTC wallet to another, and for a variety of other purposes as well. When people want to stay up to date on the newest events in the world of BTC transactions, one of the best ways to do so is to use Twitter (Grover, et al., 2019). Unfortunately, because of different caliber of people on Twitter, the information about BTC transactions that would be shared on Twitter would come from a variety of different sources, making it difficult to verify. However, some well-known people, for example, Tweets from Elon Musk about BTC, for example, have had a significant impact on the current state of BTC transactions. It is possible that other individuals who tweet about BTC will not do the same. Now, it only makes sense to figure out how important a BTC-related tweet is and how significant it is. This is associated with the transaction. The focus of this study is not associated with BTC sentiment analysis on Twitter. This study formulates this concept to focus on transactional information, which involves all the various data sources that BTC has access to. With everything pertaining to BTC transaction data, the whole system is forced into a state of heterogeneity. This means that the information contained in BTC transaction data is associated with a heterogeneous property. In the BTC market, there is a great number of speculators and tech-savvy investors on Twitter coming together in the midst of turbulent market conditions, where, they switch to a contrarian strategy when market volatility increases (Lee, Li and Zheng, 2020).

The research challenge or problem in this scope is to trace a pattern to which everyone (speculators, tech-savvy, and stakeholders) use on Twitter about BTC transactions. Whether there is an established direction of certain Tweets from information of historical investment on price depreciation, appreciation, and fluctuation. Huge dataset is expected in this situation, and if the dataset is large, the chance of it being heterogeneous increases. This has a direct relationship to a Big Data characteristic. As a result, heterogeneous dataset, from BTC transaction data associated with a tweet, can be any data that contains a high variability of data types and formats. A large amount of information can be retrieved from Twitter by using the hashtags #BTC or #BTC. Because so many people of varying calibers are responsible for posting that information, it is possible that the data will be ambiguous and of low quality, as well as containing missing value values. In fact, it should be expected to contain a high level of data redundancy as well as inaccurate and sometimes misleading information. It would be difficult to integrate all of these disparate data sources in order to achieve a comprehensive understanding of the real outcome experienced by people during a BTC transaction.

The research gaps that was established lies with the fact that whereas there are millions of people who invest in BTC, their motives for doing so are not as well understood as those for making other kinds of investments (Mattke, et al., 2021). Furthermore, because of the variety of data

ARO-The Scientific Journal of Koya University
Vol. IX, No.2 (2021), Article ID: ARO.10857, 8 pages
DOI:10.14500/aro.10857
Received: 18 August 2021; Accepted: 16 October 2021
Regular research paper: Published: 02 November 2021
Corresponding author's e-mail: nmaljojo@uj.edu.sa

acquisition techniques available, the data generated from BTC tweets is frequently of personal views (Thelwall, Buckley and Paltoglou, 2011), but there is no prevision of stashing some pattern. Similarly, those data would be on a large scale due to the fact that there would be historical data on the nature of BTC performance within a specific time frame available (Kamps and Kleinberg, 2018). When it comes to data acquisition, there would be a strong correlation between the time and space in which it was done, as well as the geographic location of each piece of information (Yue, et al., 2018). After all is said and done, the effectiveness of the data would only account for a small portion of the decision (Dwyer, 2015). Because there would be a large amount of noise data that could be collected during the acquisition, this is an important consideration in establishing pattern for better understanding.

Given that BTC can be regarded as a private currency (Dwyer, 2015), and that the data generated as a result of its transactions can multiply in an exponential manner each day (Gandomi and Haider, 2015), and that the total number of BTC owners is approximately 100 million, huge information can be generated by Twitter about its transaction. Tweeter had 199 million daily active users with 500 million tweets/day as of the first quarter of 2021, while considering that when it comes to big data, the first and, in some cases, the only dimension that comes to mind is its sheer size (Sean, 2021). This study proposes to use Big Data analytic to analyze BTC transaction tweets to establish a pattern of the direction of the BTC market investment.

II. Related Work

Big data analytic is currently a very important area of study in data science, similarly BTC dwell one of the major part of computer studies. In terms of working with both areas, it creates the possibilities to show the process by which the data were collected changed the structure of the resulting dataset to draw some conclusion.

The extreme diversity of data in the Big Data world is one of the most difficult challenges to deal with (Christophides, et al., 2020). The previous study establish that heterogeneity can be created by grouping data from different experiments together in a single analysis (Lugli, Roederer and Cossarizza, 2010). In this instance, it is assumed that an experiment designed to develop heterogeneity will be the one to bring it to light. Each experiment will be tailored to the study's specifications, which means that the data generated by these experiments is likely to have a diverse range of characteristics (Bridges, et al., 2020). On the other hand, natural heterogeneity is also established by the underlying data sources, for example, in medical data (Yue, et al., 2020), Labor market data (Kikuchi, Kitao and Mikoshiba, 2020), Landscape data (Urrutia, et al., 2020), business data (George and Kabir, 2012), all these can be attributed to "data pooling" as a well-known cause of heterogeneity in data analysis.

When performing data analysis, it is common for the data distributions attribute to be non-normal, and that is where the first problem of heterogeneity arises (Kazemi and Hassanzadeh, 2020). Variety is one of the most important characteristics of Big Data, and it is closely associated with the heterogeneous nature of data sources (Hashem, et al., 2015). Given that Big Data encompasses the various data types that make up a dataset, it includes datasets that contain structured, semi-structured, and unstructured data, among other things (Casado and Younas, 2015). The previous research identified that, the vast majority of the data is structured and extremely well organized (Schulze, et al., 2018), whereas in data science, many believe that managing unstructured data is one of the most difficult problems because of the tools and techniques that are successful with handling structured data are not as effective with unstructured data, which results in much confusion (Blumberg and Are, 2003; Malik, Burney and Ahmed, 2020). One of the shortcomings identified by previous research studies in dealing with mixed-up data is that it is not entirely within the control of the application running it (Alkatheeri, et al., 2020).

It has been highlighted that there is a chance of receiving incorrect results in any situation when working with a variety of data types (Cappa, et al., 2021). This is also associated with BTC data. Big data analytics and block chain technology has been discussed in (Krithika and Rohini, 2020). The previous studies associated with BTC and big data are mostly focusing on price prediction, typical to these studies are the work of (Lahmiri and Bekiros, 2020) who investigate a multifractal analysis of BTC's price and volume relationships involves looking at a range of time scales, from the nanosecond to the year, to illustrate various multifractal processes that are occurring at differing intervals. Similarly, Big data analytics has been found to be useful for to identifying illegal activities on BTC transactions (Kumar, et al., 2021). Prediction of price of BTC has been to more appropriate with Big data analytics (Dutta, Kumar and Basu, 2020). In same direction, Big data have been used to identify the most time-efficient and accurate model for predicting the price of BTC ((Shankhdhar, Singh, Naugraiya and Saini, 2020).

Finally, it is important to note that all of the previous research in the areas of BTC and Big data reviewed in this study is either concerned with the problems of heterogeneity in data (Christophides, et al., 2020; Alkatheeri, et al., 2020) or with the application of big data for the prediction of BTC price or some part of BTC operations (Cappa, et al., 2021; Shankhdhar, et al., 2021). Pano and Kashef, 2020, who investigated sentiment analysis of BTC Tweets, served as inspiration for this study, which identified a gap in the current state of the art by not applying analysis of BTC tweets to the general BTC operations that are associated with its transaction.

III. Methodology

In the context of this study, a more specific phrase for "research methodology" is one that is focused on problem-based methodologies. This is critical when dealing with BTC transaction data, which is dealt with one transaction pattern at a time and must be dealt with on an individual basis. The study investigates the structural heterogeneity that is associated with the abundance of data derived from the (#BTC and #btc hashtags) to better understand how minimal presentational incompleteness can be achieved on BTC-associated tweets data. One of the biggest problems in the realm of Big Data is extreme diversity. That is why Big data analytic was adopted using Hadoop platform, to test for patterns from heterogeneity data.

In performing this kind of test with Hadoop, there is an intrinsic property of data analytics that need to be considered as the major problems to solves, such as multimodality, incompleteness, and noise, which constitute a considerable obstacle to the successful deployment of revolutionary data analytics algorithms (Vaduva, Iapaolo and Datcu, 2020). This strategy is essential because it is critical to examine applying data analytical techniques, for the input data in such a way that it can be comprehensive and well-formatted for the technique to be effective; otherwise, the technique will be rendered ineffective. Many people have had trouble sorting through all of the data in big data analytics, and that can be seen in the amount of data that has been collected (Hu, et al., 2014). That is why when faced with the challenge of analyzing a huge volume of unstructured and semi-structured data, identifying meaningful patterns can be challenging because the material has nothing in common. For this current research methodology, unreliable, partial, and imprecise data in the real world about BTC transaction will be analyzed. It’s expected that the analytical technique might lead to a poor application performance under certain circumstances because the datasets are tweets.

A number of data cleaning and integration approaches are used in conjunction with data transformation to remove noise from the data. The methodology is also responsible for the data transformation of the cleaned data. This type of technique is used in the context of traditional data analytics. For the fact that BTC tweets come with numerous challenges associated with big data, new techniques for data cleaning are required to deal with the uncertainty and data quality problems that arise as a result of data noise and inconsistent data, as well as the sheer volume of data size (Dey, et al., 2019). The justification of applying the propose methodology lies with these noise-removal techniques and the quality of the data analysis from Hadoop. That is why the findings obtained from the analytic can be greatly improved as compared with the previous method that relies on sentiment analysis.

A. Dataset

The dataset used in this study was obtained from Kaggle (Kaushik, 2021), which consisted of “BTC Tweets with trending #BTC and #btc hashtags” and “BTC Tweets with trending #BTC and #btc hashtags.” The dataset for tweets can be obtained using the R-package, but due to the limited number of tweets and time that can be downloaded from each account when using the package, as well as the fact that data from Kaggle are open for use by many researchers to have some comparison, the Kaggle dataset was used in this study. The dataset is analyzed since it is associated with heterogeneity in the direction of redundancy or missing values, the results of analysis is correlated with properties of BTC values.

B. Analytical Framework

The research analytical framework is presented in Fig. 1. Pattern types are used by the project to accommodate a diverse range of data types and transaction ideas. These pattern types are described in greater detail to present data types. Among other things, classes can be classified in a variety of ways, including across multiple languages and even within a single language, among other things. Note that, despite the fact that all of the data sources make reference to the same entities, each of the data sources refers to those entities using a different name to distinguish them from one another. In situations where different domain modeling models are applied to the same subject matter in a variety of ways, the term “semantic or logical mismatch” can be used to describe the differences in domain modeling approaches. The fact that this is true even when concepts are defined through the application of a variety of axioms is what gives you the confidence to believe that it is true in the first place. This is due to the fact that their conceptualizations are diametrically opposed to one another, which allows them to be easily distinguished from one another. As a result of the various ways in which the concepts have been modeled, there is a misalignment in the concepts as a result of the misalignment.

When two or more data sources describe the same geographic region at the same level of detail but from a different point of view than the first, the coverage of those two or more data sources is defined as the sum of their coverage when the two or more data sources are combined. When two datasets that describe the same geographic region, but with finer-grained detail, are compared, the datasets with different levels of detail (granularity differences) are referred to as a dataset with different levels of detail (granularity differences) in the literature (also known as a dataset with granularity differences). Data that were generated as a result of incompatible points of view or, more specifically, differences in scope between two parties are referred to as “incompatibility data,” which is another term for it. Semiotic heterogeneity is defined as the phenomenon of different people interpreting the same entity in a way that is different from their own interpretation. The use of computers to detect and solve heterogeneity in a complex problem, such as this one, is required when dealing with a problem that can only be solved by humans in this situation. The fact is that this is a difficult task for computers to complete and is therefore not recommended.

C. The Analytic Experiment

Based on the analytical framework presented in the previous section, the Hadoop distributed data management and processing system was used in this study, specifically because it allows for distributed data management and processing. The direct impact of utilizing the Hadoop lies with, tokenizing formatted data and parsing the data during the map phase and preparing the data for the Hadoop map reduce cycle. The pre-

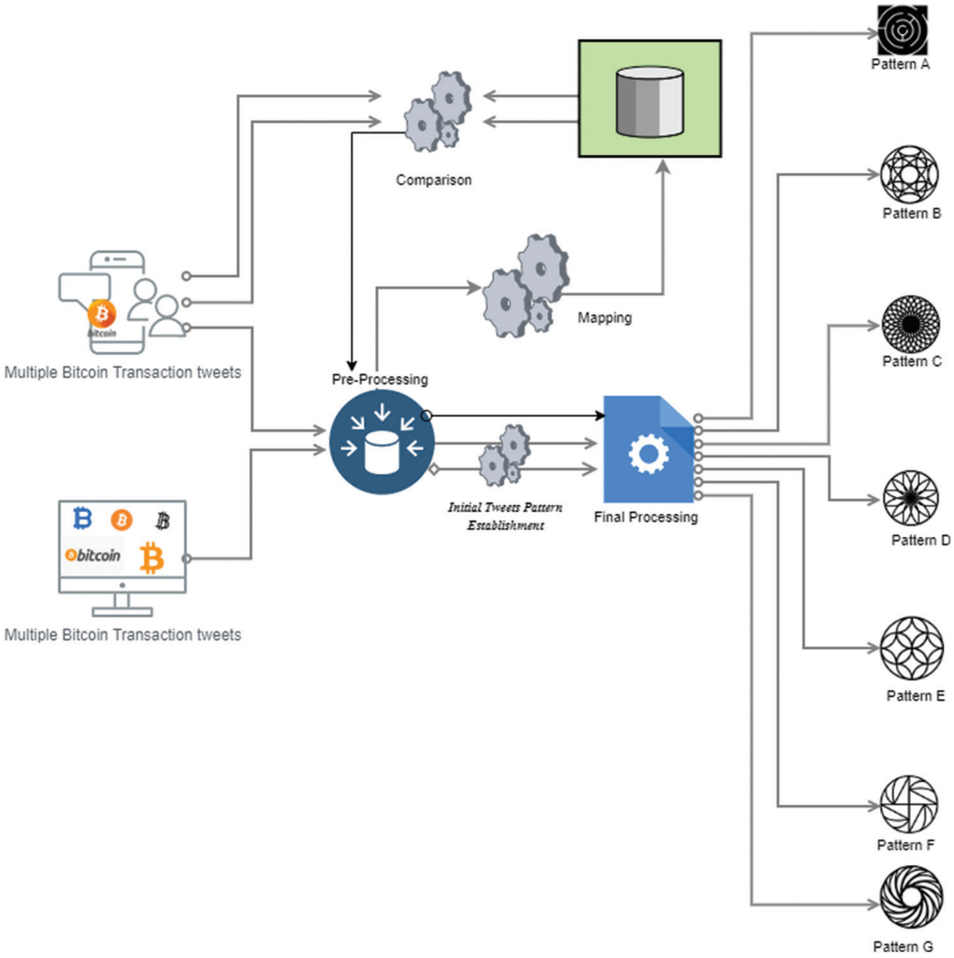

Fig. 1. The research analytical framework.

proceed data retrieved the attributes and transferred them to the Hadoop context, where they were used in the reduce step to aggregate the results for the particular data type in question. The Hadoop Distributed File System (HDFS), which was developed by Google, improved the efficiency of MapReduce processing and is well-suited analyzing BTC tweets because it is a clustered storage file system with increased bandwidth and self-healing capabilities. It is scalable with good error detection, and enhanced response handling. It was also designed with built-in redundancy, means failure is not a concern, and it is capable of achieving high availability through the use of parallel computing techniques (Abubakar, El-Gammal and Alarood, 2020).

Among other things, the cluster in Hadoop includes functionality that protects information stored on the nodes in the event of a node crash or failure, among other things. For each data node, three copies of its data are stored on the server by default, according to the default configuration. When a cluster has multiple data nodes, the HDFS manages and balances the data load among them, allowing the cluster to add or remove data nodes as needed. This study distributes our Name Nodes across multiple points to avoid having a single point of failure, so that they can continue to function even if one is lost due to a node failure.

D. MapReduce

MapReduce is used to construct the analytical framework for this study. MapReduce allows for the analysis of large datasets in a parallel manner. In this study, MapReduce was used to read each piece of BTC transaction data one at a time and record it into a hash table, with each data unit serving as the key and the value representing the number of times it occurs. When a unit of data does not fit into a predefined pattern, it is added and given a numerical value. It is also possible that a data unit exists in the hash table, in which case the value of that data unit is updated to reflect the new occurrence count. This operates in a serial fashion, with the amount of time required increasing in direct proportion to the complexity of the data and the size of the data set being processed. When dealing with a large dataset, this is, without a doubt, extremely difficult to compute. As a result, serial data possession is not feasible and takes an inordinate amount of time, and we must implement parallel data processing.

The research analysis process begins with the division of the input sequence into several files. This number has been assigned to a number in the sequence of numbers. The Mapper returns an associative key/value pair as a result of its operation. Values from different mappers that have the same key combine to form a single value. In other words, a pattern has been established and maintained. The A-G pattern was laid out in the manner depicted in Fig. 1. Each pattern has a value, which corresponds to the first Mapping in the MapReduce process, which is the first stage. The data elements in the input file are combined to form a set of key-value pairs. Each symbol is represented by two symbols when the mapping function is used to count. It is a mapper function, which means that it does not change the data in the input list but instead returns a new output list. Reducers, also known as reducers, are used to switch between the outputs of the mapping step. Reducer nodes each have a subset of keys that are unique to them. In order to generate patterns, this subset is the next reducing step (partitioning). A task has the ability to push key-value pairs to any partition it chooses. On every mapper/reducer, the result is the same. Reducers are capable of dealing with a wide range of keysets. Prior to the keys being placed into the next stage, sorting, they must first be reduced. Finally, reduce the size of the object. Each reducer key necessitates the inclusion of a single instance of user-provided code. The inputs are a loop with a key on top of it, and the output is a key. It is possible that the iterator values are in an unexpected order. Each reducer produces a single output file.

The Kaggle dataset was preprocess, after setting up MapReduce by using IntelliJ IDEA instance. This was followed by creating a Java Project and downloading Hadoop from its official website where it was extracted into the working directory. Finally, in order to extend and perform the MapReduce tasks, we will require a few jar files from the Hadoop files. They will be dependencies that are required for the experimentation and will be referenced by the code. Once the directory has been created, the next step is to set up an input directory within the project. In here, we will detail which data is required for analysis.

IV. Presentation of the Results

After initial mapping g of the dataset by #BTC and #btc, MapReduce examined one piece of the BTC transaction data at a time and recorded each value and key as the sum of all occurrences in a hash table. That is following the isolations of dataset generated by #BTC and #btc, thirteen patterns where mapped under various categories (Fig. 2), where 1 (String, 1), 2 (Boolean, 2), 3 (Text, 3), 4 (Datetime, 4), 5 (Timestamp, 5), 6 (Date, 6), 7 (Hyperlink, 7), 9 (Currency, 9), 10 (Emoji, 10), 11 (Tags,11), 12 (Floating- point number, 12), and 13 (Character, 13).

When the data are not part of a specified pattern categorized from Fig. 2, it is given a separate value. Data units can exist in the hash table, in which case the value is updated to reflect the new occurrence count. The more complex the data, the more time it takes to process it, but evaluating A hash map is used to keep track of a list of records. It is likely that several highly active processing operations are going on simultaneously, because, it is difficult to compute when the dataset is large and many operations are going through the processing pipeline at the same time. For this particular reason, the serial data processing time-consuming that is the time needed to process a certain amount of data in dataset was measured (Fig. 3).

Whereas, it requires parallel data processing to be implemented. The mapper specifically mapped [map(π,μ)→list(λ,β)] where π is the in-key and μ is the in-value for which λ represent intermediate key, β represent the intermediate value. This application makes it possible to store data in a distributed manner. It also contributes to the reduction of a significant amount of data. In MapReduce, the terms reducing and mapping are used interchangeably. This research is currently in a position where it must complete the current mapped task before moving on to the next. As a result, the complete mapping was given by [{π, 1}, {μ, 1}, {λ, 1}, {β, 1}]…, [{π, n}, {μ, n}, {λ, n}, {β, n}] where the reducer modules consolidate the intermediate map data, thereby alleviating the processing load on the underlying framework. Among the 13 patterns mapped using attributes previously stored in the Hadoop context, the three out of them that were formed after the reduced operation, includes many used of "Emoji" "no certainty, only possibilities" and "patience" which were mostly associated with web link. However, the text is only one piece of the puzzle in terms of BTC transaction interaction. Whereas analyzing the "text," it was reveals that many texts that are associated with hyperlink and uniform resource locator (URL) were connected to a secured connection. That is apart from the term "BTC" and "Crypto" web link dominated a connection with all other

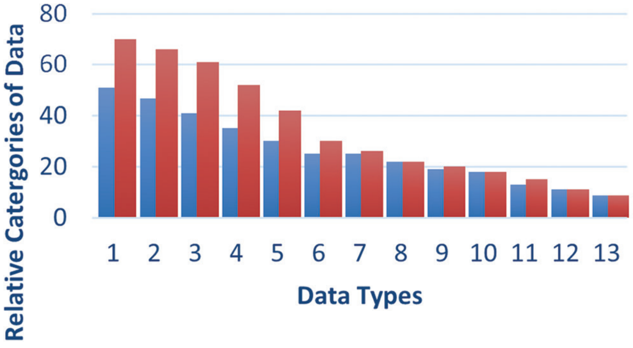

Fig. 2. The relative mapped data.

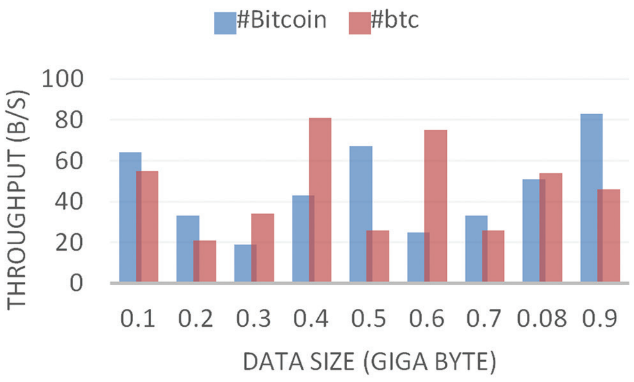

Fig. 3. The time required to process the Mapped #Bitcoin and #btc amount of data.

data types in the interaction of Bitcon transaction in BTC tweets (Fig. 4). The data-network map highlighted t.co at high frequencies and while following down the link. One of the most significant data types discovered is the Twitter automatic shorten URLs data, which is represented by the t.co. Twitter shorten URLs are displayed on any post that shares a URL on Twitter. There is no weblink in a tweet that can be used to opt out of link shortening. In tweets, you can combine URLs that are no longer than a maximum number of characters and tweet out only the shortened link, without jeopardizing the character count for the rest of the message.

It turns out there was unknown "emoji data" and other known emoji. No sure thing, just a lot of possibilities, followed by patience Even so, the text is only one of many aspects to take into consideration when considering BTC transaction interactions. Those emoji interactions were created in order to compensate for the lack of presentation of a more comprehendible data format. Typically, the finding of this study reveals that when emoji were used, the recipient was able to understand the based on emotions of the sender. Tweets that use emoji appear to be friendlier and more extroverted than those who do not. This suggests that it is possible to achieve an increase in approachability and competence in BTC transactional tweets with emoji. Many tweet rely heavily on emoji because they're vital for conveying emotions through visual expression.

Even though it has been revealed that the text is only one piece of the puzzle when it comes to BTC transaction interaction, and that one of the most important pieces is the statement, it is important to note that "no certainty, only possibilities" in BTC transactions and "patience" are the reduced form of the text outcome of the Hadoop. Hence, the cloud-tags of the text data type were with keyword clouds, also known as keyword cloud visualization are drawn (Fig. 5). The term "price" has the highest frequency a part from the term "BTC," "Cryto," and "BTC" Lots of tweets lies with BTC's price and its volume, and the data types associated with this interaction are mostly "Datetime," "Timestamp," and "Date."

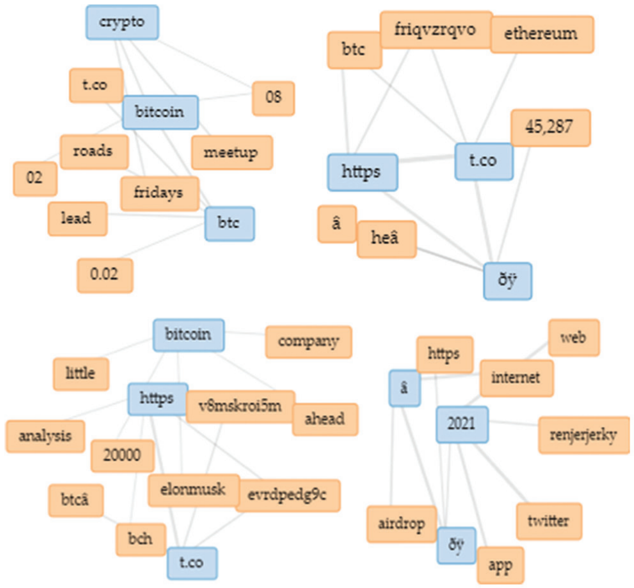

Fig. 4. Networked pattern of the Bitcoin transaction tweet.

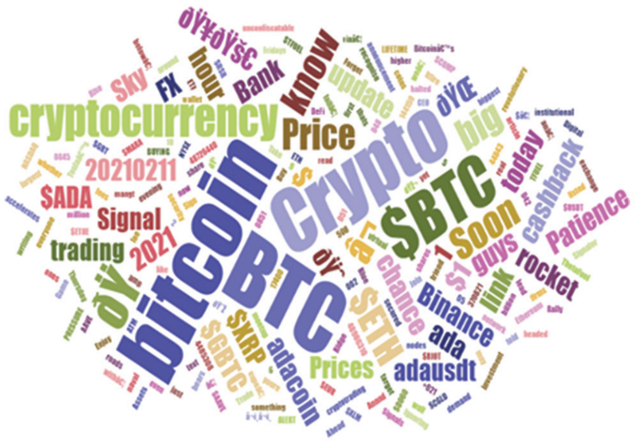

Fig. 5. The keyword cloud visualization of the Bitcoin tweet.

V. Discussion

Whereas many study deals with text analysis concerning characters or strings, this study found it important to include other data types in analyzing the communication between people in regards to BTC transactions. MapReduce has been used, and it mapped out thirteen key attributes of other data types used with the inclusion of string and text. These data were used for interaction about The BTCs transaction. One of the key terms that comes out about the data is the word "Patience." This has been used when the transaction was in turbulent case. Other data types that are mostly not been considered by previous research is emoji. It was found that emojis are attached to messages. The senders' motivations for sending a message may be different.

In light of the massive amounts of data that are being collected as a result of real-time transactions that occur around the clock, this study first supports the perceptions of the technical community about the massive amount of new data that are being generated every day, and it is referred to as possessing the characteristics of Big Data in the scientific community. Big Data analytics has made a significant contribution to the success of data science, especially when taking into consideration the ability to have an analytical technique that can deal with massive amounts of information. To draw some reasonable conclusions from BTC transactions, Big data can be used to handle the data from the transactions. Whereas Big Data analytics can provide a variety of benefits in general, as well as to the decision-making process for BTC transactions in particular, there are some limitations to the way heterogeneous data is handled when the technique is employed. This means that when working with BTC data, there will be a plethora of different sources that will be required to be examined with each transaction based on how it appears in the pattern of transactions. The data associated with BTC tweets is examined in this study, in contrast to the previous research, to determine the amount of minimal presentation incompleteness that can be achieved using the data. When the hashtags for BTC and BTC volume were used on Twitter, large amounts of information could be retrieved, which was then mined to uncover a pattern that was used by everyone.

It is critical to use Twitter to communicate BTC transaction information in order to ensure a successful transaction. As a result, BTC transaction tweets are influenced by data and information from the previous day. It is proposed to use Big Data analytics to identify a pattern in Twitter message transactions to track down BTC communications in order to track down BTC communications. The workflow that resulted in the development of the map reduce system was created with the goal of handling the tokenization and parsing of structured data during the map phase of the process. As a result, only 13 patterns were discovered, one of which was the Emoji data, which had previously been overlooked in previous Hadoop-related research efforts. Although the text is significant, only one piece of the BTC transaction puzzle, namely, the statement "There is no certainty, only possibilities," appears in the transaction details, regardless of its significance.

VI. Conclusion

The volume of Big Data is increasing every day. Big Data analytics has progressed substantially as a result of data science. Big Data analytics may be incredibly useful when dealing with massive amounts of data. Before a transaction can be finalized, it must be thoroughly examined. The usage of the majority of those who have invested in BTC and used Twitter to acquire information are still confused what tweets are correct or incorrect. The data storage technique used by BTC is incapable of maintaining any of the patterns that provide additional information. Previous research did not include several data types required for understanding people's views on BTC transactions, thus it's possible that information concerning BTC's history will be used in the future. When working with BTC data, a transaction pattern must be evaluated independently. This investigates how the pattern that was developed was established using Twitter data connected to BTC transactions. To find a pattern, Big Data analytics was utilized to follow BTC transaction communications in tweets. Hadoop's map reduce workflow was designed in the map step of the procedure, and the mapper components mapped thirteen sets of patterns, which were reduced by the reducer into three patterns using the attributes previously stored data in the Hadoop context, one of which is the Emoji data, which was left out in previous research discussions but is critical to unlock the BTC transaction key communication. Along with web links, the phrase "No certainty, only possibilities" and the word "patience" were prominently featured. In BTC transactional tweets, URLs and linked text were commonly used. The majority of BTC-related tweets include a link to another website. The impact of BTC tweets produced patterns that centered on continuing to wait or exercising patience in regards to BTC price and volume.

References

Abubakar, A., El-Gammal M.T. and Alarood, A.A., 2020. End-to-end fully-informed network nodes associated with 433 MHz outdoor propagation environment. *International Journal of Computing and Digital Systems*, 10, pp.1-19.

Alkatheeri, Y., Ameen, A., Isaac, O., Nusari, M., Duraisamy, B. and Khalifa, G.S., 2020. The effect of big data on the quality of decision-making in Abu Dhabi Government organisations. In: *Data Management, Analytics and Innovation*, Springer, Singapore, pp.231-248.

Blumberg, R. and Atre, S., 2003. The problem with unstructured data. *Dm Review*, 13(42-49), p.62.

Bridges, D., Pitiot, A., MacAskill, M.R. and Peirce, J.W., 2020. The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8, p.e9414.

Cappa, F., Oriani, R., Peruffo, E. and McCarthy, I., 2021. Big data for creating and capturing value in the digitalized environment: unpacking the effects of volume, variety, and veracity on firm performance. *Journal of Product Innovation Management*, 38(1), pp.49-67.

Casado, R. and Younas, M., 2015. Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience*, 27(8), pp.2078-2091.

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G. and Stefanidis, K., 2020. An overview of end-to-end entity resolution for big data. *ACM Computing Surveys*, 53(6), pp.1-42.

Dey, N., Das, H., Naik, B. and Behera, H.S., 2019. *Big Data Analytics for Intelligent Healthcare Management*, Academic Press, Cambridge, Massachusetts.

Dutta, A., Kumar, S. and Basu, M., 2020. A gated recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, 13(2), p.23.

Dwyer, G.P., 2015. The economics of Bitcoin and similar private digital currencies. *Journal of Financial Stability*, 17, p.81-91.

Gandomi, A. and Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, pp.137-144.

George, R. and Kabir, R., 2012. Heterogeneity in business groups and the corporate diversification firm performance relationship. *Journal of Business Research*, 65()3, pp.412-420.

Grover, P., Kar, A.K., Janssen, M. and Ilavarasan, P.V., 2019. Perceived usefulness, ease of use and user acceptance of blockchain technology for digital transactions insights from user-generated content on Twitter. *Enterprise Information Systems*, 13(6), pp.771-800.

Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A. and Khan, S.U., 2015. The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, pp.98-115.

Hu, H., Wen, Y., Chua, T.S. and Li, X., 2014. Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, 2, pp.652-687.

Kamps, J. and Kleinberg, B., 2018. To the moon: defining and detecting cryptocurrency pump-and-dumps. *Crime Science*, 7(1), pp.1-18.

Kaushik, S., 2021. *Bitcoin Tweets, Tweets with trending #Bitcoin and #btc hashtag.* Available from: https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets [Last accessed on 2021 May].

Kazemi, I. and Hassanzadeh, F., 2020. Modelling multivariate, overdispersed count data with correlated and non-normal heterogeneity effects. *Statistics and Operations Research Transactions*, 1, pp.335-356.

Kikuchi, S., Kitao, S. and Mikoshiba, M., 2020. Who suffers from the COVID-19 shocks? Labor market heterogeneity and welfare consequences in Japan. *Covid Economics*, 40, pp.76-114.

Krithika, D.R. and Rohini, K., 2020. Blockchain with bigdata analytics. In: *Intelligent Computing and Innovation on Data Science*, Springer, Singapore, pp.403-409.

Kumar, A., Abhishek, K., Nerurkar, P., Khosravi, M.R., Ghalib, M.R. and Shankar, A., 2021. Big data analytics to identify illegal activities on bitcoin blockchain for IoMT. *Personal and Ubiquitous Computing*, 1, pp.1-12.

Lahmiri, S. and Bekiros, S., 2020. Big data analytics using multi-fractal wavelet leaders in high-frequency Bitcoin markets. *Chaos, Solitons and Fractals*, 131, p.109472.

Lee, A.D., Li, M. and Zheng, H., 2020. Bitcoin: Speculative asset or innovative technology? *Journal of International Financial Markets*, 67, p.101209.

Lugli, E., Roederer, M. and Cossarizza, A., 2010. Data analysis in flow cytometry: The future just started. *Cytometry Part A*, 77(7), pp.705-713.

Malik, A., Burney, A. and Ahmed, F., 2020. A comparative study of unstructured data with SQL and NO-SQL database management systems. *Journal of Computer and Communications*, 8(4), pp.59-71.

Mattke, J., Maier, C., Reis, L. and Weitzel, T., 2021. Bitcoin investment: A mixed methods study of investment motivations. *European Journal of Information Systems*, 30(3), pp.261-285.

Pano, T. and Kashef, R., 2020. A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19. *Big Data and Cognitive Computing*, 4(4), p.33.

Schulze, P., Unger, B., Beattie, C. and Gugercin, S., 2018. Data-driven structured realization. *Linear Algebra and its Applications*, 537, pp.250-286.

Sean, B., 2021. *Twitter Hits 199 Million Users, Reports "Solid" Q1 Revenue.* Available from: https://www.thewrap.com/twitter-hits-199-million-users-reports-solid-q1-revenue [Last accessed on 2021 May].

Shankhdhar, A., Singh, A.K., Naugraiya, S. and Saini, P.K., 2021. Bitcoin price alert and prediction system using various models. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 1131. IOP Publishing, p.012009.

Thelwall, M., Buckley, K. and Paltoglou, G., 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2), pp.406-418.

Urrutia, A.L., González-González, C., Van Cauwelaert, E.M., Rosell, J.A., Barrios, L.G. and Benítez, M., 2020. Landscape heterogeneity of peasant-managed agricultural matrices. *Agriculture, Ecosystems and Environment*, 292, p.106797.

Vaduva, C., Iapaolo, M. and Datcu, M., 2020. A Scientific Perspective on Big Data in Earth Observation. In: *Principles of Data Science*, Springer, Cham, pp.155-188.

Yue, L., Tian, D., Chen, W., Han, X. and Yin, M., 2020. Deep learning for heterogeneous medical data analysis. *World Wide Web*, 23(5), pp.2715-2737.

Yue, X., Shu, X., Zhu, X., Du, X., Yu, Z., Papadopoulos, D. and Liu, S., 2018. Bitextract: Interactive visualization for extracting bitcoin exchange intelligence. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), pp.162-171.