International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 2, January-February 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Twitter Big Data Analysis

Hamza Pekdogan, Dr. Atilla Ergüzen

Department of Computer Engineering, Kirikkale University, Kirikkale, Turkey

ABSTRACT

Internet usage is increasing today. As a result, internet use changes their lifestyle. Many activities such as shopping, meeting, social environment take place on the internet. Thus, a large amount of data is generated. As the amount of data increased, studies were carried out to store this data. Storage is generally used to analyze data. Data analysis studies are used to realize the advertisements and investments of organizations in the right area and in the right way. Therefore, big data analysis has a place in many areas. The data analyze people personally, except to give information specific to only one area. As a result, the desired effect is created on the analyzed people. Therefore, one of the most used fields is the political field. The most used platform for those who want to express their opinions in the political field is Twitter. Therefore, my aim is to obtain big data via Twitter, to store this data, and to analyze the political approach of the person on the stored data.

KEYWORDS: *PostgreSql, Big Data, Twitter Analysis, Artificial Intelligence, Java Programing*

How to cite this paper: Hamza Pekdogan | Dr. Atilla Ergüzen "Twitter Big Data Analysis" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-2, February 2021, pp.50-53, URL: www.ijtsrd.com/papers/ijtsrd38266.pdf

IJTSRD38266

I. INTRODUCTION

The internet, which covers the whole world today, was initiated by the US army during the cold war periods with the ARPA (Advanced Research Projects Agency) project. The computer network spread across the country was designed. Later, IBM company created the BITNET system, which is considered the father of the internet. Interest in the internet increased after the end of the cold war. Thus, internet usage has gone beyond the military fields. [1]

At first, there was no internet in homes or offices. Because the internet could be used as a result of learning a complex system. Therefore, internet usage was used only in scientific studies and by engineers. It was first used in the field of electronic mail. Later Tim Barnes Lee invented the world wide web (www). As a result of this invention, the aim was to provide visual design and information sharing called "hypertext". As a result of these developments, the foundation of today's internet sites was laid. [2]

The arrival of the Internet in our country started in the 1990s. The first connection was made at Middle East Technical University in April 1993. Turkey has been used primarily in the academic field. It started to be used commercially in 1999. Bunun sonucunda As a result, a formation called TTnet emerged. With the introduction of the internet into the commercial field, the internet started to develop very rapidly in our country. [2] It is stated that the number of internet users, which was 1.785.000 in 2000, was around 2.5-3 million in 2003. There are more than 12 million users since 2005. [2] [3] In 2015, there were 2.5 billion active internet users in the world. In our country, there are 40 million social media accounts in 2015. [4]

Social media usage is increasing as a result of the increase in internet usage. Increasing use of social media creates big data. Studies are carried out on the resulting data. Within the scope of the "Data never sleeps 4.0" project prepared by Domo (2016), data on social media usage in June 2016 was published. According to these data, 400 hours of video uploads take place on the youtube video sharing site within a minute. On Google, 69.500.000 words are translated within a minute. It is stated that American users use 18,000 GB of data with mobile devices per minute. It is stated that Facebook messenger users shared 216,302 photos in a minute and made 2,430,555 likes for the images shared by Instagram users. On the Amazon website, it is stated that a sales of $ 222,283 per minute is made. On Twitter, 9,678 emoji tweets are sent within a minute. [5] Considering the usage values of social media, it is seen that there is a field where good big data can be obtained. Many industrial enterprises and areas are becoming different big data sources used when performing studies with new data generation and quantification of existing data. [6]

It is seen that people who use Twitter in particular share in text form. There are many information and valuable inferences in these text-based posts. [5] Many industrial enterprises and areas are becoming different big data sources used when performing studies with new data generation and quantification of existing data.

II. FORMULATION OF THE PROBLEM

First of all, it was aimed to obtain data with political implications as a result of data storage, data analysis and processing of the latest data after obtaining the data.

2.1. Data Storage Technologies

Data collected after data collection is usually large and this data must be stored. When data needs to be stored, both relational and no SQL database come into play. These databases are read from a specific program and the program's common routines use them. The main databases used in the big data field are [7]:

- MySql
- NoSql
- PostgreSql
- MongoDb
- Cassandra

Relational databases such as PostgreSql and MySql have basic features. It is the basic feature provided in classical relational database systems that practitioners use or are familiar with, and is an acronym for Atomicity Consistency Isolation Durability. [8]

ACID structure includes the following features [8]:

- Atomicity
- Consistency
- Isolation
- Durability

Expressed by the acronym of NoSQL BASE (Basically Available, Soft state, Eventually consistent) versus ACID operability used by relational databases [8].

- Basically Available: Copies are used to solve data access problems. These copies ensure the security of the data. In addition, data can be accessed in divided and shared form. It provides both speed and security. [9]
- Soft State: In ACID logic, data consistency is a must. This is a requirement. However, NoSQL systems allow the hosting of inconsistent and discontinuous data. [9]
- Eventually consistent: Applications are concerned with instantaneous consistency. However, it is assumed that NoSQL systems will be consistent at some point in the future. The consistency required by the ACID is guaranteed to occur at a time that is not defined in NoSQL. [9]

NoSql databases are widely used for storing big data. Since it is not relational, it works fast on big data. [7] Therefore, big data is stored unrelated. Relationships between them are provided by software developers. Some NoSql databases keep a data on more than one node when it is disconnected. As a result, the data is compatible with the parallel search process. The advantage of parallel search is fast results. Therefore, it is used in areas where speed is important. In addition, keeping data in too many nodes prevents data loss. The deleted damaged node fixes its damage by retrieving its data from the undamaged node. There are also disadvantages to keeping a data on multiple nodes. One of the disadvantages of keeping a data on multiple nodes is that it uses a lot of storage space. Necessary choices are made considering these situations. [7]

Thanks to their scalable structure, fast access to data is no longer a problem.

2.2. Data analysis

The data obtained through social media are not only digital, but also in many types such as photographs, pictures, videos, audio, text and various sizes for each. Our aim is thus to obtain meaningful and valuable information from such a large, fast and diverse data collection. The term "Big Data Analysis" is used for the methods developed for this purpose. [5]

In order for the data to enter the data analysis process, it must meet at least one of the following five criteria.: [5]

Diversity: The data should have integrity and their contents should be interchangeable and convertible. [5]

Speed: Data must be obtained quickly. Social media feeds are an example. [5]

Volume: A large amount of data is required to qualify as big data. [5]

Verification: It should be verified whether the data collected are reliable and reliable. Some pieces of data are unacceptable to be incorrect. [5]

Value: Data should share the interests and morale of the area it serves. [5]

If the data meets one of the 5 criteria above, it can be considered big data.

2.3. Data Analysis Technologies

After the collected data are stored, analysis is made on them. Some deductions are made as a result of the analysis process. The following technologies are used in this field [10]:

- Spark
- Hadoop

Nowadays, Apache Hadoop and Apache Spark technologies offer the main methods used in big data problem solving. [10]

Spark is an open source software programmed in 2009 with Scala Lang. Performs parallel computing and processing on big data. It is widely used in the analysis of big data. [11]

Hadoop is written in the Java programming language. It is developed by Apache and is the open source MapReduce framework maintained by Apache. It is a framework that will allow developing consistent, scalable and distributed projects on it. It is a software architecture that allows great advantages by processing very large data distributed. [12]

Hadoop is software that performs high level calculations on big data. It runs in parallel on multiple servers instead of running on a single server. It analyzes big data very quickly thanks to its ability to work in parallel. Parallel working structure is in the form of keeping a data in more than one node. When a data resides in more than one node, its own node performs its own searches. For this, search speed and wasted storage space are increasing. Since one data is stored in more than one node, there is no data loss in a data damage. The damage is removed from the other node where the damaged data is stored. Hadoop is advantageous in both data security and speed. However, it is disadvantageous in terms of data storage space. [11]

Generally, over 70 companies such as Yahoo, IBM, Facebook, Adobe, mainly in the industrial and academic fields, use Hadoop software.[13]

Apache Spark works in-memory, unlike Apache Hadoop technology. In this way, it provides almost 100 times improvement in performance in big data analysis. [14]

It is selected according to the requirements of the study.

After the data are analyzed, they should be interpreted. The interpretation process is done through natural language processing. Natural language processing is done by processing sound and text. It is performed in 2 ways [15]:

- Outputting text to text input.
- Convert voice input to text (speech to text) and then convert text output to audio output.

The most widely used technology in this field is NLTK and Python programming languages. Artificial intelligence technology is also frequently used because it provides the deep learning process [15].

Tensor flow is a software library used mostly with the python language. There are areas of use such as machine learning and deep learning models. It can perform numerical calculations. It is an open source software library that uses data stream graphs. It was developed by the Google Brain Team, which is part of Google's Machine Intelligence Research Organization. It works on multidimensional data arrays that communicate with each other. One of the most important features of Tensor Flow is its ability to run any process on a multi-dimensional array. The first version of Tensor Flow, which can be applied in a wide variety of areas, was published in February 2017. Apart from the contributions of Google Brain Team, it continues its rapid development with the contribution of other users. [16]

Tensor Flow uses two hardware components [15]:

- PROCESSOR; If the number of cores is low, it will run slower than the GPU. [15]
- GPU; Its core size is high, so it works fast and in parallel. [15]

When the artificial intelligence learned with deep learning is ready, the data is ready to be processed.

In deep learning, there is a structure based on learning data or representations of more than one feature level. [15]

The data is examined by artificial intelligence with its positivity or negativity and an analysis report is created. This report is then converted into statistics. As a result, data analysis is completed.

Based on the background described above, the formulation of the question is as follows:

1. How accurate is the analysis on Twitter?
2. Is it possible to use big data analytics on Twitter instead of polls?
3. Does artificial intelligence developed with deep learning give accurate results?

RESEARCH PURPOSES

The research objectives described against the background supported by the formulation of the problem are as follows:

1. Collecting accurate results from Twitter data in the political arena.
2. Analyzing the tweets people tweet.
3. To get the correct results from these tweets.

RESEARCH BENEFITS

To be able to distinguish the positive and negative results produced by artificial intelligence.

Collecting Twitter messages and tweets instantly and live.
For storing big data.
To analyze big data.

RESEARCH METHODOLOGY

In order to collect and collect data in this study; Java programming language, Twitter streaming API is used. Collected data is added to and stored in a PostgreSQL database.

Both positive and negative comments are collected from various websites and these data are used in the deep learning process of artificial intelligence.

With deep learning, an artificial intelligence is created that can distinguish positive and negative comments.

Statistics are obtained after the artificial intelligence interprets the data stored in the PostgreSQL database.

CONCLUSION

Data can be stored via the Twitter API using the Java programming language. It can be recorded in a continuous stream without losing data.

PostgreSQL can work with big data. It appears to be sufficient for this study.

Artificial intelligence manages to learn with positive and negative data. Afterwards, the positive and negative comments obtained from social media successfully conclude.

Two results return to the comments given to artificial intelligence. Returns 1 or 0. 1 means positive and 0 means negative. Comments were taken from the subjects and the results are below.

Subject Comments;

- Party x hired only its own supporters to create staffing. (0)
- Party x ignores the country's health problems. (0)
- Party x reduced municipal and unnecessary costs.(1)
- Party x did not keep any of its pre-election promises. (0)
- Party x ignores the country's defense problems. (0)
- Party x does not say anything except the national vision. (0)
- They see the head of party x as unskilled. (0)
- Party x ignores the country's unemployment problems. (0)

In this study, the data obtained from Twitter and the data created by the surveys obtain approximate values. In line with these results, it is proven that political views can be deduced from Twitter.

SUGGESTIONS

In this study, it is seen that statistically correct data and statistics on politics and people's political views can be collected successfully with big data analysis. Twitter also gives correct results in these areas. It is also observed that the in-depth learning of artificial intelligence produces successful results in distinguishing negative and positive comments.

REFERENCES

[1] BÖLÜKBAŞ K. İNTERNET KAFELER VE INTERNET BAĞIMLILIĞI ÜZERINE SOSYOLOJIK BIR ARAŞ-TIRMA: DIYARBAKIR ÖRNEĞI, YÜKSEK LISANS TEZI, DICLE ÜNIVERSITESI, SOSYAL BILIMLER ENSTITÜSÜ, SOSYOLOJI ANABILIM DALI 2003.

[2] PSIKIYATRIDE GÜNCEL YAKLAŞIMLAR-CURRENT APPROACHES IN PSYCHIATRY 2009; 1:55-67

[3] XII. "TÜRKIYE'DE İNTERNET" KONFERANSI 8-10 KASIM 2007, ANKARA

[4] İNTERNET VE SOSYAL MEDYA KULLANICILARININ İNTERNET GÜVENLIĞI VE ÇEVRIMIÇI GIZLILIK ILE İLGILI KANAATLERI VE FARKINDALIKLARI*

[5] BÜYÜK VERI ANALIZINDE YAPAY ZEKÂ VE MAKINE ÖĞRENMESI UYGULAMALAR

[6] BÜYÜK VERI: UYGULAMA ALANLARI, ANALITIĞI VE GÜVENLIK BOYUTU

[7] VERI BÜYÜKLÜKLERININ VERITABANIYÖNETIMSISTEMLERINDE MEYDANA GETIRDIĞI DEĞIŞIM: NOSQL

[8] WEB İÇERIK YÖNETIM SISTEMI TASARIMI VE GERÇEKLEŞTIRILMESI

[9] VERI BÜYÜKLÜKLERININ VERITABANI YÖNETIM SISTEMLERINDE MEYDANA GETIRDIĞI DEĞIŞIM: NOSQL

[10] APACHE SPARK TABANLI DESTEK VEKTÖR MAKİNELERİ İLE AKAN BÜYÜK VERİ SINIFLANDIRMA

[11] APACHE SPARK TABANLI DESTEK VEKTÖR MAKİNELERİ İLE AKAN BÜYÜK VERİ SINIFLANDIRMA

[12] APACHE HADOOP VE DAĞITIK SİSTEMLER ÜZERİNDEKİ ROLÜ

[13] İNTERNET: HTTP://WIKI.APACHE.ORG/HADOOP/POWEREDBY(2010)

[14] APACHE SPARK, ERIŞIM 16 EKIM, 2015, HTTP://SPARK.APACHE.ORG/

[15] DOĞAL DIL İŞLEMENATURAL LANGUAGE PROCESSING

[16] TENSORFLOW İLE SEBZE-MEYVE HALLERİNDE NESNE TAKİBİ