Research Article

Twego Trending: Data Analytics Based Search Engine Using Elasticsearch

Vedant Karmalkar[1], Kanchan Bhalerao[2], Gaurav Kaje[3], Asra Anjum[4], Smita Kasar[5]

[1,2,3,4,5]Department of Computer Science and Engineering, Maharashtra Institute of Technology, Aurangabad, MH, India

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Twitter monitoring enables firms to consider their market, stay on track of what is being said regarding their company and contenders, and uncover emerging market trends. Twego Trending is a platform where data will be viewed and structured by an automated procedure of analyzing and processing tweets data and classifying it into various hash statistics and visualizations. Implementing Twego forecasting analysis on Twitter data using various technologies may help businesses know how consumers talk about their product. Twitter has more than 340 million active users and almost 500 millions tweets are posted every day. This social media platform helps companies to reach a large audience and communicate without intermediaries with consumers. The aim is to build a Search Engine in which , when someone will type in a query , it will return back tweets as well as do data analytics on the results and provide visualizations.

Keywords: Data Analytics, ElasticSearch, Search Engine, Twitter Data

1. Introduction

Opinion is an emotion, mood, or perception and Opinion mining is studying people's sentiments. Internet is an ingenious place in regards to sentiment data. From the viewpoint of user, people can upload their own content via different social media, such as forums, blogs, or various social networks. It won't be inaccurate to imply that social media is something we're living with. Social media has been the game breaker of the modern era, whether it be advertisement, activism or globalization, data has been anticipated to increase more than ever before. In the past couple of years, more data has been produced than ever in the history of this planet. It is clear that the number of internet users is now rising from millions to billions.

Data analytics is the process of applying structured statistical techniques to identify, recap, check and analyze data. It is a multiphase process that involves collection, processing, sorting and analysis. The objective of the proposed system is to perform analysis of the enormous amount of data readily available from social media (Doshi, Nadkarni, Ajmera & Shah, 2017). Data from Twitter is special than data shared by most other social networks because it represents data that people want to share publicly.

In this Proposed System, A highly scalable Data Analytics Based Search Engine will be designed for Twitter's trending Data Sets. Data will be viewed and structured by automated processes of analyzing text data, processing it, sorting it into various statistics and visualizing it using various dashboards. The Search Engine will be powered by using Elasticsearch.

Elasticsearch is an open source search engine which works with all kinds of data. It is a NoSQL database and search engine which will be used for Storing and Searching of Data. It is known for its basic REST APIs and scalability property. Elasticsearch is the prime integrant of the ELK (Elastic Stack), a collection of softwares for data analysis, storing and visualization. It is regarded as one of the best search engines able to deal with structured and unstructured data (Bhatnagar, Suba Lakshmi & Vanmathi C, 2020). Additionally, it's an open-source product; therefore, it's always been supported and developed by programmers and engineers.

1.1. Necessity

With the rise in Internet and technology, Companies have begun to focus on Opinion Mining more than ever. Genuinely, it's their best means of knowing their clientele extensively, and their purchase behavior. Analyzing social media can allow businesses in any area to identify their consumers' demands and suggestions, which ultimately makes organizations ramp up their offerings. This framework is needed because sources such as social media produce increasingly voluminous data and managing and analyzing such big data is critical.

Figure 1. Necessity

2. Proposed Approach

A Web application will be designed to ameliorate the work of past researchers in this field, which will essentially function like a search engine. JavaScript will be used to create this web application. It will also consist of various dashboards and visualizations. Just like any other search engine, user can search for a term. Then, based on the tweets of that specific term, analytics will be provided using Dashboards.

Twitter data will be used for analysis. It will be retrieved from Developer Accounts using Twitter's official Stream API. This collected information will then go through a series of processes that will purify the data. After purification, procedures for key analysis will be followed. In our method, Python is used extensively for data processing, analysis and storage. Only the necessary tuples of a tweet will be extracted and stored (Doshi, Nadkarni, Ajmera & Shah, 2017).After the Feature Extraction is done, then the processed data will be stored in Elasticsearch.

Data storage and searching can be achieved with the assistance of Elasticsearch. It is an open source, distributed, readily-scalable, NoSQL database and enterprise-grade search engine published under the terms of the Apache License. It is based on Java and is designed to run in real time. It can search and index data of diverse formats. The Web Application that will be built on top of Elasticsearch will fetch data according to the query hit by a user (Bhatnagar, Suba Lakshmi & Vanmathi C, 2020). Visualization Dashboards will be created by using Kibana and JavaScript. Google Maps API will be used for Map Visualizations

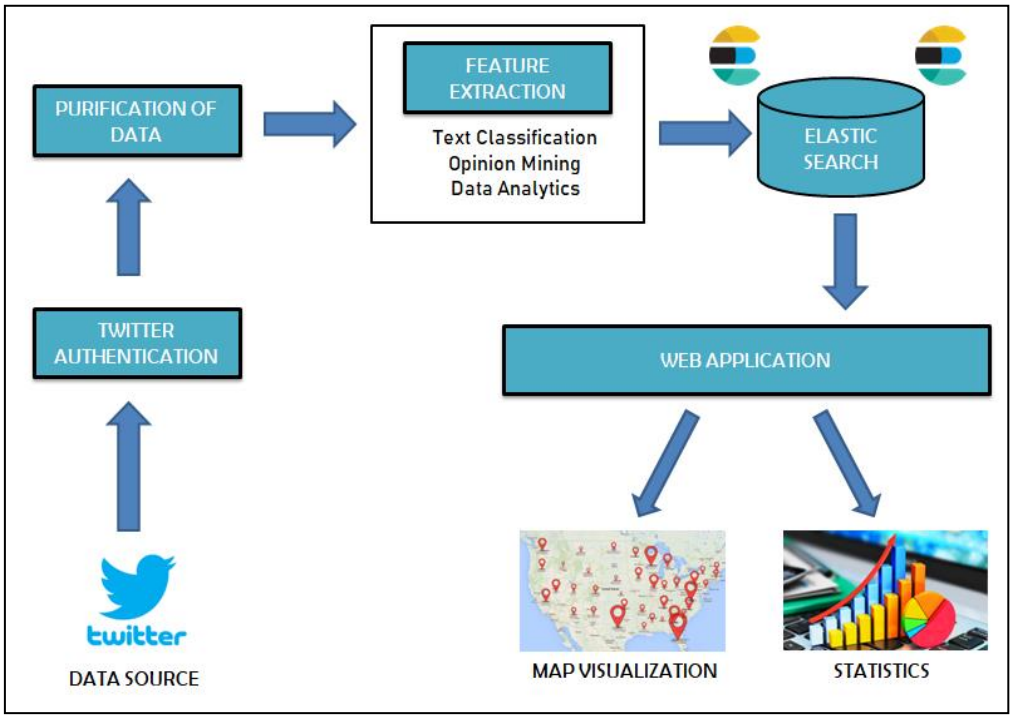

Figure 2. System Architecture

2.1. Algorithms/Techniques/Procedures Used

The System is divided into two major Functional Blocks. These blocks illustrate the working of system in detail.

2.1.1. Writer Part

Writer Block is associated with all the processes that download the Twitter data, then the process of extracting, analyzing and storing data into Elasticsearch. This is a Top-Down Process.

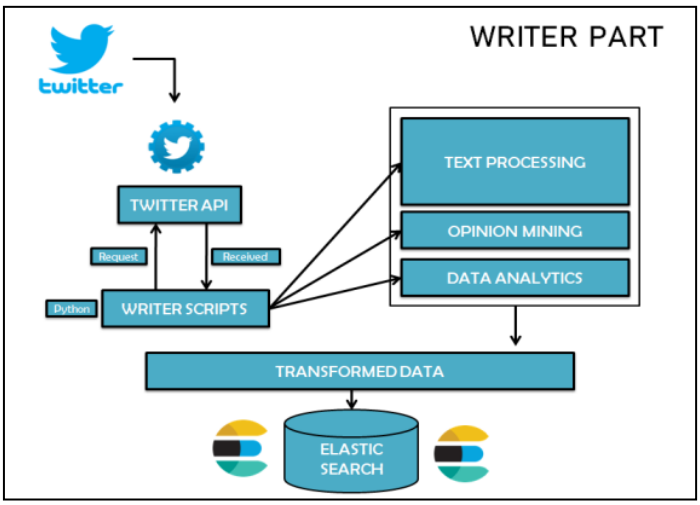

Figure 3. Writer Part

2.1.2. Reader Part

Reader Block is associated with all the processes that the Search Engine performs. This is a Bottom-Up Process. It illustrates how the system will process when user searches for a query in the Search Engine Application.

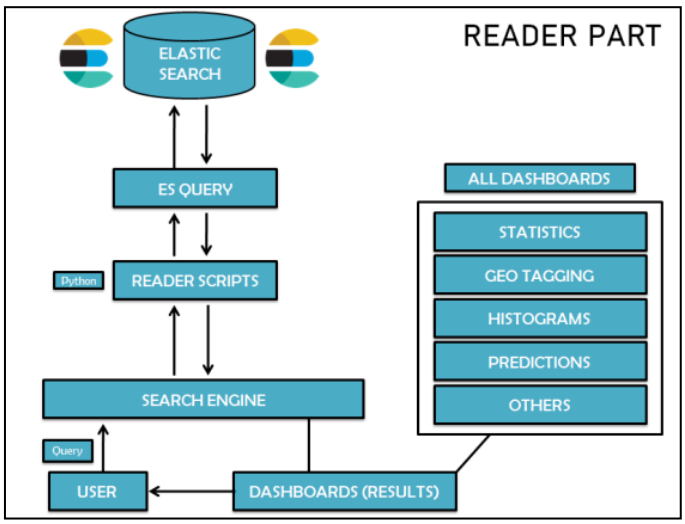

Figure 4. Reader Part

3. Process, Results and Discussion

3.1. Twitter Authentication

Creating a Twitter application using Twitter Developer Account is the first step in collecting data sets from Twitter. When an application is created, four important tokens are generated. These tokens are used for authenticating with Twitter.

3.2. Streaming Real-Time Tweets

After authentication, Data is extracted using the Twitter Stream API. To detect and gather tweets in real-time, SteamListener object is a critical component. The tweet data is in JSON format.

3.3. Python Integration

In our method, Python is used extensively for data processing and analysis. Three Major Python Libraries will be used, namely, Tweepy for Accessing the Twitter API , NLTK for performing analysis and Elasticsearch-py which provides functions for integration with Elasticsearch . For Web application support, Python Flask is used.

3.4. Data Storage

Elasticsearch will be used as the Database for this project. It is a NoSQL database and search engine which will be used for Storing and Searching of Data. The python library Elasticsearch-py is a high-level library aimed at helping to write and run queries for Elasticsearch.

3.5. Data Visualization

Data Visualization is done using two major tools i.e. Kibana and JavaScript. Kibana is an open source tool used for visualizing Elasticsearch data. Various categories of visualization dashboards such as Top Trends, Top Hashtags, Bar Charts, Pie Charts, Time Charts, Histograms, Heat Maps, Map Visualization, etc. are presented for users.

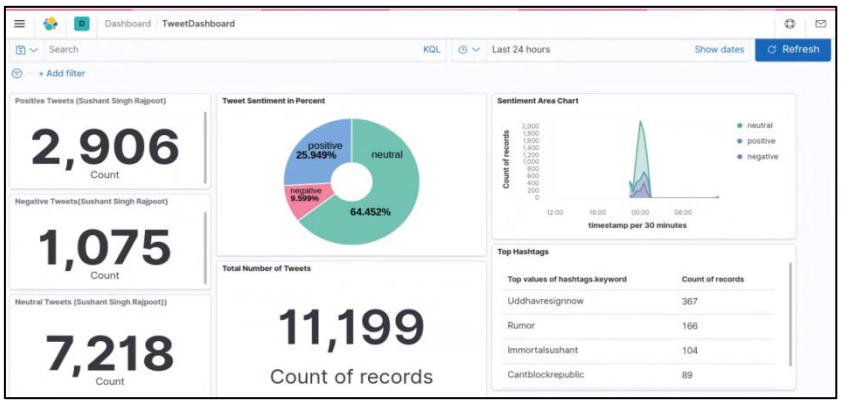

Figure 5. Dashboards

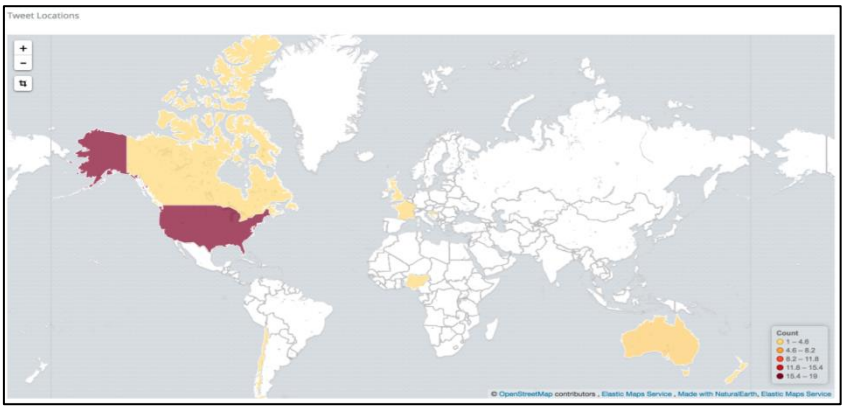

Figure 6. Geo Location Visualization

4. Real World Practices

Twego Trending can be implemented into different complex applications instantly, including:

4.1. Business Intelligence

These days, when releasing a specific product, the general practice is often to introduce it widely through a hashtag. The marketing and analytics team of the organization would then use heat map analysis to actively observe customer reactions across multiple countries. If sales do not pick up in a given country, they will zoom in to find consumer tastes and plan their next update accordingly. In such low sales areas, they can also introduce new initiatives and offers to encourage further sales.

4.2. Stock Market

In the estimation of stock market developments, analysis of Twitter data will assist. Research shows that the stock market can be hugely influenced by news stories and social media. It was noted that news of overall positive sentiment referred to a significant price rise. On the other hand, negative news is shown to be correlated to a decrease in rates, except with more prolonged ramifications.

4.3. News

Users will have the privilege by receiving live news, whether it's an incident or some other emergency. For instance, a person could use the geo visualization feature to get news of an occurring cricket match by zooming in to the precise location of the game and viewing tweets sent by the match audience. This way, geo location visualization will let a far-off event to be a virtual experience for the user.

4.4. Reviews from Websites

A huge range of reviews and feedback on almost all is accessible on the Internet today. This involves reviews of goods, opinions on political issues, suggestions on services, etc. There is therefore a need for a method for sentiment analysis that can extract feelings about a specific product or service. This will help us automate the provision of feedback or rating for the product, object, etc. provided. This will satisfy both the consumers' and the vendors' needs.

4.5. Sub Component Technology

In recommending systems, a sentiment predictor system may also be helpful. Items getting a lot of negative reviews or fewer ratings would not be recommended by the recommender system. We come across offensive language and other negative factors in online communication. These can simply be identified by recognizing a strongly negative sentiment and taking action against it accordingly.

4.6. Career Counseling

Academic establishments, teachers and mentors should use this as a way to determine which skill sets are actually emerging in a given discipline and to advise their wards appropriately to seek for better career prospects for that expertise.

5. Conclusion

A highly scalable Data Analytics based Search Engine is designed for Twitter’s trending Data Sets. Data is viewed and structured by automated processes of analyzing text data, processing it, sorting it into various statistics and visualizing it using various dashboards through Kibana and JavaScript. The Search Engine is powered by using Elasticsearch.

References

1. Zeel Doshi, Subhash Nadkarni, Kushal Ajmera, Prof. Neepa Shah (2017). Tweet Analyzer: Twitter Trend Detection and Visualization. International Conference on Computing Communication Control and Automation (ICCUBEA)
2. Dhruvi K. Zala, Ankita Gandhi (2019). A Twitter Based Opinion Mining to Perform Analysis Geographically. 3rd International Conference on Trends in Electronics and Informatics (ICOEI)
3. Divyesh Bhatnagar, R Jaya SubaLakshmi, Vanmathi C (2020). Twitter Sentiment Analysis Using Elasticsearch, Logstash, Kibana. International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)

4. Andleeb Aslam, Usman Qamar (2019). Opinion Mining Using Live Twitter Data. IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)
5. Maragatham G, Shobana Devi A (2018). Twitter Analysis Using Apache Spark and Elasticsearch. International Journal of Engineering & Technology, 7 (3.12), 314-321
6. Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde (2014). Sentiment Analysis of Twitter Data through Big Data. International Journal of Computer Science and Information Technologies, Vol. 5 (3), 3098 - 3100