STM JOURNALS

Journal of
Advanced Database
Management & Systems

ISSN: 2393-8730
Volume 11, Issue 3, 2024
September—December
DOI (Journal): 10.37591/JoADMS

https://journals.stmjournals.com/joadms

Review JoADMS

Social Media Analysis Using Big Data

Krutika Monde[1], Altaf Khan[2,*], Om Ugale[3],
Omkar Tonde[4], Shilpali Bansu[5]

Abstract

Social media has become an essential aspect of everyday life across different age groups, serving functions ranging from sharing personal updates to staying informed about social developments. Our objective is to harness big data to extract meaningful insights from the immense and ever-expanding volume of data generated on social media platforms. We will collect and analyze data from various social media sources, including text, images, and user interactions, using cutting-edge big data technologies. By employing advanced analytics like semantic word analysis and topic modeling, we aim to uncover valuable trends, user sentiments, and key influences, offering practical applications for businesses, governments, and researchers. Moreover, ethical considerations and privacy safeguards will be integral to our approach to ensure responsible data usage. By combining the extensive reach of social media with the scalability of big data analytics, this project aims to deliver a comprehensive understanding of the social media landscape, offering a wealth of actionable insights. Whether it's shaping marketing strategies, tracking public sentiment, or enhancing academic research, the results of this project will empower stakeholders to make well-informed decisions in an era where social media is a crucial aspect of our digital lives.

Keywords: Social media analysis, big data, sentiment analysis, ethical considerations, topic modeling

INTRODUCTION

The advent of the digital age ushered in an unprecedented era of data proliferation with social media platforms at its epicenter. Every day, billions of users across the globe generate an immense volume of content spanning text, images, videos, and interactions. This rich and dynamic source of information represents a goldmine of potential insights; however, it also presents a formidable challenge in terms of scale, complexity, and ethical considerations [1].

In response, our project seeks to explore the vast intersection of big data and social media analysis, aiming to unlock the invaluable knowledge concealed within the digital conversations and interactions of individuals and organizations. Social media analysis involves the examination and derivation of valuable insights from the enormous volume of data produced on social media platforms. It involves studying user-generated content, interactions, and behaviors to understand the trends, sentiments, and impact of social media on various aspects of our lives. A significant aspect of social media analysis is the abundance of the available data. Users share text, images, and videos, and engage in various forms of interaction on platforms such as Facebook, Twitter, and Instagram. These data offer valuable resources for businesses,

***Author for Correspondence**
Altaf Khan
E-mail: altafarshadkhan@acpce.ac.in

[1-4]Student, Department of Computer Engineering, A. C. Patil College of Engineering, Navi Mumbai, Maharashtra, India
[5]HoD and Assistant Professor, Department of Artificial Intelligence & Data Science, A.C. Patil College of Engineering, Navi Mumbai, Maharashtra, India

Received Date: July 01, 2024
Accepted Date: August 12, 2024
Published Date: September 11, 2024

Citation: Krutika Monde, Altaf Khan, Om Ugale, Omkar Tonde, Shilpali Bansu. Social Media Analysis Using Big Data. Journal of Advanced Database Management & Systems. 2024; 11(3): 35–46p.

researchers, and individuals. Companies can use social media analysis to understand consumer behavior, enhance their marketing strategies, and track their brand's online presence. It enables them to track customer feedback, analyze sentiments, and keep an eye on competitors. In market research, social media analysis complements traditional methods by providing real-time feedback from a diverse audience.

This aids in comprehending market trends and discovering new opportunities. In times of crisis, social media analysis can be crucial for managing and reducing reputational damages. Rapid responses to negative sentiments on social media are crucial for businesses and organizations. Social media analysis also plays a significant role in political and social movements. It can be used to gauge public sentiment regarding political issues, track the spread of fake news, and understand how social media influences public opinion. Content creators and influencers utilize social media analysis to customize their content to match audiences' preferences. They can detect trending topics and refine their growth strategy. Sentiment analysis, a core element of social media analysis, employs natural language processing and machine learning to assess whether social media posts convey positive, negative, or neutral sentiments. This is essential to understanding public opinion, product reviews, and brand reputation. Ethical considerations are critical in social media analyses. Privacy concerns, data security, and responsible data usage are important aspects that researchers and analysts must consider. Social media platforms and algorithms are continually evolving, requiring social media analysis tools and methodologies to adapt and remain up-to-date with the latest changes and trends in the field.

LITERATURE SURVEY

Data from social media have various uses in many fields [2]. Data on the Internet is increasing at a rapid pace because of structured data or unstructured data, and various MNCs use structured and unstructured data to grow companies and make profits from it. These types of data are essentially in large quantities and are considered big data. As big data cannot be processed and stored by normal computational devices and technologies, it is introduced to various big data frameworks to perform the processing and storing part; one such framework is HADOOP, which has all the big data fundamentals and uses functions such as MapReduce to perform the processing work, which can be useful for sentiment analysis research. Sentiment analysis is a popular technology today. Most studies have been conducted in this field. The following are the most popular approaches in today's world: Most research has been conducted in this area of analysis.

The increasing importance of social data in various fields and social media big data mining has gained attention from researchers in government, academia, and industry [3]. Sentiment analysis of news data is a crucial aspect of social media big data as it helps determine the emotions and reactions associated with news events. This explains that existing sentiment analysis methods are often based on sentiment dictionaries or supervised methods, which may not be scalable for analyzing vast amounts of social media data, we are moving forward than that and making a robust semantic analysis ML model that uses the social media and give out the most accurate result of the sentiment of a user according to the sentence.

Social media has transformed individuals from passive consumers of information to active content producers [4]. With the advent of Web 2.0, people have become more eager to express their opinions on various aspects of life and share their attitudes towards events, products, activities, and entities. This surge in user-generated content on social media has led to the generation of massive volumes of textual data, which require automated methods for analysis and knowledge extraction. Emotions are crucial in human life and significantly affect decision-making and social relationships. Analyzing user-generated content to recognize emotional content is essential, as emotions can provide valuable insights into a person's personality, status, and behavior and can help understand the public mood and attitude towards various events. Emotion recognition from big social data can greatly enhance our understanding of people's states and offer valuable insights into collective human behavior, applicable in domains such as product review analysis, marketing campaigns, and political stance detection.

In particular, when dealing with large-scale social media platforms such as Facebook and Twitter, interactions involve not just dyadic relations, but various individual associations [5]. The benefits of employing set theory for computational social science are outlined, including its capacity to conceptualize vagueness, deal with both categorical and dimensional constructs, analyze multivariate associations and align with most social science theories. Set theory can also effectively combine quantitative variable-centered analytical methods and qualitative case study methods.

A study conducted in China showed a big data system for collecting, storing, and retrieving public opinion event information from various sources, such as Weibo, WeChat, foreign media, and domestic media [6]. MongoDB is a NoSQL database that stores text data in real time and provides a flexible and scalable data model. Elastic Search is a distributed search engine that creates and queries the inverted index of text data and provides fast and accurate full-text search capabilities. Spring framework to develop the web interface of the system, which allows users to manage, query, and analyze public opinion event information.

PROPOSED SYSTEM

Social media has become a vital component of the modern digital world, transforming how we connect, communicate, and share information. With billions of users worldwide, platforms such as Facebook, Twitter, and Instagram have not only transformed the way we interact but have also generated an immense amount of data. This influx of user-generated content presents a goldmine of insights waiting to be uncovered, and that is where social media analysis comes into play. This unprecedented level of connectivity and the sheer volume of data generated by these interactions have given rise to an innovative and transformative field known as social media analysis.

The methodological procedure entails acquiring, scrutinizing, and deciphering information sourced from various social media platforms. This dynamic field employs a range of tools and methodologies to extract meaningful information from a vast ocean of digital conversations, posts, and interactions. It serves a multitude of purposes, from tracking trends and sentiment analysis to understanding consumer behavior and shaping marketing strategies. Social media analysis is not confined to the realm of business or academia; it profoundly influences daily lives.

It influences our news consumption, product selection, and even our perceptions of reality. By studying social media content and user interactions, we can appreciate the impact of these platforms on our personal and collective consciousness. In this age of digital interconnectedness, social media analysis is not just a tool but also a lens through which we can better comprehend the world around us and the digital footprints we all leave behind.

ARCHITECTURE

The proposed social media analysis architecture is explained in this section.

Algorithm to Perform ETL of Data from Various Social Media Platforms

This algorithm efficiently extracts, transforms, and loads (ETL) data from various social media platforms, enabling seamless integration and analysis of diverse social media datasets, as shown in Figure 1.

In this research endeavor, the first step involved the meticulous selection of three social media platforms to serve as primary data sources. The data extraction process is executed through the adept utilization of API services, web crawlers, or dedicated scrapers [7]. Following the identification and setup of the chosen platforms, an extraction mechanism was configured and implemented to seamlessly obtain pertinent data from these sources.

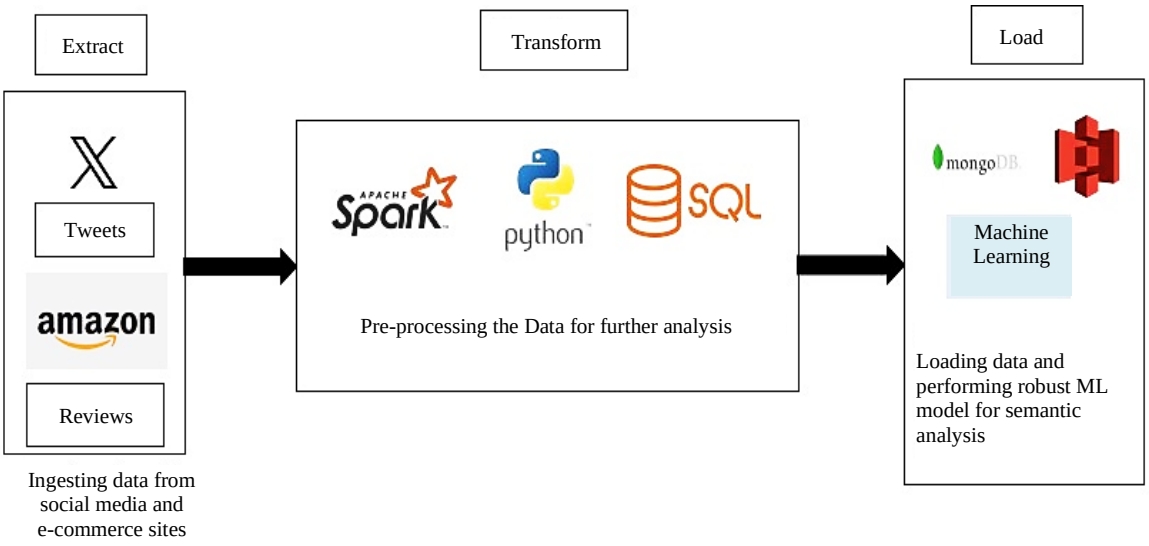

Figure 1. Architecture for social media analysis.

Upon the successful extraction of data, the second phase centers on the establishment of an adept storage system. MongoDB, a NoSQL database, was selected for its document-based architecture, which proved instrumental in efficiently storing and managing distributed data across multiple commodity servers [8]. This strategic decision is intended to improve data retrieval and boost the overall system performance.

The subsequent stage is dedicated to the meticulous pre-processing of the extracted data. This involves a comprehensive cleansing of the textual content to eliminate stop words, hyperlinks, and extraneous elements. Furthermore, a linguistic analysis was conducted to determine the root forms of words, thereby fostering a deeper understanding, and facilitating subsequent analysis. The employment of Apache Spark, a robust big data processing tool, is instrumental in executing these pre-processing tasks efficiently, especially when dealing with large-scale data.

The subsequent implementation entails the development of the PySpark code, harnessing the rich functionality of its libraries. This code serves as the backbone for data processing, ensuring a systematic and reliable approach to model creation. PySpark's capabilities are leveraged to handle intricate data processing tasks, thereby fortifying the robustness of the subsequent analysis.

In the final phase, the processed data were stored within the MongoDB cluster for seamless access and retrieval. These data, which are now refined and cleansed, are then deployed in a machine learning model tailored for semantic analysis. The model aims to deliver accurate and insightful results aligned with the intrinsic characteristics of the curated dataset. This comprehensive algorithm outlines a systematic methodology for extracting, storing, pre-processing, and analyzing social media data with a focus on semantic analysis, amalgamating the strengths of PySpark and MongoDB in this multifaceted process (Figure 2).

Algorithm to Create a Robust Semantic Analysis ML Model

Data Splitting

The preprocessed data were meticulously divided into training, validation, and testing sets. The training set forms the foundation, and the validation set acts as a crucial checkpoint to prevent overfitting during training. The held-out test set provides a final evaluation of the model's generalizability to unseen data.

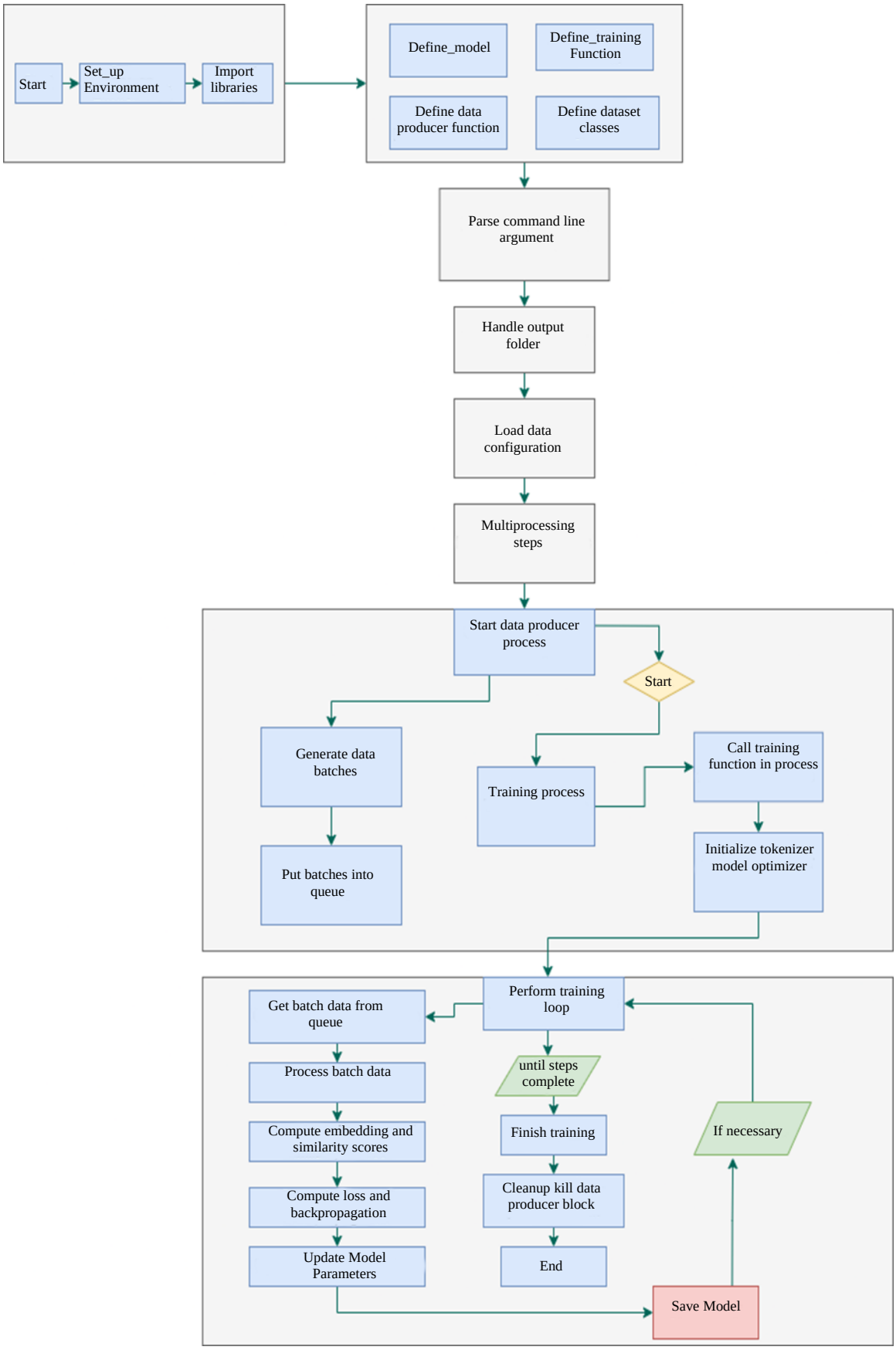

Figure 2. Data flow diagram.

SentenceTransformer Model Construction

The SentenceTransformer model is constructed by leveraging the transformer architecture, which is known for its adeptness in understanding contextual relationships in text. It extends transformer models such as BERT or RoBERTa to generate fixed-size embeddings for whole sentences rather than individual words. By training on large datasets, the model learns to produce high-quality sentence embeddings and capture nuanced semantic information. These embeddings facilitate a range of subsequent tasks including semantic similarity analysis and text classification.

Model Training

Training a sentence transformer model involves several key steps. First, a large corpus of text data is collected, which can include various sources, such as books, articles, and web pages. Next, these data are preprocessed, which typically involves tokenization, cleaning, and sometimes augmentation to increase diversity. Then, the preprocessed text is fed into the model, which consists of transformer architecture layers such as BERT or RoBERTa.

Model Evaluation

During training, the model learns to generate high-dimensional representations (embeddings) for each input sentence, thereby capturing semantic and contextual information. These embeddings were refined by minimizing a loss function that quantifies the difference between the predicted and target embeddings. This process is usually performed using large-scale computing resources and takes advantage of techniques such as mini-batch gradient descent and backpropagation. Finally, the trained model can be adjusted for specific tasks, such as sentiment analysis or semantic similarity, to tailor it to specific applications [9].

Model Tuning (Optional)

Hyperparameters such as the learning rate and number of training epochs can be adjusted based on the evaluation results to potentially improve the model's performance. This may involve iterative training, evaluation, and refinement of the model configuration to achieve optimal sentiment analysis accuracy.

RESULT

In the culmination of this research endeavor, the proposed methodology showcases a robust and systematic approach to extracting, storing, and analyzing social media data for semantic analysis, as shown in Figure 3. The amalgamation of sophisticated data extraction techniques from carefully selected platforms, coupled with the efficiency of MongoDB as a NoSQL database and the power of PySpark for large-scale data processing, contributes to a comprehensive solution for handling diverse and dynamic datasets.

In Figure 4, the presented visual representation illustrates a pie chart depicting the sentiment analysis results. The user has the flexibility to select the preferred visualization type through a drop-down menu located on the left side of the interface. This interactive feature improves the user experience by enabling dynamic exploration of sentiment distribution.

The model, highlighted in Figures 5 and 6, operates by selecting tweets relevant to a specific topic, determined by the user's input of keywords, such as airline names. The model then evaluates the sentiments conveyed in these tweets. Notably, the simplicity of the model is evident in its streamlined process of extracting user sentiments, underscoring its efficiency and ease of use.

To enhance the precision of the results further, the application incorporates various prompts that are accessible within the interface.

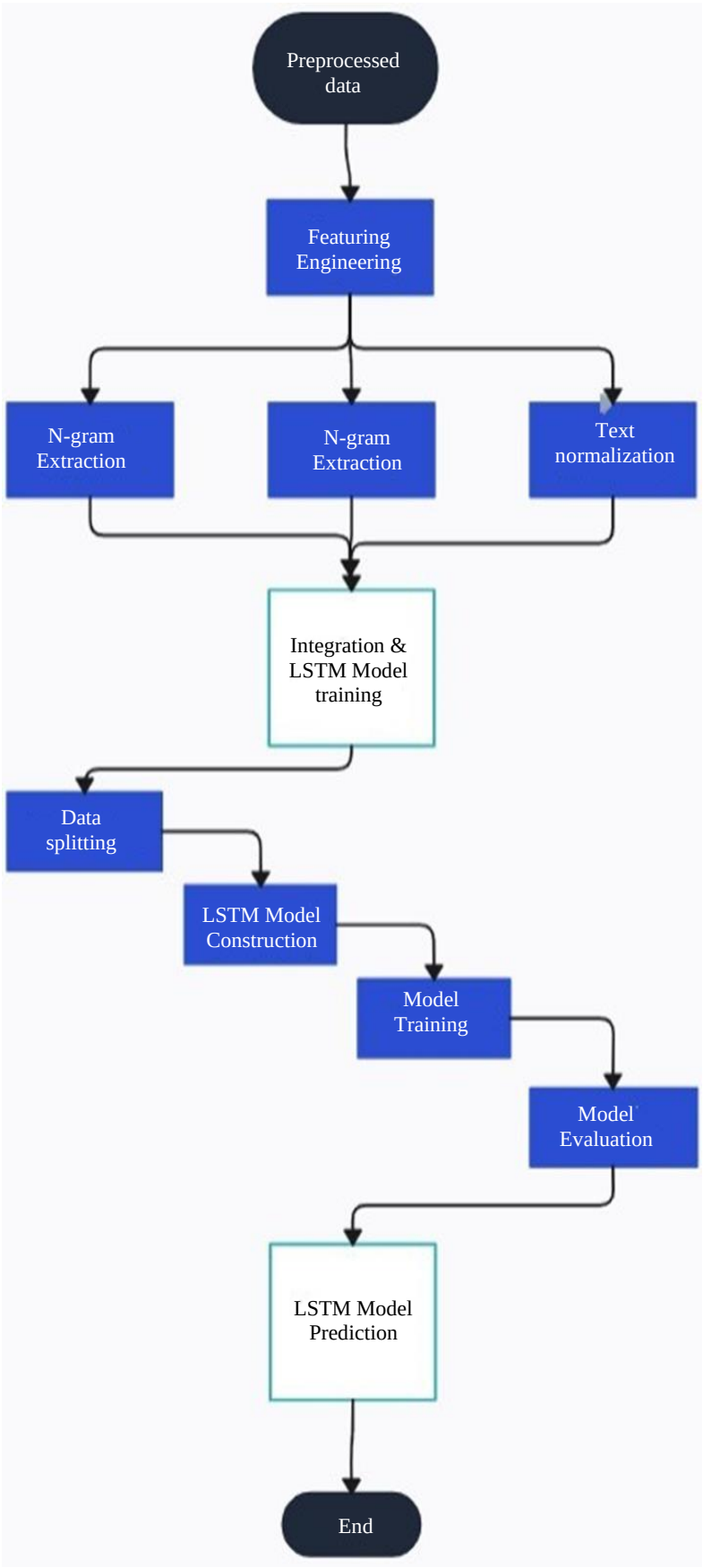

Figure 3. ML model workflow.

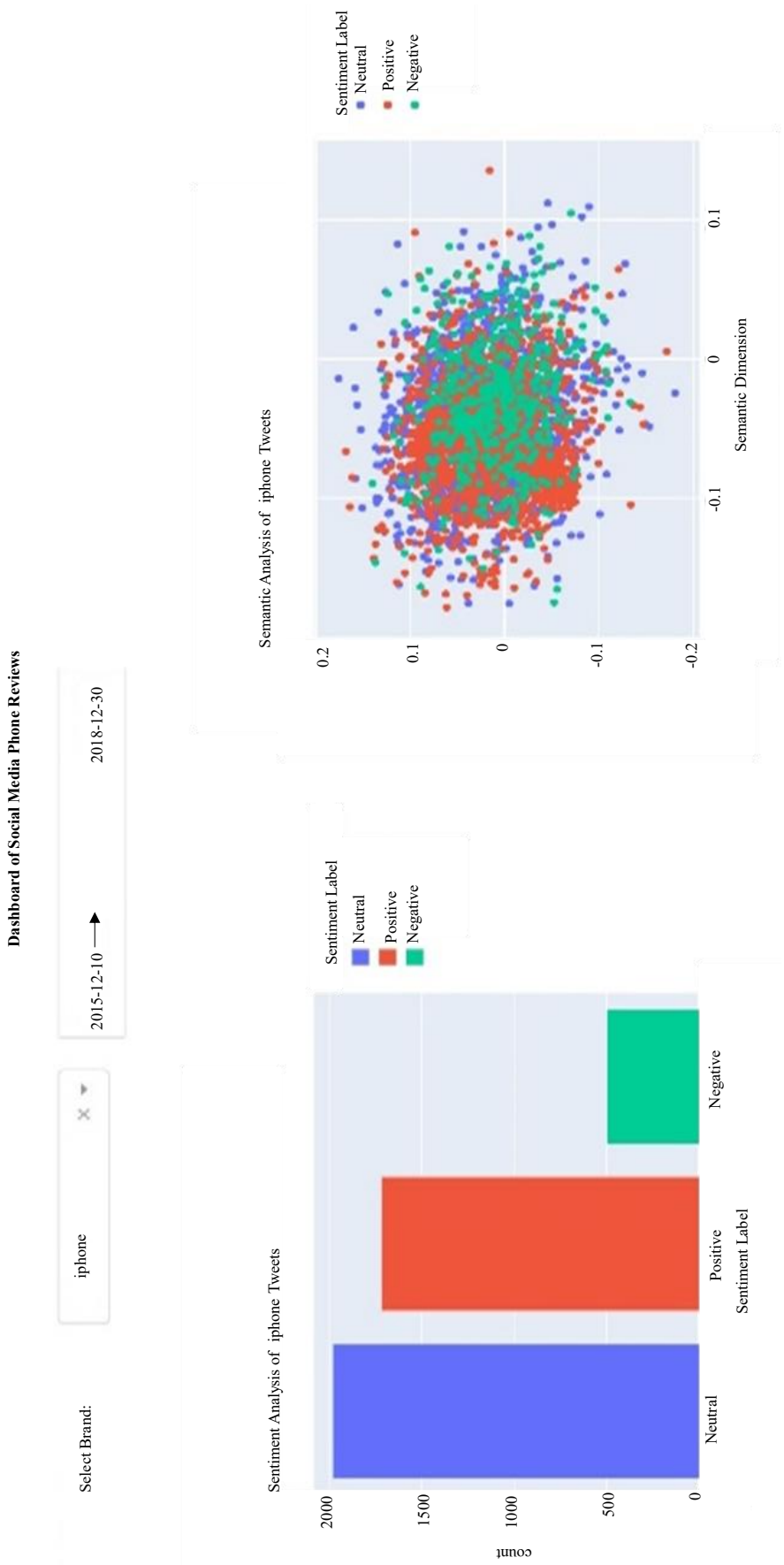

Figure 4. Dashboard.

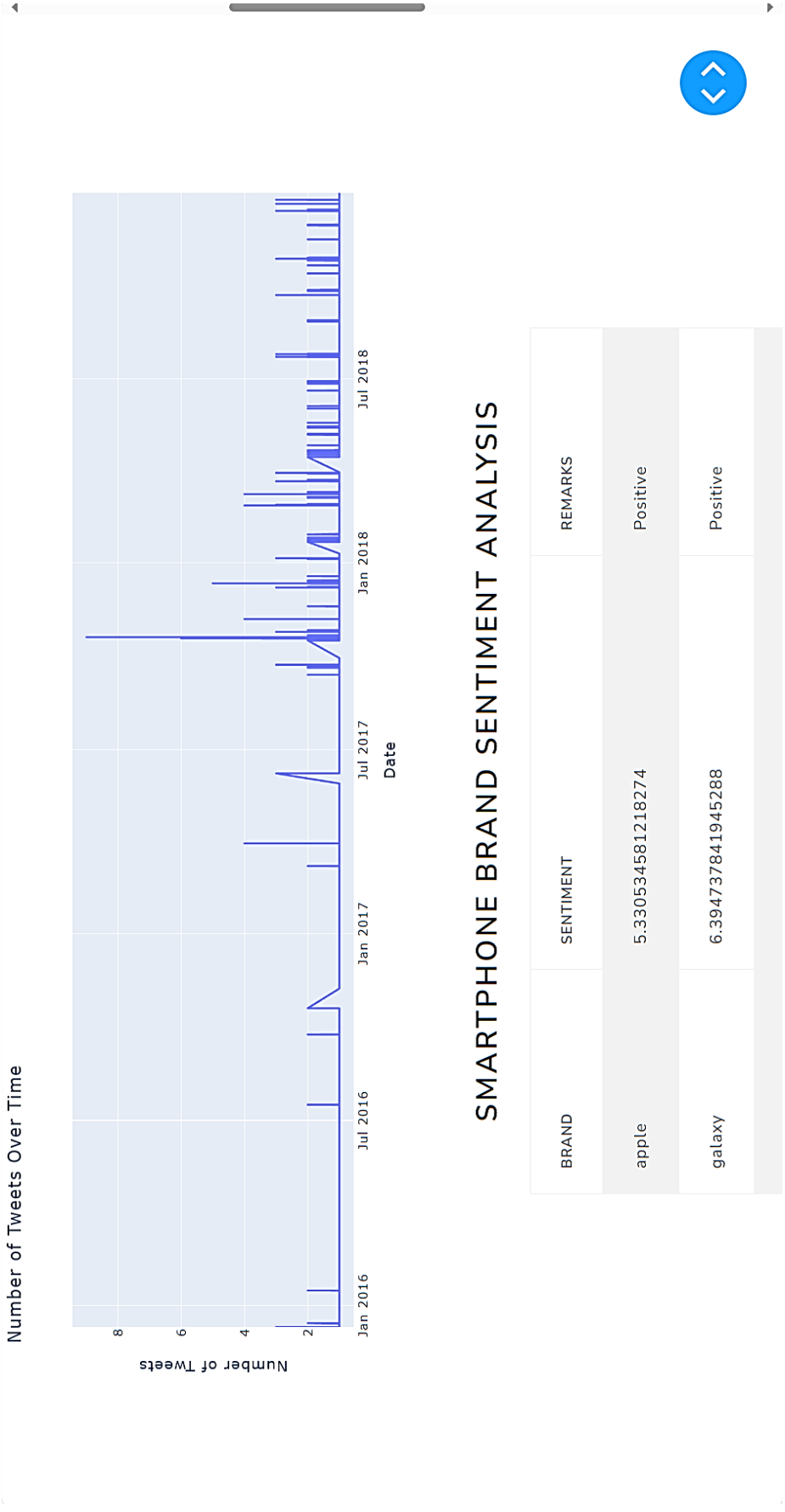

Figure 5. Sentiment analysis.

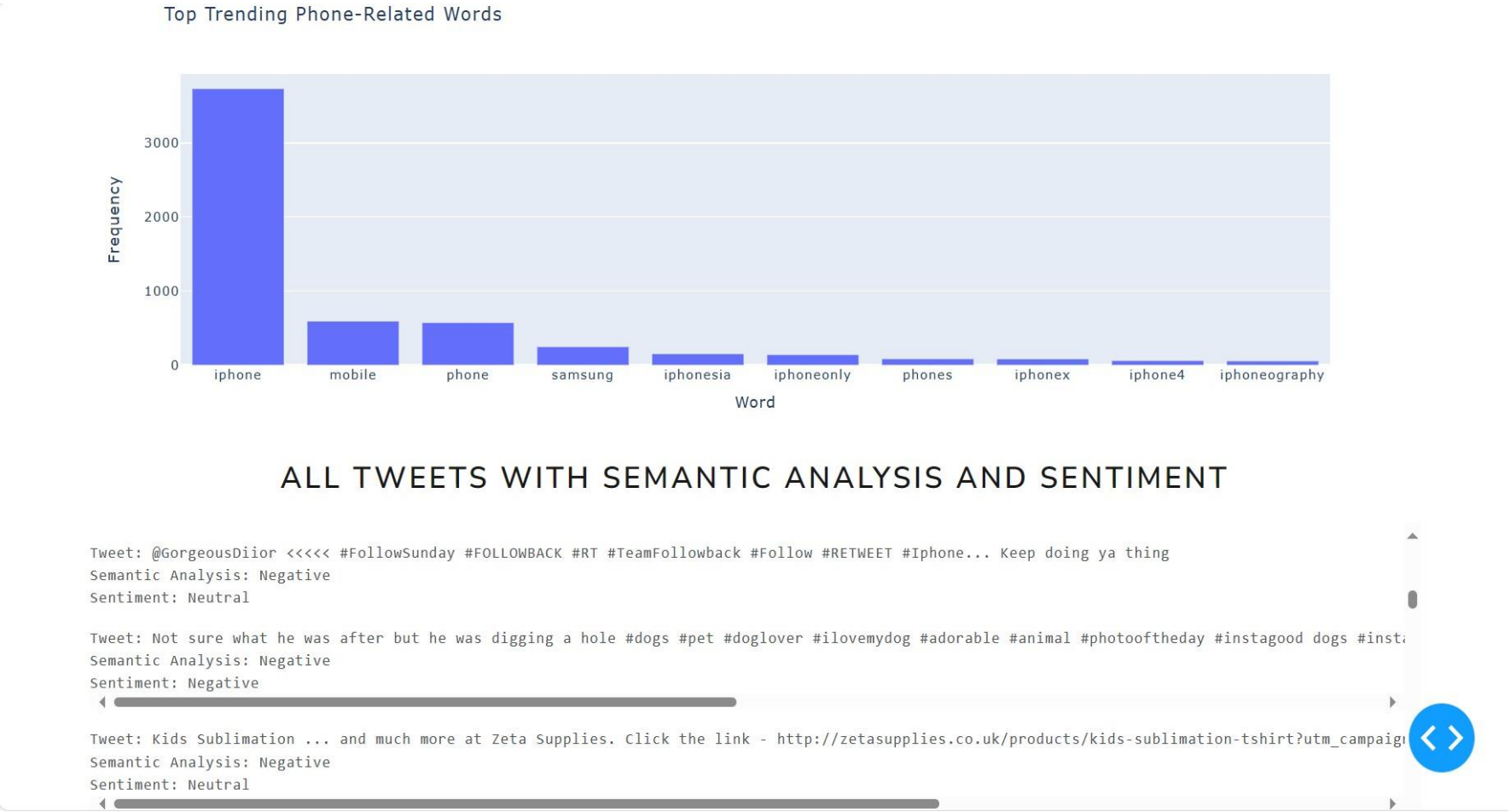

Figure 6. All tweets with semantic analysis and sentiment.

These prompts guide user interactions, providing a structured and informed approach for obtaining sentiment analysis outputs. This feature enhances the overall accuracy and reliability of the presented methodology [10].

The utilization of MongoDB as the storage backbone facilitates seamless data retrieval and access, thereby providing a robust infrastructure for the subsequent deployment of a machine learning model. This model, finely tuned to the intricacies of a curated dataset, aims to deliver nuanced and accurate results in semantic analysis, thereby contributing to the ever-evolving landscape of social media data analytics. The holistic approach presented herein establishes a framework that addresses the complexities of real-world data and underscores the significance of a well-integrated solution in advancing the field of data science and analytics [11].

CONCLUSION AND FUTURE WORK

Social media has become a vibrant tapestry woven from billions of conversations. This project approached this dynamic landscape as a symphony of signals, each post, comment, and interaction, offering a note that contributes to the broader melody of public opinion. Through the lens of big data analytics, we aimed to decode this symphony by extracting underlying trends, sentiments, and influential voices. Although the vastness of social media data presents a challenge, it also signifies an unparalleled opportunity. By harnessing the power of semantic analysis and topic modeling, we can move beyond basic sentiment analysis and delve deeper, uncovering nuanced conversations and hidden connections that shape online discourse. A deeper understanding empowers stakeholders to make informed decisions.

For businesses, these insights can be key to crafting targeted marketing campaigns and fostering stronger customer relationships. Governments can gain a real-time pulse of public sentiment, enabling them to address concerns and build trust with their constituents. Researchers can leverage these data to explore the social fabric of our digital age, shedding light on social movements, cultural trends, and the evolving dynamics of human interaction. Finally, the future of social media analysis lies in its continuous innovation. As platforms and user behavior change, our analytical tools must also be considered. By staying at the forefront of technological advancements and prioritizing ethical data practices, we can ensure that this symphony of social signals continues to be a source of valuable knowledge, shaping a more informed and connected digital landscape. The accuracy of the model cannot be determined as the data are expanding as we extract it daily and will have the final accuracy result after collecting an adequate amount of data.

REFERENCES

1. Sponder M. Social Media Analytics: Effective Tools for Building, Interpreting, and Using Metrics. McGraw Hill Professional: New York; 2011.
2. Sehgal D, Agarwal AK. Sentiment analysis of big data applications using Twitter data with the help of HADOOP framework. In: International Conference on System Modeling & Advancement in Research Trends (SMART); 2016. p. 251–5. doi: 10.1109/SYSMART.2016. 7894530.
3. Shahare FF. Sentiment analysis for the news data based on the social media. In: International Conference on Intelligent Computing and Control Systems (ICICCS); 2017. p. 1365–70. doi: 10.1109/ICCONS.2017.8250692.
4. Perikos I, Hatzilygeroudis I. A framework for analyzing big social data and modelling emotions in social media. In: IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService); 2018. p. 80–4. doi: 10.1109/BigDataService.2018.00020.
5. Flesch B, Vatrapu R, Mukkamala RR, Hussain A. Social set visualizer: a set theoretical approach to big social data analytics of real-world events. IEEE International Conference on Big Data (Big Data); 2015. p. 2418–27. doi: 10.1109/BigData.2015.7364036.

6. Hu Q, Zhang Y. An effective selecting approach for social media big data analysis—taking commercial hotspot exploration with Weibo check-in data as an example. 3rd International Conference on Big Data Analysis (ICBDA); 2018. p. 28–32. doi: 10.1109/ICBDA.2018.8367646.
7. Mitchell R. Web Scraping with Python: Collecting more Data from the Modern Web. 2nd Edition. Sebastopol, CA: O'Reilly Media, Inc.; 2018.
8. Reis J, Housley M. Fundamentals of Data Engineering. Sebastopol, CA: O'Reilly Media, Inc.; 2022.
9. Bozonier J. Test-Driven Machine Learning. Packt Publishing Ltd; 2015.
10. Liu, B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. 2nd ed. Cambridge: Cambridge University Press; 2020. DOI: https://doi.org/10.1017/9781108639286.
11. Ngaboyamahina M, Yi S. The impact of sentiment analysis on social media to assess customer satisfaction: case of Rwanda. In: 4th International Conference on Big Data Analytics (ICBDA); 2019. p. 356–9. doi: 10.1109/ICBDA.2019.8713212.

