Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

Mining Social Media Data Streams for Sentiment Analysis

Prof. Rahul Patil [1],Kaustubh Ingale[2],Vishwajeet Patil [3],Soham Puranik [4]. Swapnil Halwalkar[5]
Dept. of Computer Engineering,
Pimpri Chinchwad College of Engineering, Pune,India

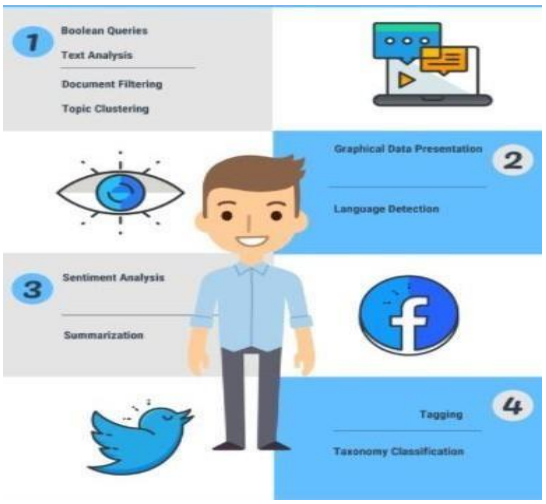
***Abstract*:-Web clickstreams, information retrieval, customer data, user evaluations, and business data, as well as network event logs, transactions, and information overflow from social networking sites, are some of the additional sources. As a result, streaming data is created, which is a continual flood of data at various rates Content analysis jobs that require organisation include emerging topic recognition, Indexing, analysing, and collecting hidden information from a vast are all instances of indexing, evaluating, or obtaining secret information from a huge dataavalanche. It's difficult to watch and analyse the valuable and important material from such a wide stream of data due to the massive volume of data flowing from numerous data sources (from social media platforms). As a result, we may collect data from various sources and employ algorithms to generate trending and relevant information.**

Keywords:- Sentiment Analysis, Apache Spark, Social Networks, Big Data, Machine Learning

INTRODUCTION:-

In terms of analysis, prediction, extract information, and opinions, social media is one of the most essential and appropriate large data sources for machine learning research.One feature of social media data, Twitter messages, for example, offer a wealth of structured information about the persons involved in the conversation. It has the potential to be successful. to lead to more precise methods for extracting semantic data.It allows for the empirical investigation of social interaction features. Governments and businesses generate massive amounts of streaming data, necessitating the use of effective data analytics and machine learning approaches to help them make predictions and choices. The continuously changing environment of new products, new markets, and new customer behaviours, on the other hand, unavoidably leads to the emergence of a concept drift problem. How to give more trustworthy datadriven predictions and decision tools in an everchanging and large data environment has become a critical topic.

OUTLINE :



The components listed below comprise the working model of our system are:

- To study different tools like Apache spark, Apache kafka, ML Algorithms(webscraper [trendseries/time. series analyses[arima model]]trend detection[outlayer detection]) which would be useful in data stream mining.
- For sentiment analysis, several data stream mining algorithms should be implemented.
- To develop system to know the trending topics on social media and to process further for positive, negative and neutral.

SENTIMENT ANALYSIS:-

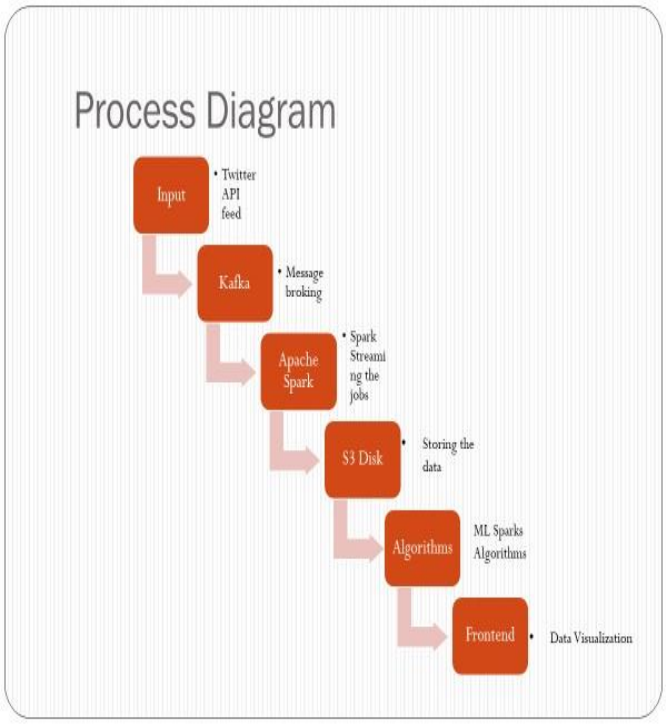
Natural language processing (NLP) is used to track social media conversations and uncover additional context about a topic, business, or theme. The goal of sentiment analysis is to categorise the polarity of a text.A text-based tweet, for example, can be classified as "good," "negative," or "neutral." A model can be trained to predict the proper sentiment given the text and related labels. Our brand passion index and net sentiment score illustrate how users feel about your brand and how it compares to your competitors. Machine learning, lexicon-based approaches, and hybrid methods are all used in sentiment analysis. Multimodal sentiment analysis, aspect based sentiment analysis, fine-grained opinion analysis, and language

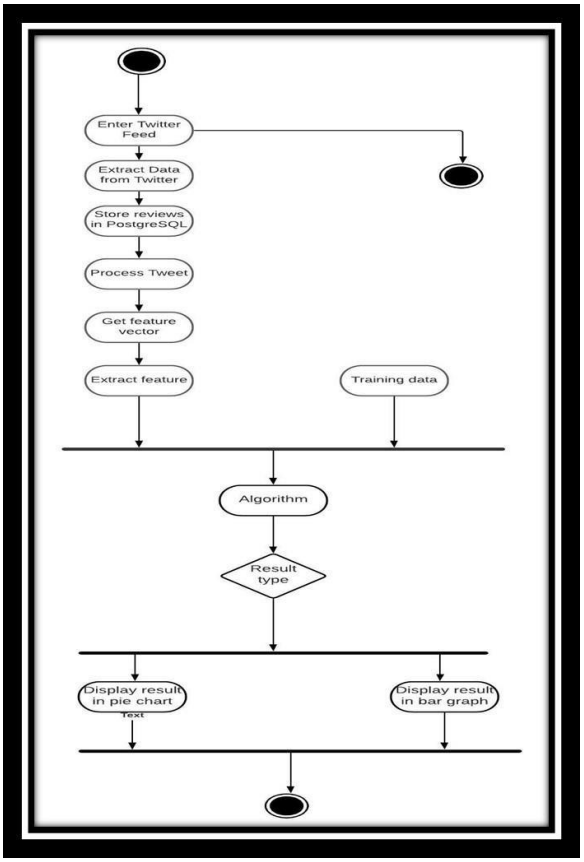
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

specific sentiment analysis are some of the subcategories of sentiment analysis study.

More than 4.5 billion individuals, or roughly 49 percent of the world's population, use social media. Users send over 550,000 Tweets and 530,000 Facebook comments every minute, and a huge portion of these messages offer valuable business insights about how customers feel about products, companies, and services. Sentiment analysis allows businesses to mine this data and extract the emotions that drive social media discussions, allowing them to have a deeper understanding of how and why people discuss a product or issue.

Because 51% of social media users have 'called out' a company on social media, brands are rethinking how they connect with their customers. To turn unpleasant situations into wonderful experiences, they must respond promptly, effectively, and individually. Restaurants, retail stores, internet providers, and airlines – which are frequently the most mentioned industries on social media – may use sentiment analysis tools to swiftly discover disgruntled consumers, prioritize answers, and categorize concerns by urgency. . Real-time sentiment analysis can even put you one step ahead of a potential public relations disaster you to intervene before a customer's negative experience becomes widely publicized.





SYSTEM DESIGNS:-

Initially, the Twitter API will be called by the Nodejs Module and passed to the Kafka broker, which is responsible for event streaming and connecting with the system. Later, batch streaming will be written for the data that is arriving and will be stored into the S3 Minio docker container. Finally, machine learning algorithms will be applied, such as Time series for predicting the analysis part, which will be stored in Postgres SQL and later shown to UI Dashboard Services or API with React framework.

Tools Descriptions:-

APACHE KAFKA:-

Apache Kafka is a streaming system that is free and open-source.Kafka is a programming language that is used to build real-time streaming data pipelines for transferring data across multiple systems or applications on a regular basis You can do the following with itLet's take a look at some of the basics and see what we can learn from them. How Kafka accomplishes these goals. Data is published by producers on issues of their choosing. Subscribers can receive messages and subscribe to topics. Now and again, we need a system that can process streams of events as they arrive, on the fly, and then take some action depending on the results of the processing, whether it's an alert or a notice that must occur in real time.

APACHE SPARK:-

Spark is a large-scale distributed computing open source project.Spark may be used to create real-time and near-real-time streaming applications that alter or react to data streams in real time or near-real time. Spark is comparable to Map

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

Reduce, but it is significantly more powerful and faster since it enables more sorts of operations than simply map and reduce, employs the Directed Acyclic Graph execution paradigm, and is almost totally memory-based. In the latest Spark release, it now supports both micro-batch and continuous processing execution modes. Spark can be used standalone or in conjunction with a number of schedulers, such as Hadoop Yarn, Apache Mesos, and Kubernetes. Batch computations, Spark Streaming and Structured Streaming, Mllib machine learning, and graph computations are all available. with GraphX are all possible with Spark SQL.

FRAMEWORK AND BACKEND PART:-

ReactJS is an open-source JavaScript library for designing single-page user interfaces. It allows us to create reusable UI components since it is declarative, fast, and adaptable. React components are difficult to reuse, and it's utilised as the foundation for single-page, massive, interactive online projects. The virtual DOM approach in React takes a long time to write and is inaccurate. In a React application, each component is in charge of rendering a small, reusable chunk of HTML. By stacking components within other components, complex applications can be created utilising simple building pieces.

Node.js is a free and open-source bridge default setting for processing JavaScript code even in a browser. It's important to keep in mind that NodeJS isn't a programming language or framework.. The majority of people are baffled, supposing it to be a framework or programming language. Back-end services like APIs, Web Apps, and Mobile Apps are typically built with Node.js.

PostgressSql PostgreSQL is a free and robust object-relational database system. It has a well-known design and it has been actively developed for more than 15 years, giving it a great reputation. Reliability, integrity of data, and accuracy are all hallmarks of the company. For Linux, UNIX (AIX, BSD, HP-UX, SGI IRIX, Mac OS X), and other operating systems Macos, Solaris, or Windows, to name a few. PostgreSQL is a database management system.

LITERATURE REVIEW:-

The authors of this paper[1] A real-time big data sentiment analysis for iraqi tweets using spark streaming. The volume of data broadcast through social media platforms like Twitter is increasing at a breakneck speed. In terms of research and forecasting, extract information, and opinions, Twitter is one of the most essential and appropriate large data sources for machine learning research. People make use of The flow of Iraqi dialect has expanded in recent years, particularly via the Twitter platform. In data science research, opinion mining for numerous dialects has become a prominent challenge. In this research, we will use spark streaming to try to construct a real-time analytic model for sentiment analysis and opinion mining on Iraqi tweets. They developed a way for creating a quick analysis framework for Twitter data collection, processing, prediction, and visualisation.

The suggested method has a number of characteristics that set it apart from earlier sentiment analysis frameworks. The proposed system solves the problem of evaluating hundreds of tweets per second that arrive in the system memory. The framework uses Spark's HDFS storage to provide a solution to the large amount of data. For scalable stream data, they provide parallel data collection nodes and parallel processing nodes. The challenge of managing incoming tweets is solved via Kafka.

The authors of this paper[2] Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning. In the derived strategy, machinelearning techniques are used to analyse Twitter data from the 2014 World Cup soccer tournament in Brazil in order to discover global sentiment. Sentiment polarity was estimated based on emotion words discovered in user tweets after filtering and analysing the data using natural language processing techniques. The dataset was normalised to be used by machine learning algorithms, and natural language processing techniques such as word tokenization, stemming and lemmatization, part-of-speech (POS)tagger, To extract emotions, the textual data of each tweet was analysed using name entity recognition (NER) and a parser.

Machine learning techniques were utilised in this study to offer a novel way to sentiment analysis on a linguistic dataset using machine learning techniques. The techniques for reducing noise, data filtering, and preparing linguistic data were demonstrated using natural language processing (NLP) methodologies.. The tweets were filtered and processed to produce more accurate data while lowering the dataset's size To find sentiment scores for the tweets, The study used tagging, stemming, tokenizing, and part-of-speech analysis. To further analyse the data, we used machine learning approach such as knn Classifier, SVM, random forest, and KNN.

[3] COVID-19, developed by the authors of the paper, was used to examine social media sentiment. In today's environment, everyone is involved in social networks on a regular basis. These days, we can use social media data to conduct a lot of research and statistics. Finally, we use Natural Language Processing and Recurrent Neural Networks to analyse Twitter users' thoughts and expressions (like, hastags, pests, tweets) based on the primary trends (in this case, the keyword 'covid' and the coronavirus issue). We analyse, integrate, visualise, and summarise statistics in this phase to prepare them for further processing. They created a model that used a recurrent neural network to predict emotions, look for word connections, and identify them as positive or negative in a vast number of tweets. They've divided the numerous books into a far more detailed emotional strength classification (weakly positive/negative, very positive/negative) than simply positive and negative extremes. This has been paired with a custom keyword-based data scraper. allowing us to use our previously learned RNN model on the scraped datasets They focused on the coronavirus and the emotional shifts and fluctuations it causes, and they discovered that the typically positive manifestation and presence on social media platforms persisted throughout the epidemic. Of course, there are both negative and positive aspects. Although there is a natural growth in negative emotion, positive sentiment has grown stronger over time.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

Sentiment Analysis Methods for Social Media Data [4], by the author. A large amount of data is generated by users of social media websites, which plays an important role in decision making. Because it is impossible to read the entire text, sentiment analysis simplifies the process by assigning polarity to the content and categorizing it into positive and negative categories. Different algorithms can be used to do classification tasks, each with a different level of accuracy. The goal of the survey is to give participants an overview of the various methodologies for sentiment analysis. The review also included a comparison of several sentimental analysis approaches as well as their performance evaluation. This review paper focuses on the fundamentals of sentiment analysis and categorization methods. The numerous sentiment analysis methodologies, as well as their performance characteristics, were investigated in a systematic review. A lot of work has been done in recent years to uncover semantic relationships using word embedding methods and categorization using artificial neural networks. Because similar words usually express the same polarity, the semantic relationship must be checked. The researchers compare their suggested work using SVM with naive Bayes as a reference model. These two algorithms achieve high accuracy using feature selection methodologies.

The author in this paper [5], The use of social media data and sentiment analysis in election forecasting is becoming more common. This research offers and examines several platforms study on the power of various aspects, emotions, and Methodologies for predicting key judgments from online social media using social networks. According to the majority of studies, the election results can be anticipated using Twitter. Streaming APIs are used in a wide range of applications. data gathering via investigation. The majority of the time, researchers have used sentimentoriented data from social media to forecast election results and assess political campaign positions. This comprehensive study provides an overview of sentiment analysis and related methods. In addition, the article showcased some cutting-edge research on sentiment analysis utilising deep learning and word embedding approaches.

Challenges and Answers for Handling Real-Time Big Data Streams [6], by the author: A Systematic. Because a number of businesses are preparing to obtain a competitive advantage, real-time data warehousing (DWH) and big data streaming have become commonplace. In terms of real-time stream processing, data warehousing is empowered by the capacity to organise huge data in an efficient manner to reach a business conclusion. This study gives a comprehensive assessment of the literature on real-time stream processing systems, ; developments and challenges in real-time stream processing systems and can be used as a guide for the implementation of real-time stream processing frameworks for all types of data streams. Their research identifies the most important publication channels for real-time stream processing research in the real-time DWH sector and other big data applications, including: (IoT, Social media, Google etc). In addition, their literature includes implementation issues, established approaches/toolsfor real-time data integration in all of the service categories listed, as well as assessment data.

In this study [7], Streaming Big Data Analytics- Current Status, Challenges and Connection of unbounded data Processing platforms. Big data analytics is an approach for investigating vast dimensions of organised, unstructured, and semistructured data sources. Data generated continuously from a variety of data sources such as Internet-of-Things (IoT) devices, mobile applications, Embedded Sensors, web clicks, and many others must be stored, processed, and analysed in a very short period of time in order to extract meaningful insights and make appropriate decisions as soon as the need arises. Analyzing streaming big data (continuous flow or unbounded data) is, nevertheless, a difficult task. Continuous data streams have become a requirement for a wide range of commercial and scientific applications; yet, the present HadoopMapReduce technology is not suitable for massive data stream processing. The problems and benefits of streaming big data, as well as its architecture, are examined in this research. on the various open source streaming processing platforms that are available to handle large amounts of data quickly.

The author in this paper[8], Using data minig techniques to extract key factors in mobile live streaming. Smartphones are becoming more powerful, and cellular networks are offering broadband access, thus expect a rise in live streaming data traffic for mobile devices in the future years. On the other hand, despite the fact that video streaming works and is a very popular Internet application. It is usual to discover criticism of the classic client-server approach in the literature, such as that the Internet was not designed to support multimedia applications or that mobile networks are not appropriate for streaming. This study uses the association rule, a data mining approach, to examine popular live streaming sessions in order to determine what elements may influence the broadcasts.

The author in this paper[9], A Study on Sentiment Analysis Techniques of Twitter Data. It relies on vector of displacement of object to adjust rectangle to encompass accuarately the tracked object.It compares 2 set of images groups before and after change of object trajectory and detects a remarkable change if theta is high.Once a person is find suspicious the behavious detection by predicting the intention of person takes place.

The author of this paper[10] used NLP and a supervised KNN classification system to analyse sentiment on twitter tweets concerning COVID-19 vaccinations. In this study, they conduct a sentiment analysis of tweets posted on Twitter in the second half of 2008. They assess the sentiment of each tweet using an updated version of the Profile of Mood States (POMS), then connect the results to a timeline of major events that occurred during that time period. They discovered that considerable, albeit delayed, shifts in public mood levels across a range of mood dimensions are linked to social, political, cultural, and economic events.

In this study, the author[11] Modeling Public Mood and Emotion: Twitter Sentiment and SocioEconomic Phenomena: data preparation, model training, and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

inference. The framework is made up of two neural networks: CNN and RNN (Recurrent Neural Network). The CNN algorithm is used to extract high-level features from images in order to minimise the input's complexity. For categorization, RNN is utilised, which is ideally suited for video stream processing. The suggested system makes use of a VGG-16 (Visual Geometry GroupOn the ImageNet dataset, a pre-trained model was created. The algorithm is now being trained to predict behaviour based on the movie. In the footage utilised to enhance the monitoring process, the model can predict suspicious or typical human behaviour.

Reference	Consensus used	Advantages
[1] A real-time big data sentiment analysis for iraqi tweets using spark streaming	Framework used as Spark's HDFS storage to provide a solution to the large amount of data.	This method ensures that all locations in close proximity are grouped together.
[2] Sentiment Analysis on Twitter Data of World Cup Soccer Tournament Using Machine Learning	Natural language processing techniques are used to filter and analyse the data.	For your convenience, social networking platforms offer a variety of language options.
[3] Social media sentiment analysis based on COVID-19	RNN.	Big data analytics and data science are becoming the research focal point in industries and academia
[4] Sentiment Analysis Techniques for Social Media Data: A Review	SVM with naive Bayes as a reference model	By analyzing these reviews one can extract the information about their area and can do improvement.
[5]The emergence of social media data and sentiment analysis in election prediction	Streaming APIs are commonly used in studies for data collecting	There isn't a single multilingual model that can analyse emotions.
[6] Challenges and Solutions for Processing Real-Time Big Data Stream: A Systematic	DWH	It shows good precision in robust brightness due to brightness guided technique
[7] Streaming Big Data Analytics- Current Status, Challenges and Connection of unbounded data Processing platforms	(IoT)	This research gives a thorough examination of massive data stream analysis.
[8] Using data minig techniques to extract key factors in mobile live streaming	Traditional client-server model.	Video streaming is a popular Internet application that works.
[9]A Study on Sentiment Analysis Techniques of Twitter Data	It is possible to detect suspicious behaviour by predicting a person's intention.	By predicting the intention of person takes place.
[10]. Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm.	NLP and supervised KNN classification algorithm	Changes in public mood levels across a variety of mood dimensions.
[11] Modeling Public Mood and Emotion:Twitter Sentiment and Socio-Economic Phenomena:	CNN and RNN(Recurrent Neural),	In the footage utilised to enhance the monitoring process.

CONCLUSION:-

One of the most important aspects of text analysis is sentiment analysis.The amount of data being streamed through social media sites like Twitter is growing at an exponential rate. In terms of analysis, prediction, extract information, and opinions, Twitter is one of the most essential and appropriate large data sources for machine learning research. People utilise the Twitter network on a regular basis to express themselves, which is a basic fact. In three basic phases of data feeding, data processing, and data visualisation, a framework was created to acquire, filter, and mine streams of data. The fluctuations in the effectiveness of sentiment analysis algorithms when numerous features are evaluated is an interesting subject for further research. To put it another way, integrating diverse characteristics improved performance in most cases but resulted in subpar performance in others. As a result, looking into the causes of these performance instability would be a fascinating line of research to pursue in the future. Another alternative is to investigate the topic of data sparsity using both batch and hybrid methods. The goal is to assess the resilience of various Twitter sentiment techniques in light of data scarcity.

REFERENCES:-

[1] Zhang Jinagyi," Design of Intelligent Hive and Intelligent Bee Farm Based on Internet of Things Technology", June 2019
[2] Ashutosh Chauhan," Apache Hive: From MapReduce to Enterprise-grade Big Data Warehousing", -July 5, 2019,
[3] Scott Wares1 ,John Isaacs1 ,Eyad Elyan," Data stream mining: methods and challenges for handling concept drift" SN Applied Sciences (2019).
[4] Salman_Salloum, Ruslan_Dautov, Xiaojun Chen, "Big data analytics on Apache Spark" International Journal of Data Science and Analytics volume 1, pages145–164 (2016)
[5] Frederic Stahl1,2 and Atta Badii," Building Adaptive Data Mining Models on Streaming Data in Real-Time"(2020)
[6] Streaming Data. Discovery Science, Lecture Notes in Computer Science, 2010,
[7] Alec Go, Richa Bhayani and Lei Huang. Twitter Sentiment Classification using Distant Supervision. Project Technical Report, Stanford University, 2009.
[8] Patrick Paroubek and Alexander Pak Twitter as a Sentiment Analysis and Opinion Mining Corpus In: LREC 2010: Proceedings of the International Conference on Language Resources and Evaluation.
[9] Albert Biffet and Eibe Frank. Sentiment Knowledge Discovery in Twitter

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 04, April-2022

[10] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner and Isabell M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2010
[11] Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002.
[12] Efthymios Kouloumpis, Theresa Wilson and Johanna Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of AAAI Conference on Weblogs and Social Media (ICWSM), 2011.
[13] Luciano Barbosa and Junlan Feng. Robust Sentiment Detection on Biased and Noisy Data on Twitter. 2010 International Conference on Computational Linguistics (COLING).