

Continuous Tweet Stream Analysis using Apache Spark Streaming API & Kafka

Nisha A G¹, Shivaprasad N², Nischitha C B³, B.M.Shashikala⁴, M Sujithra⁵ and Yogesh Chaba⁶

¹Student, Electronics & Communications, Engineering, JSS Science & Technology University, Mysuru
Email: nishaag9903@gmail.com

²Assistant Professor, Dept. of Electronics & Communications Engineering, Sri Jayachamarajendra College of Engineering, JSS Science & Technology University, Mysuru
Email: shivaprasad_n@sjce.ac.in

³Assistant Professor, Dept. of Computer Science, St.Joseph's First Grade College, Mysore
Email: nischitha1512@gmail.com
Mysuru

⁴Assistant Professor, Dept. of Computer Applications, Sri Jayachamarajendra College of Engineering, JSS Science & Technology University, Mysuru
Email: shashikalabm@sjce.ac.in

⁵Associate Professor, Department of Computing -Data Science, Coimbatore Institute of Technology, Coimbatore, India

⁶Professor, Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India

Abstract—Social media has become a trend nowadays people are spending more time in social media platform like Facebook, Twitter, Instagram to share their opinions, creating new interests, ideas and generate awareness. These provide a major platform for collecting information which helps us to analyze them for our needs. One such microblogging social network is twitter which allows information collaboration and quick information sharing that provides relief from rising emails. Twitter has several millions of people tweeting and expressing their perspectives on a due to its distinctively brief and straightforward expressing technique. This enormous volume of raw data from tweets can be utilised for analysis and processing based on industrial or business needs. In this technical paper, the application of tweet sentiment analysis by extracting a vast number of tweets using continuous tweet stream provided by Spark and Kafka has been done. This system fetches live tweets from Twitter based on the search criteria, process the tweets, by using text mining techniques and classify them based on different labels.

Index Terms— Twitter Sentimental Analysis, Apache kafka, Spark Streaming.

I. INTRODUCTION

Nowadays millions of people use survey [1] social media to share their ideas, opinions mainly through online platform like blog spots, review websites and different chatting applications These applications help people [2] to build communication network between other people in an easy and timely fashion and also allows them to share different kinds of information. One such application is Twitter, due to the advent of twitter, people can view opinions of different people in a global scale. With the development of these real-time networking services,

such as Twitter, collections of opinions on nearly every worldwide entity of interest have emerged. With these available data, an analysis of these data can be done and it as many applications in different domain, for example business product can get feedback from user reviews. Business company depends mainly upon the user generated opinions or reviews over online to a greater extent for decision making of their products, for example if any user want to buy a product online or prefers to use liked service from the business companies, user first looks up its online reviews compare it with the competitors and discuss about it before purchasing or before making any decision. The amount of data generated by users is very vast for a human to analyse. So, there is a need to fetch the data in streams, various sentiment analysis techniques are widely used. Opinion and sentimental mining have been studied on this aspect and various approaches and research fields [3] have been discussed. As twitter may provide an amazing platform for opinion creation and expression, it poses newer challenges the procedure is ineffective without correct equipment for scrutinizing those opinions. Recently, various research projects have been conducted that utilises sentiment analysis to find out the public's perspectives on political matters on Twitter [2].

Due to the increase of hateful and negative communications and opinions over social networking sites like Instagram, Facebook and Twitter, Government of India tried to mitigate concerns over censorship of these sites, where web users continued to express against any proposed restriction on posting of content. This paper follows the following outline, Section I deals with the introduction of the project. Section II discuss the perusal of data and its analysis. Section III deals with the software's that are made use to obtain the results. Section IV deals with the methodology. Section V deals with the results and discusses about it. Finally, Section VI concludes our paper.

II. PERUSAL OF DATA

A. Data Description

Twitter being a social media network allows individual bearing an account to post messages in real time. It also provides a service called the microblogging [4] that ensures the restriction to words in the messages that is up to 150 words approximately. These messages are called the tweets that expresses meanings crisply. It must be noted that there can be mistakes due to this nature of microblogging . These tweets also give the privilege of their visibility to only the users list ensuring the privacy security. The user can tweet to a group of people who are also having the similar interests via the hashtags. Hashtags are the terms associated with tweets that mark the highlights of the topic. Some other terms related to tweets are the Emoticons and the target. In this paper the data that is used for analysis are real time streaming data's that targets the user comments classification based on the sentiment analysis without limiting it to language made during Streaming API [5]. It can be a Hashtag or the Emoticons. Emoticons are the pictorial facial expression of the human behavior that can't be expressed in words. Preliminarily the words are restricted to ensure the visibility of the tweets which is achieved by labeling them

B. Sentiment Analysis

Analyzing the factual data present based on the objective approach is a textual information evaluation technique the main drawback of such a approach is that it fails with respect to the subjective analysis thus came the sentiment analysis [6]. This sentimental analysis can be defined as the process of evaluating the current state of mind of a user based on the subjective analysis of the opinions, attitude etc. of the user. Twitter having the limitation of usage of words, user will be more prone to the use of short forms which includes the pictorial representations and the text information's. Among the terminology that tweets encounter Emoticons and hashtags play a major role in this analysis. This paper focusses on subjective classification of the real streaming data from the server as positive or negative. Table 1 shows some examples of the positive and negative expressions. Emoticons in the absence of prosody serves to gain the receivers noticing temper of the sender, that briefs out impact of interpretation on verbal communication being changed. For example, “ :(“represents that the current state of mind of user on object is disappointed or sad or pity thus the short messages are of great use to draw out the feelings out of emoticons to analyze the state of mind.

TABLE I. SAMPLES FEATURES USED TO CLASSIFY

Type	Positive	Negative
Expressions	:) :-) :o :) :-3	:(:-(:/ :o[
Words	Love, lol, good, great, thankyou	Shit, hell, hate, enemy

C. Related Work

A few works have been previously discussed on sentimental analysis. Sentimental analysis comes into the form of opinion mining, where opinion mining comes to a broader area of NLP(natural language processing), computational linguistics and emotions expressed as texts these are combinedly specified as sentiments [7]. Sentiment analysis and opinion mining as has various kind of applications which includes technology, entertainment, education, politics etc. [8] . Twitter being a popular online blogging service which allows user to share curtailed information known as tweets, this online platform as a character limitation of 140 [9,10]. User tweets based on their mood, expression and topics relating to their daily lives. So, based on all these, twitter has become a platform for public opinions and reviews on certain issues [11]. With more number of people using twitter millions of messages are generated, delivered, this platform has become a valuable asset for many industries, organization to examine the value of their product or company by scrutinizing sentiments from the user tweets about their product, brand and about competitors[12]. Sentiment analysis can be done on tweets which can provide meaningful indicators for different useful purposes[13]. On analyzing tweets, they can be classified into neutral, positive and negative words [14]. In paper [15] similar work has been carried out to classify the tweets as neutral, positive and negative. The most widely read newspapers and magazines' tweets were acquired in order to create a corpus of objective posts. They have classified based on naïve bayes. Many others have used twitter to detect other social and physical event[16]. In paper [17] researchers were able to detect earthquakes using twitter feed. In paper [18] researchers were able to identify current news topics. However, they were only able to detect such calamities or current news only hours later and in a delayed way, none of the above can detect targeted event in seconds.

III. TOOLS FOR SENTIMENT ANALYSIS

The sentiment analyses for a real time streaming data can achieved with the use of the following tools as discussed below to establish the server connections and fetch the tweets and then classify them. The following tools are been chosen among the all because of its reliability, robustness and accuracy.

A. Twitter API

Twitter streaming API provides access to all tweets which are published on twitter [19]. More than 400 million SUMs are shared every day by Twitter, which is the most popular microblogging service online today, with more than 600 million active users [20]. On an average about 5,000 tweets per second are posted on twitter and usually developer/subscribers will get a small portion of it. The streaming API is one of the two types of twitter API's. The other API is REST API which is more suitable for singular searches such as user profile information, historic tweets or posting tweets. The streaming API is little limited to real- time tweets whereas the REST API gives historical tweets up to a week or more. To start with the twitter API users, need to have developer account and obtain credentials. Real time data in tweets are sampled and filtered, these filtered tweets are evaluated approximately to the predicted emotion and are returned as public tweets. This match is performed based on the keyword or the user id respectively to have the expected accuracy.

B. Apache Kafka

Apache Data may be transferred with low latency and high throughput using the open source distributed streaming platform known as Kafka [23]. One of the better tools for reading data that is continuously flowing is Kafka [21]. Figure 1 illustrates how Kafka operates using a publish-subscribe strategy, with many desirable properties such as horizontal scalability and fault tolerance[22]. Whenever there is message there are grouped into topics, when messages are consumed, they are removed from Kafka as shown in figure 1.

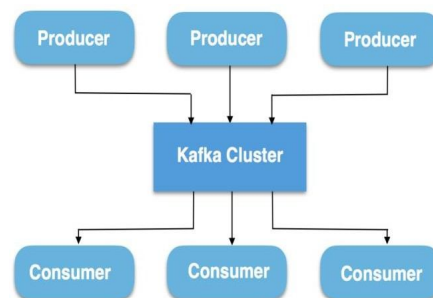


Figure 1: Kafka Publish Subscribe Model

C. Spark Streaming API

Spark streaming an extension of core Spark API which enables scalable, adaptable, high-throughput, fault-tolerant stream processing of live data streams [24]. Data can be inputted from different sources like Kafka, Flume or TCP sockets and can be processed using complex algorithms as shown in figure 2.



Figure 2: Spark Streaming Architecture

Initially, Spark receives live input data streams and divides data into batches as shown in figure 3, these batches of data are then processed by spark engine to generate final stream of results in batches.



Figure 3: Batches of Data

IV. METHODOLOGY

Initially to fetch the tweets we need twitter developer account, after enrolling for developer account, twitter will give four important credentials which allows us to consume information.

1. Consumer Key (API Key)
2. Consumer Secret (API Secret)
3. Access Token
4. Access Token Secret

This twitter account provides an individual token for authentication in the user privacy settings. this information is made use for the accessing the real time streaming tweets [25]. All twitter streaming data is processed in near real time implemented in kafka. It is important to notice that to successfully connect to the twitter Streaming API one must have a valid Oauth tokens. Kafka is tool that is used to read the Streaming API data [26], Kafka is basically a publish subscriber model which is predominantly used in real time architectures which creates topic that consists of the publisher data [27]. Usage of kafka is an added benefit as it enables horizontal scaling as stream size vary over time.

Twitter requests are packaged in a kafka publisher that contains a topic, in our case only specific twitter search terms. Publisher sends topic to broker that partition topics and send this message to subscribers [28]. Here Spark Streaming is our subscriber. The processing layer will be implemented in Apache Spark as it provides high computational performance in distributed framework [29].

The core data processing of transformation ,aggregation and sentiment analysis will be done in this layer [30]. This Spark Streaming API divides the topic into batches of data, this also converts the unstructured data into structured format mainly the data coming from twitter will be in a JSON format. The database schema consists two entities: twitter_raw_json. The twitter_raw_json table is used to store unaltered Twitter JSON messages captured by the data injection layer [31]. Now the analysis is carried on these raw JSON file. In order to create a model to classify Twitter message by sentiment, a sentiment corpus is required. The dataset should include training data and the assigned sentiment labels as shown the procedure in figure 5. Below are some examples: Positive = Achievable, Adore, Eager, Fortune, Healthy etc. Negative= Abolish, Bias, Bully, Condemn, Deny, Enslave etc. Apart from this our approach also supports emotions embedded in messages. For instance, :) indicates positive emotions, and :(indicates negative emotions. # Define positive and negative emoticons positive_emoticons = [":)", ":-)", ":)", ":D", "=)"] ; negative_emoticons = [":(", ":-(", ":("];

we have created a sentiment corpus based on different topics.

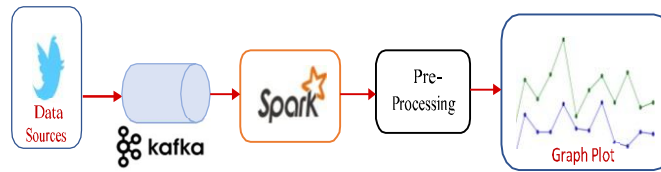


Figure 5 : Architectural Flow

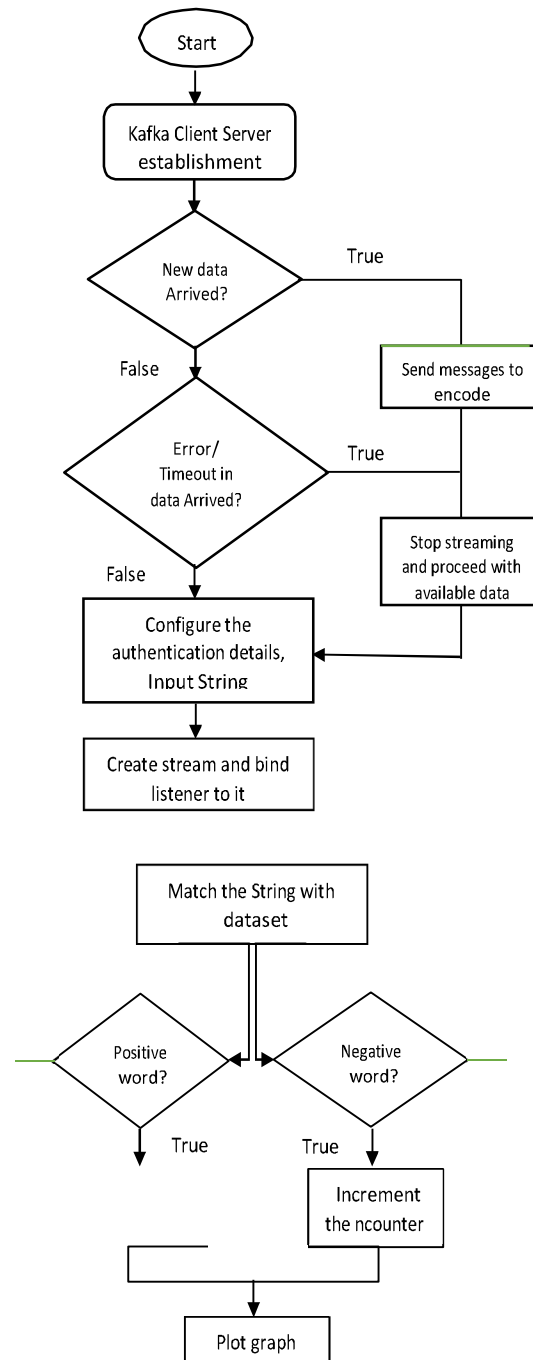


Figure 4: Flowchart of Twitter Sentiment Analysis

For the purpose of this application. We use the following key words.

Specify key words keywords = ["Election", "Narendra Modi", "Lok Sabha Elections", "Rahul Gandhi", "Chowkidar", "Prime minister election", "politics"]. We use the Tweepy API to create a simple Python program to collect and classify Twitter message by emotion based on the following emoticons. The analysis program is written in Python and is designed to run within the Apache Spark framework. Despite Spark's superior in-memory and parallel processing capabilities, it is substantially quicker than MapReduce. Additionally, it integrates with Hadoop and includes built-in support for data streaming [32]. The NLTK (Natural Language Toolkit) library, a toolkit for processing natural language, is used to construct the machine learning model [33]. With the help of the library, the implementation of a Naive Bayes model is greatly simplified. This simulation is done using the python language [34]. Detailed flow of running sentimental analysis is shown in figure 4 and steps to it are listed below.

Step 1. Initially the process begins with gaining Oauth tokens from twitter developer account this data will be included in our python script, after gaining access to twitter stream API the connection will be established between kafka and twitter stream API.

Step 2. After the connection establishment, check whether there are messages coming from the kafka topic.

Step 3. If there are any messages, encode the messages, if there are any time out in arrival of messages, continue with the available ones.

Step 4. Authentication and filtering of words are done during this step and create stream and bind listener to it.

Step 5. In this step the publisher will be twitter streaming API and subscriber will be the Spark Streaming API, which handles the data streams in the form of batches with an interval of 10 seconds

Step 6. Batches of data streams are now analyzed based on python NLTK libraries. The messages are now examined by comparing with the available corpus and conclude whether the messages weigh positive or negative contexts.

V. RESULTS AND DISCUSSIONS

The results of execution of sentimental analysis using Kafka and Spark Streaming is shown in the figure 6.

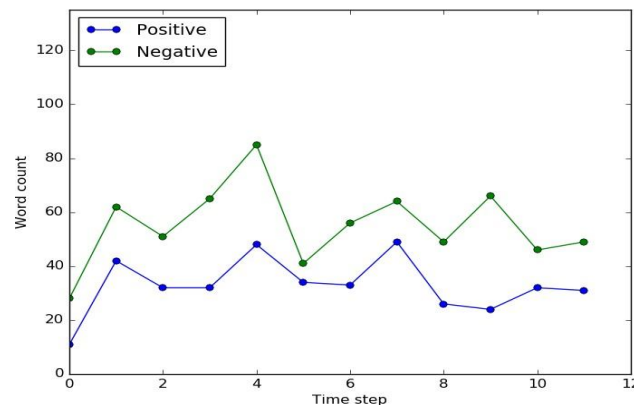


Figure 6: Sample Graph of Twitter Data

The graph indicates the number of positive and negative word counts along the time step. The main sentimental analysis of data happens from our python code which include NLTK libraries and the datasets to judge whether the words are positive and negative.

We have considered a case study on UBER and implemented that using our model, Uber which is a pioneer multinational transportation network company, uber being operational in many cities worldwide and handling a huge database of users, uber grabs a lot of suggestions, feedbacks, complaints, requests, issues and emergency needs. Nowadays social media is more preferred platform to raise/requests such issues.

The huge amount of approaching data is difficult to handle, analyze, categorize and to generate insights will be more challenging. We scrutinized conversation happening about UBER on twitter regarding payment issues, safety concerns, transaction issues and service. We took data from the latest tweets mentioning UBER as show in figure 7. From figure 8, graph indicates the number of positive and negative word counts along the time step.

The analysis is carried out for the above-mentioned tweet. Based on the tweets our model analyses the sentiments with the use of NLTK libraries and the python code, further these can be automated to raise tickets and further help customer to solve the issues.

@Uber_Support driver never ended my trip & I was charged \$113 for a 5 min ride. Map follows her across the state!! How do I fix this?!



Figure 7 : Tweets Mentioning UBER Support

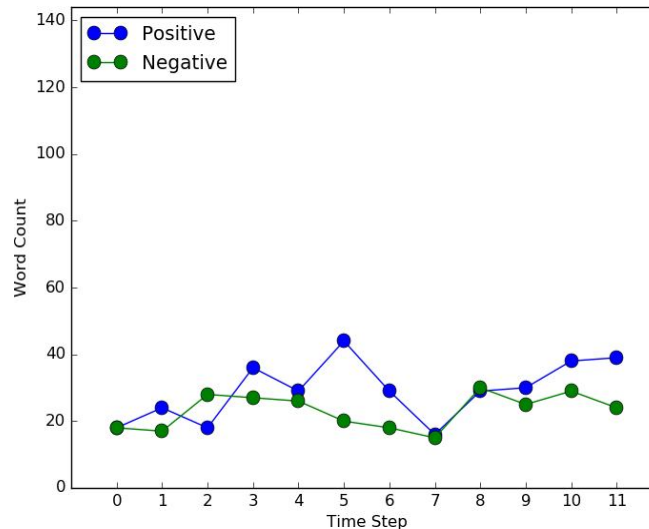


Figure 8 : Sample Graph of UBER's Twitter Data

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a model which analyses tweets in a real-time fashion. Sentiment analysis is a broad area for research. We have covered some of the important aspects. The major aspect of our project is use of Apache Kafka, Spark Streaming API's and use of machine learning methods, such as naïve Bayes. Here naïve Bayes showed the highest accuracy that can be learned as the baseline learning methods. However, there is a real need for fast responses from streaming system, even if the answer is not accurate. The machine learning algorithm naïve Bayes that we have used in our model can be further improved by experimenting with various other models and better feature selection. As our application was run only for a short period of time during testing, scaling was not a major issue.

However, the data that we have used for analysis will keep running beyond a foreseeable terminating point. These inputs can be taken for future work.

REFERENCES

- [1] Akshi Kumar, and Teeja Mary Sebastian, "Sentiment Analysis on Twitter," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
- [2] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, "Real-Time Detection of Traffic From Twitter Stream Analysis," Ieee Transactions On Intelligent Transportation Systems, August 2015.
- [3] Aliza Sarlan, Chayanit Nadam, Shuib Basri, "Twitter Sentiment Analysis," International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia.
- [4] Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Micro-blogging as online word of mouth branding. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, pp. 3859–3864 (2009).
- [5] M.Trupthi, Suresh Pabboju and G.Narasimha, "Sentiment Analysis On Twitter Using Streaming API", IEEE 7th International Advance Computing Conference, 2017.
- [6] A.M.Abirami and V.Gayathri, "A Survey On Sentiment Analysis Methods And Approach", IEEE Eighth International Conference on Advanced Computing (ICoAC)
- [7] T. Carpenter, and T. Way, "Tracking Sentiment Analysis through Twitter,". ACM computer survey. Villanova:

Villanova University, 2010.

- [8] D. Osimo, and F. Mureddu, "Research Challenge on Opinion Mining and Sentiment Analysis," Proceeding of the 12th conference of Fruct association, 2010, United Kingdom.
- [9] A. K. Jose, N. Bhatia, and S. Krishna, "Twitter Sentiment Analysis". National Institute of Technology Calicut, 2010.
- [10] P. Lai, "Extracting Strong Sentiment Trend from Twitter". Stanford University, 2012.
- [11] A. Pak, and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," Special Issue of International Journal of Computer Application, France: Universitede Paris-Sud, 2010.
- [12] H. Saif, Y. He, and H. Alani, "Semantic Sentiment Analysis of Twitter," Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. United Kingdom: Knowledge Media Institute, 2011.
- [13] M. Annett, and G. Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs," Conference on web search and web data mining (WSDM). University of Alberia: Department of Computing Science, 2009.
- [14] R. Prabowo, and M. Thelwall, "Sentiment Analysis: A Combined Approach," International World Wide Web Conference Committee (IW3C2), 2009. United Kingdom : University of Wolverhampton
- [15] Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data", In Proceedings of the 13th International Conference on Discovery Science, Berlin, Germany: Springer, 2010, pp. 1-15.
- [16] Siqi Zhao, Lin Zhong, Jehan Wickrama suriya and Venu Vasudevan, "Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games", Technical Report TR0620-2011, Rice University and Motorola Labs, June 2011.
- [17] Y. Qu, C. Huang, P. Zhang, and J. Zhang, "Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake," in Proc. ACM 2011 conference on Computer supported cooperative.
- [18] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand: news in tweets," in Proc. ACM SIGSPATIAL, 2009.
- [19] Rui Li, Kin Hou Lei, Ravi Khadiwala and Kevin Chen-Chuan Chang, "TEDAS: A Twitter-based Event Detection and Analysis System", IEEE 28th International Conference on Data Engineering, 2012.
- [20] J. F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," Comput. Intell., vol. 31, no. 1, pp. 132–164, 2015.
- [21] Godson Michael D'silva, Azharuddin Khan, Gaurav and Siddhesh Bari, "Real-time Processing of IoT Events with Historic data using Apache Kafka and Apache Spark with Dashing framework", 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017.
- [22] Bhole Rahul Hiranman, Chapte Viresh M and Karve Abhijeet C, "A Study of Apache Kafka in Big Data Stream Processing", International Conference on Information, Communication, Engineering and Technology (ICICET), Aug 29-31, 2018.
- [23] Rishika Shree, Tanupriya Choudhury, Subhash Chand Gupta and Praveen Kumar, "KAFKA: The Modern Platform for Data Management and Analysis in Big Data Domain", 2nd International Conference on Telecommunication and Networks (TEL-NET 2017), 2017.
- [24] Eman Shaikh, Iman Mohiuddin, Yasmeen Alufaisan and Irum Nahvi, "Apache Spark: A Big Data Processing Engine", 2nd IEEE Middle East and North Africa COMMUNICATIONS Conference (MENACOMM), 2019.
- [25] Sushree Das, Ranjan Kumar Behera, Mukesh kumar and Santanu Kumar Rath, "Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction", International Conference on Computational Intelligence and Data Science (ICCIDS 2018), 2018.
- [26] Babak Yadraniaghdam, Seyedfaraz Yasrobi and Nasseh Tabrizi, "Developing a Real-time Data Analytics Framework For Twitter Streaming Data", IEEE 6th International Congress on Big Data, 2017.
- [27] Han Wu, Zhihao Shang, Katinka Wolter, "TRAK: A Testing Tool for Studying the Reliability of Data Delivery in Apache Kafka", IEEE International Symposium on Software Reliability Engineering, 2019.
- [28] Apostolos Lazidis, Euripides G. M. Petrakis, Spyridon Chouliaras and Stelios Sotiriadis, "Open-Source Publish-Subscribe Systems: A Comparative Study", International Conference on Advanced Information Networking and Applications, 2022.
- [29] N. Ahmed, Andre L. C. Barczak, Teo Susnjak and Mohammed A. Rashid, "A comprehensive performance analysis of Apache Hadoop and Apache Spark for large scale data sets using HiBench", Ahmed et al. J Big Data (2020) 7:110 <https://doi.org/10.1186/s40537-020-00388-5>, 2020.
- [30] Vateekul, P.; Koomsubha, T. A study of sentiment analysis using deep learning techniques on Thai Twitter data. In Proceedings of the 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), Khon Kaen, Thailand, 13–15 July 2016; pp. 1–6.
- [31] Chen W, Yeo CK, Lau CT, Lee BS (2017) A study on real-time low-quality content detection on Twitter from the users' perspective. PLoS ONE 12(8): e0182487. <https://doi.org/10.1371/journal.pone.0182487>.
- [32] Bilal Akil, Ying Zhou and Uwe Rohm, "On the Usability of Hadoop MapReduce, Apache Spark & Apache Flink for Data Science", IEEE International Conference on Big Data (BIGDATA), 2017.
- [33] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat and Priyanka Badhani, "Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python", International Journal of Computer Applications (0975 – 8887) Volume 165 – No.9, May 2017.
- [34] M. Vadivukarassi, N. Puviarasan and P. Aruna, "Sentimental Analysis of Tweets Using Naive Bayes Algorithm", World Applied Sciences Journal 35 (1): 54-59, ISSN 1818-4952, 2017.