



Real-Time Sentiment Analysis of Reddit Comments Using “RoBERTa”: A Natural Language Processing Pipeline Leveraging Amazon Kinesis and “SageMaker”

Hiba J. Aleqabie^{1,2}

¹Artificial Intelligence Engineering Department, College of Information Technology Engineering,
Al-Zahraa University for Women. Karbala, Iraq.

²Computer Science Department, College of Computer Science and Information Technology,
University of Kerbala, Kerbala, Iraq.

Article's Information	Abstract
Received: 09.03.2025 Accepted: 07.02.2026 Published: 15.09.2026	With the explosion of social media platforms, a great volume of user-generated content is available, creating a need for effective ways to analyze public opinion. While this process has and continues to provide marketers and researchers with valuable insights that can be utilized, the technology of yesterday pales in comparison to the staggering amount of data now available. This study attempts to identify and overcome the above challenges by using deep learning techniques for real-time sentiment analysis of Reddit comments, facilitated by AWS Kinesis and SageMaker. Using the RoBERTa transformer model, we built a resilient NLP pipeline to stream, process, and analyze user comments. The dataset is prepared for 100,000 Reddit comments via AWS Kinesis Data Streams in the preprocessing step and sentiment classification. Module RoBERTa gave strong results for the task and reached an accuracy of 92% with 90% precision and 89% recall. The model has an f1 score of 89.5% and an ROC-AUC score of 0.95, which says — it is a good model for sentiment classification (i.e., to predict positive sentiments). Thus, our approach in taking only 4 hours to train and utilizing state-of-the-art transformer-based architectures can be applied in cloud-based infrastructures for interaction with the public sentiment and discourse at an inference runtime of just 30 milliseconds per comment.
Keywords: Amazon, AWS Kinesis, Real-time sentiment analysis, SageMaker, RoBERTa, Reddit comments.	

<http://doi.org/10.22401/ANJS.29.3.17>

*Corresponding author: hiba.jabbar@alzahraa.edu.iq



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Thousands of comments are posted by Reddit users every minute, expressing a range of viewpoints on subjects from politics to technology. However, businesses, scholars, and policymakers have a rare chance to quickly assess public mood by analyzing this enormous influx of data in real-time[1]. Furthermore, this volume and speed of information can no longer be handled by conventional data processing techniques. Thoughtful of the mood, context, and the refinements of textual information as it emerges is now essential to meet the increased need for real-time analytics, which goes beyond just keyword searches [2]. Organizations run the danger of losing out on important insights that could affect crucial decisions

if they don't use cutting-edge technologies. Rapid and effective data analysis is made possible by the combination of state-of-the-art machine learning techniques with stream-processing systems [3, 4]. There is tremendous investment in this space for good reasons, and without leveraging advanced technologies, it isn't easy to get meaningful data insights quickly [5]. Data is indispensable in associations, which are readily available in large numbers to take advantage of decisions and quick predictive abilities [6]. Stream-processing platforms and machine-learning techniques are at the foundation of speedier aggregation and processing jobs [7]. In downstream jobs such as auctions and thresholds, some models learn based on running actual systems, and this data could be provided with

low latency [8]. Sentiment analysis [9], frequently called opinion mining, is a natural language processing process for detecting the positivity, negativity, or neutrality of data [10]. Sentiment analysis is opinion mining, which analyzes people's emotions and their attitudes toward a topic [11]. The NLP pipeline speeds up effective cloud-based real-time text analytics for both batch and streaming analysis. Machine learning, whether supervised or unsupervised techniques, clustering models, NER, OIE, and novel classification, are outputs of text analytics levels at the receiver's end [12]. As a real-time monitoring device, our submission of streamlined NLP passes in the details of algorithms or from standard shards to fully optimize our NLP model. Large companies are anticipating a daily influx of data that hits their methods or, for large or central financial institutions, even an intra-day process to do frontline risk management [13]. They must release these details and report them in an audit, detailing their capabilities to managers and governments to ensure that the information is entered in a reliable, consistent, and state-of-the-art encryption system that also provides low-latency responses [14]. NLP pipelines typically make use of a series of stages or computational processes that take input data from its 'raw' form, perform one or more processing steps on it, and then (ideally) output information that is highly usable and useful [15]. Since slow systems can lag behind data in real-time applications, and affect business intelligence and highly competitive fields' decision-making Optimizing Natural Language Processing (NLP) pipelines is essential for quick text analytics [16]. Data pipeline processing must be done efficiently to lower expenses and boost throughput. With speedier semantic typing, businesses may create error-seeking tools like real-time search capabilities [17]. Saving time, money, and processing power can be achieved by encoding a global or time-sensitive data at the pipeline's front end. Optimizing now will help businesses remain relevant in the rapidly evolving tech world [13, 17]. Real-time text analytics run continuously, processing new data and producing immediate, responsive results based on what the new data states [18]. Some large companies across industries use very large-scale NLP processing to help guide them and even dictate actions. The ability to detect sentiment is particularly crucial in the Web 2.0 age when a few disgruntled souls can, in minutes, upload their gripes to a chat or blog. Machine learning models might also be embedded into a real-time pipeline to increase predictive capabilities based on user behavior on a real-time basis [19]. When implementing real-time systems, the main concerns

include how large volumes of incoming data are processed within an acceptable cloud budget, as well as how continuously low latencies are obtained [17]. In this research, creating a strong Natural Language Processing (NLP) pipeline for sentiment analysis of Reddit comments in real-time we address these difficulties. Our method uses AWS SageMaker for scalable model training and deployment, as well as AWS Kinesis for high-throughput data streaming. We therefore achieve precise and effective sentiment categorization by utilizing RoBERTa, a cutting-edge transformer-based model that can analyze massive amounts of social media data with no delay[20], showcasing its capability to analyze complex social media text effectively. This research focuses on analyzing the sentiment of Reddit comments in real-time by employing RoBERTa, an advanced transformer-based model, while utilizing AWS Kinesis for streaming data and AWS SageMaker for deploying and executing the model. The remainder of this paper is organized as follows: Section 2, Integrating AWS Kinesis and SageMaker for NLP Pipelines; Section 3 describes the state of the art of related work; Section 4 describes the proposed system architecture, including the block and hardware assignments and software implementation; Section 5 presents the main functionalities and services of each module, and Section 6 provides a discussion and final considerations. Section 7 illustrates the results and performance metrics. Finally, Section 8, the discussions and conclusions respectively.

2. Integrating Kinesis and SageMaker for NLP Pipelines

AWS Kinesis is a managed cloud service that receives, processes, and stores hundreds of gigabytes of data per second from real-time data streaming [21]. The hallmark of AWS Kinesis is its streams and shards[22]. The stream acts as a channel that allows its users to connect to high-throughput data sources consisting of several producers and consumer applications [23]. The producers send incoming data to the stream, and each incoming record is arranged as a record, which can be batched and stored in shards [24]. A shard is a satisfactory storage unit and represents a sequence of one or more records. Many records can serve as inputs to one or more applications. AWS Kinesis can batch a set of records that will be received and sent together to the stream. More interestingly, a Kinesis shard holds up to 1 MB worth of data and can be read twice. Finally, the

records inside a shard are guaranteed to be stored for up to 7 days, with the option of extending the storage duration [24]. A Kinesis stream is a central concept that allows you to continuously collect data arriving in real-time and process it in real-time [25]. Each stream is composed of one or more shards, and each shard serves as a stable, manageable unit of consumption of data records[26]. AWS Kinesis can also automatically split or merge shards as time goes on and the amount of data coming from your source grows and shrinks. This architecture distributes Kinesis stream records across many shards, so many consumers can process the data with good performance. Kinesis also supports the ability to scale and partition data across multiple consumers [27]. The two services are combined in the architecture to provide scalable analytics and real-time, low-latency data streaming [28]. The endpoint output is an extra feature, and the machine learning model is improved by this circular cycle [25]. Because of the platform's format awareness, fewer ideal record modifications are required [29]. By integrating [27], these pipelines, real-time data processing for analytics use cases is made possible, enhancing the platform for batch inferencing and mid- to longer-term analytics. This generates an analytical step that may be readily integrated into applications like access logs, chatbot suggestions, fraud detection, and real-time customization[30]. New models can be implemented more quickly with a hybrid method, which enables one-click deployment to AI-powered APIs. In the quickly evolving field of natural language processing, improved real-time responsiveness is crucial because new word embeddings [20, 31].

3. Related work

Recent studies have explored sentiment analysis techniques for social media data, particularly focusing on Reddit comments [32]. Evaluating various models on the GoEmotions dataset, finding that RoBERTa outperformed, other models in fine-grained sentiment classification[31] developed a real-time sentiment analysis system for online retailers using Reddit data, implementing a Naive Bayes classifier and Kafka streaming [33]. Applied RoBERTa and graph attention networks to analyze COVID-19 sentiments in college subreddits, revealing increased negative sentiment during the pandemic [34]. Compared VADER and RoBERTa.

However, these studies have shown that advanced models like RoBERTa are highly effective at analyzing sentiment across multiple domains. This manuscript highlights the vital role of real-time processing in enabling a better understanding of public opinion and consumer behavior.

4. Materials and methods

This study will utilize an NLP pipeline with AWS Kinesis and SageMaker to process data representing a continuous flow of information, such as social media posts, customer reviews, or chat messages. Consequently, this paper will focus on developing a system for real-time sentiment analysis of Reddit comments (Figure 1), leveraging on the Pushshift API for Data collection, AWS Kinesis for rapid flow processing, and AWS SageMaker for model training and development. The processing is initiated by collecting Reddit comments in real time through the Pushshift API, which provides access to both current and historical data. The collected comments are subsequently transmitted to AWS Kinesis, where large volumes of data are handled with low latency, permitting efficient and scalable processing of continuous data streaming. Subsequently, a preliminary preprocessing stage is applied, during which unnecessary elements, including special characters, URLs, and stop-words, are removed. Emojis are converted into textual representations to preserve emotional information. The cleaned text is then transformed into a numerical representation using the RoBERTa encoder, which is especially designed to process informational language and abbreviations commonly found in online conversations. The data was divided into three groups: 80% for training, 10% for model validation, and 10% for testing to accurately measure performance. During the training phase, RoBERTa was fine-tuned on the AWS SageMaker environment, utilizing distributed training and hybrid accuracy techniques, which helped reduce the training time to approximately two hours. The model was then quantized to improve execution speed and reduce its size, and subsequently deployed as an endpoint on SageMaker to process requests instantly. During inference, AWS Lambda processes feedback in batches of 10 items and sends them directly to the endpoint to generate sentiment predictions. The resulting classifications (positive, negative, neutral) are then stored in AWS S3 for future analysis. The system demonstrated an accuracy of 92.2% on the test set, with a response time of up to 15 milliseconds per comment and the ability to process 1000 comments per second, making it an efficient and scalable system for working in real-time applications.



Figure 1: The proposed system

5. Results

The performance results of the real-time sentiment analysis pipeline built with AWS Kinesis and SageMaker are presented in this section in Table 1. It displays the performance, efficiency, and scalability of the proposed model. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score are clarified along with real-time processing and benchmarks. Table 1. The RoBERTa-base model, after fine-tuning, reached an accuracy of 92.2% on the test dataset, along with a precision of 0.91, recall of 0.90, and an F1-score of 0.

90, which shows its ability to classify sentiments precisely. The model's ability to distinguish between positive and negative sentiments is further assisted by an ROC-AUC score of 0.94. In terms of efficiency, the system processes each comment in 15 milliseconds and handles 1000 comments per second, meeting the requirements of real-time claims. These results prove the system's reliability, speed, and scalability, making it an effective solution for analyzing large amounts of social media data in real time.

Table 1 The Proposed System Results

Metric	Training Set	Validation Set	Test Set	Real-Time Inference
Accuracy	93.8%	92.5%	92.2%	91.7%
Precision	0.93	0.91	0.91	0.90
Recall	0.92	0.90	0.90	0.89
F1-Score	0.92	0.90	0.90	0.89
ROC-AUC Score	0.95	0.94	0.94	0.93
Inference Latency	-	-	-	15 ms/comment
Data Throughput	-	-	-	1000 comments/sec
Training Time	2 hours	-	-	-
Model Size	250 MB	-	-	-

6. Discussions

The assessment results indicate that the proposed real-time assumption investigation framework achieves greater execution on assessment metrics. The fine-tuned RoBERTa-base demonstrate accomplishes 92. Accuracy on the test dataset and accurately classifies estimations in most of the cases, which is exceptionally great considering the casual and equivocal nature of social media content that regularly employs slang, mockery, and truncations. The precision of 0.91 shows that the model is highly reliable for class 1 positive predictions (i.e., it has low false positive rates). This is crucially useful in applications wherever bogus a hundred% accuracy can impose severe costs with faulty experiences. A

Recall value of 0.90 demonstrates that the model is able to capture most significant positive and negative feedback, although there is room for improvement in handling misclassified negative feedback. An F1 score of 0.90 ~ 0.95 is high enough to ensure that the overall performance, considering recall and accuracy, captures hypothesis complexity with reasonable correctness. Additionally, with an ROC-AUC score of 0.94, we can get a sense that the model does really well in distinguishing between positive and negative feedback at various points. Such discriminative power is very important to applications that need strict hypothesis testing, such as open feedback monitoring and brand reputation management. In terms of efficiency, it creates a pipeline of good

performance for real-time processing. The pipeline supplies low latency within real-time processing scenarios. It boasts a comment latency of around 15 milliseconds and can process up to 1,000 comments per minute, an ideal candidate for real-time applications that operate on rapid data streams like social media platforms. These measurements show how flexible and responsive the system is, enabling it to process high volumes of data without losing performance or execution time. These outcomes were achieved by optimization methods, including quantization and distributed training to minimize both training time and inference latency, all at a high level of precision. The ROC-AUC score of 0.94, which speaks to the discriminatory power of the model, visually demonstrates this point. This value means that the model has a large ability to separate positive and negative predictions on each of the domain

limits. This property is especially useful in use-cases like open-output monitoring and brand reputation management that require precise prediction analysis. The RoBERTa-based model used in this study achieved an accuracy of 92.2% and an F1 score of 0.90, outperforming [34] Hassan et al. In previous works (similar studies). 2024 achieved an accuracy of 85.0%. The mean inference time of 15 ms per comment shows a clear advantage of our model over the other converter-based models, which belong to [15] Klein et al. Hasan et al. [34] (40 ms) and (2021) (100 ms), using optimization strategies (quantization + deployment → AWS SageMaker [13]). It is established that it can handle 1000 comments in one second, presenting its highly scalable real-time applications [37]. The system analyzes the informal and ambiguous text of Reddit comments, as illustrated in Table 2.

Table 2: Comparison of Real-Time Sentiment Analysis Studies

Study	Model	Accuracy	F1-Score	Inference Latency	Dataset
This Study	RoBERTa-base	92.2%	0.90	15 ms/comment	Reddit Comments
Hasan et al. (2024)	VADER vs. RoBERTa	85.0% (RoBERTa)	0.83 (RoBERTa)	100 ms/comment	Amazon Reviews
Klein et al. (2021)	BERT	90.1%	0.89	40 ms/comment	COVID-19 Tweets
Biswas et al. (2022)	LSTM	88.3%	0.86	30 ms/comment	General social media

7. Conclusions and Future Works

The proposed pipeline processes and analyzes the sentiment of Reddit comments in real-time using Pushshift for data acquisition, AWS Kinesis for high-speed data streaming, and AWS SageMaker for training and deployment. The RoBERTa-based optimization model achieves 92.2% accuracy in social media sentiment classification. The evaluation results also demonstrate strong model performance, with an accuracy of 0.91, a recall of 0.90, and an F1 score of 0.90. In addition, the model achieves an ROC-AUC of 0.94, indicating excellent discriminative capability. The pipeline requires only 15 milliseconds for inference and processes up to 1000 comments per second, making it suitable for real-time applications. These findings confirm the system's capability to efficiently process large-scale social media data while providing meaningful insights into public opinion and online discussion. Future work could improve the model's ability to capture sarcasm and context-

dependent sentiment, which are common challenges in social media text. Integrating multimodal data, such as user metadata or post titles, could also provide richer insights and improve accuracy. Active learning strategies could be used to prioritize ambiguous comments for retraining. Domain-specific fine-tuning could adapt the model for specific applications such as finance or healthcare. Ethical considerations, such as data privacy and bias, should also be addressed to ensure fairness and transparency. Finally, edge computing solutions could be investigated to reduce latency and enable real-time analysis on devices with limited resources, further extending the pipeline's applicability.

Conflict of Interest: Authors declare no conflict of interest regarding this work.

Funding: No funding was received for this work.

References

- [1] Medhat, W. I.; Hassan, A.; Korashy, H.; "Sentiment Analysis Algorithms and Applications: A Survey". *Ain Shams Eng. J.*, 5(4): 1093–1113, 2014.
- [2] Khan, S. S.; Alharbi, Y.; "Sentiment analysis of movie review classifications using deep learning approaches," *Int. J. Adv. Appl. Sci.*, 11(8): 146–157, 2024.
- [3] Pham, K.; Kathala, K.C.R.; Palakurthi, S., *Reddit Sentiment Analysis on the Impact of AI Using VADER, TextBlob, and BERT*. *Procedia Comp. Sci.*, 258(4): 886–892, 2025.
- [4] Gill, S.S.; Wu, H.; Patros, P.; Ottaviani, C., Arora, P., Casamayor Pujol, V., Haunschild, D., et al., "Modern computing: Vision and challenges". *Telemat. Informatics Reports*, 13: 100116, 2024.
- [5] Lele, V., Kumari, S., White, G., "Streamlining Production: Using Big-Data's CRM & Supply Chain To Improve Efficiency In High-Speed Environments". *Int. J. Curr. Sci.*, 13(2): 2250–1770, 2023.
- [6] Guerra, P.C., Meira Jr., W., Cardie, C., "Sentiment analysis on evolving social streams: How self-report imbalances can help". In: *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 24–28 February 2014, Baeza-Yates, R., Bennett, P., Diaz, F., Cambazoglu, B.B., Castillo, C. ACM: New York, NY, USA, 2014, pp. 443–452.
- [7] Taherdoost, H., "An Overview of Trends in Information Systems: Emerging Technologies that Transform the Information Technology Industry". *Cloud Comput. Data Sci.*, 4(1): 1–16, 2022.
- [8] Sharma, S., Bvuma, S., Thakkalapelli, D., "Comparative Analysis of Data Storage Solutions for Responsive Big Data Applications". *Int. Peer Rev./Ref. Multidisc. J (EIPRMJ)*, 12 (2): 244–250, 2023.
- [9] Khan, M., Alqahtani, A., Alharbi, N., *Real-Time Sentiment Analysis on social media Streams Using Deep Learning Models*. *IEEE Access*, 11(1): 74215–74228, 2023.
- [10] Meshram, S., Balani, N., Jha, P., "Sentiment Analysis of Statement". *IOSR J Eng*, 9 (5): 46–49, 2019.
- [11] On, J., Park, H.A., Song, T.M., "Sentiment analysis of social media on childhood vaccination based on ontology". *J. Med. Internet Res.*, 21(6): e13012, 2019.
- [12] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M., "A View of Cloud Computing". *Commun. ACM*, 53(4): 50–58, 2010.
- [13] Biswas, S., Wardat, M., Rajan, H., "The Art and Practice of Data Science Pipelines: A Comprehensive Study of Data Science Pipelines in Theory, In-The-Small, and In-The-Large." *Proceedings of the 44th International Conference on Software Engineering (ICSE 2022, Pittsburgh, PA, USA, May 21–29, 2022, ACM, New York, NY, USA, pp. 1-13, 2022.*
- [14] Kolluri, S. "Automating Data Pipelines with AI for Scalable, Real-Time Process Optimization in the Cloud." *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, 10(6): 2070–2070, 2024.,
- [15] Klein, A.Z., Magge, A., O'Connor, K., Flores Amaro, J.I., Weissenbacher, D., Gonzalez Hernandez, G., "Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set". *J Med Internet Res.*, 23(1): e25314, 2021.
- [16] Orosz, G., Szabó, G., Berkecz, P., Szántó, Z., Farkas, R., "Advancing Hungarian Text Processing with HuSpaCy: Efficient and Accurate NLP Pipelines". In: *Text, Speech and Dialogue: 26th International Conference on Text, Speech and Dialogue, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Ekštejn, K., Pártl, F., Konopík, M., Eds., Springer: Cham, Switzerland, pp. 58-69, 2023.*
- [17] Stylos, N., Zwiegelaar, J., Buhalis, D., "Big data empowered agility for dynamic, volatile, and time-sensitive service industries: the case of tourism sector". *Int. J. Contemp. Hosp. Manag.*, 33(3): 1015–1036, 2021.
- [18] Tao, D., Yang, P., Feng, H., "Utilization of text mining as a big data analysis tool for food science and nutrition". *Compr. Rev. Food Sci. Food Saf.*, 19(2): 875–894, 2020.
- [19] Asudani, D.S., Nagwani, N.K., Singh, P., "Impact of Word Embedding Models on Text Analytics in Deep Learning Environment: A Review". *Artif. Intell. Rev.*, 56(9): 10345–10425, 2023.
- [20] Asudani, D.S.; Nagwani, N.K.; Singh, P.; "Impact of Word Embedding Models on Text Analytics in Deep Learning Environment: A Review". *Artif. Intell. Rev.*, 56(9): 10345–10425, 2023.
- [21] Deng, J., Liu, Y., "Research on Sentiment Analysis of Online Public Opinion Based on RoBERTa–BiLSTM–Attention Model". *Appl. Sci.*, 15(4): 2148, 2025.

-
- [22] Kulvanitchaiyanunt, A.A., "A Theoretical and Practical Exploration of Generative AI in AWS Architectures Enhancing Cloud-Based Data Processing and Intelligent Service Deployment". *Int. J. Sci. Comput. Sci. Inf. Technol. Res.*, 6(1): 13–23, 2025.
- [23] Taherdoost, H., "An Overview of Trends in Information Systems: Emerging Technologies that Transform the Information Technology Industry". *Cloud Comput. Data Sci.*, 4(1): 1–16, 2022.
- [24] Ali, H., Tanveer, U., Saeed, A., Alkahtani, H.K., Alzahrani, K.J., Akbayan, B., "Cloud-based machine learning for scalable classification of software requirements: Insights from the PROMISE dataset". *Syst. Soft Comput*, 7: 200405, 2025.
- [25] Karimi-Mamaghan, M., Mohammadi, M., Meyer, P., Karimi-Mamaghan, A.M., Talbi, E.-G., "Machine learning at the service of meta-heuristics for solving combinatorial optimization problems: A state-of-the-art". *Eur. J. Oper. Res.*, 296(2): 393–422, 2022.
- [26] Pacios, D., "Amazon Web Service–Google Cross-Cloud Platform for Real-Time Image Processing and Machine Learning using AWS Lambda and SageMaker". *Information*, 16 (5): 381, 2025.
- [27] Mishra, A., *Machine Learning in the AWS Cloud: Add Intelligence to Applications with Amazon SageMaker and Amazon Rekognition*, 1st ed., John Wiley & Sons: Hoboken, NJ, USA, 2019.
- [28] Germann, U., Liepins, R., Gosko, D., Barzdins, G., "Integrating Multiple NLP Technologies into an Open-source Platform for Multilingual Media Monitoring." *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2018.Melbourne, Australia, July 2018; Germann, U., Koehn, P., Zeman, D., Eds, *Association for Computational Linguistics: Melbourne, Australia*, 2018, pp. 47–51.2018.
- [29] Kreuzberger, D., Köhl, N., Hirschl, S., "Machine Learning Operations (MLOps): Overview, Definition, and Architecture". *IEEE Access*, 11: 31866–31879, 2023.
- [30] Daniel, F., Papadopoulos, G.A., Thiran, P., "Smarter Mobile Apps through Integrated Natural Language Processing Services." In: *Mobile Web Information Systems, MobiWIS 2013, Paphos, Cyprus, 26–28 August 2013*. Daniel, F., Papadopoulos, G. A., Thiran, P., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2013.
- [31] Saloot, M.A., Pham, D.N., "Real-time Text Stream Processing: A Dynamic and Distributed NLP Pipeline". In: *Proceedings of the 2021 International Symposium on Electrical, Electronics and Information Engineering*, New York, NY, USA, 2021, ACM International Conference Proceedings Series, pp. 575–584.
- [32] Li, S., Xie, Z., Chiu, D.K.W., Ho, K.K.W., "Sentiment Analysis and Topic Modeling Regarding Online Classes on the Reddit Platform: Educators versus Learners". *Appl. Sci.*, 13(4): 2250, 2023.
- [33] Silva, D. de O., da Silva, R.M.A., Lima, P.V. de S., Batista, J.C.P., Xavier Junior, S.F.A., "Reddit Comment Analysis: Sentiment Prediction and Topic Modeling using VADER and BERTopic". *Socioecon. Anal.*, 2(1): 130–136, 2024.
- [34] Yan, T., Liu, F. "COVID-19 sentiment analysis using college subreddit data". *PLoS One*, 17(11): e0275862, 2022.
- [35] Hasan, Z., Singh, A., Yadav, R.K., "Comparison of Sentiment Analysis using VADER and RoBERTa". *Int. J. Innov. Res. Comput. Commun. Eng.*, 12 (Special Issue): 103–109, 2024.
-