AI FOR REAL TIME SOCIAL MEDIA DATA ANALYSIS

Name: Yash Pardeshi Dept: Artificial Intelligence and Data Science, ADYPSOE, Lohegaon Pune

Roll No.: 58 Email: yashpardeshi9545@gmail.com

Guide Name: Prof. Priyanka Kute

Dept: Artificial Intelligence and Data Science, ADYPSOE, Lohegaon Pune

Email: priyankakute35@gmail.com

Abstract

Social media sites like Facebook, Instagram, and Twitter have grown exponentially, creating enormous, ongoing data streams. This data offers a wealth of information that may be used to detect disinformation, analyze new trends, and comprehend public opinion [1]. Its unstructured form and real-time nature, however, make analysis extremely difficult. Deep learning and natural language processing, two areas of artificial intelligence, offer effective methods for drawing conclusions from this data[3]. A framework for AI-powered real-time social media data analysis is presented in this study. The platform combines sophisticated NLP models like BERT and LSTM with big data tools like Apache Kafka and Spark Streaming. Low-latency trend identification, efficient spam/bot filtering, and high sentiment classification accuracy are all achieved by these models, according to results from prototype and literature research [5].

Keywords: Social Media, Artificial Intelligence, Real-Time Analytics, NLP, Sentiment Analysis, Big Data

Introduction

One of the most important routes of communication in the modern world is social media. Millions of posts are made every minute on social media sites like Facebook, Instagram, and Twitter, covering everything from politics and entertainment to healthcare and emergency situations [5][7]. Organizations and governments may learn a lot about customer behavior, public opinion, and worldwide trends from this flood of fast-moving, unstructured data. However, manual analysis is not feasible due to the overwhelming amount and complexity of this data [10].

Automatically extracting insights from massive datasets is made possible by artificial intelligence (AI), particularly machine learning (ML), deep learning (DL), and natural language processing (NLP) [1][2]. AI is capable of sentiment analysis, subject classification, disinformation detection, and real-time crisis prediction

[3][6]. AI models like LSTM, BERT, and GPT can extract context, sequence, and semantics from text, producing precise and timely results in contrast to conventional techniques that depend on batch processing and human interaction [1][8].

Literature Review

Early research on social media analytics relied on classical machine learning techniques such as Naïve Bayes and Support Vector Machines (SVM), which performed adequately but struggled with contextual understanding and slang-heavy user-generated text [16][17]. The development of word embeddings such as Word2Vec and GloVe improved feature representation but still failed to capture deeper semantic meaning [12].

The introduction of deep learning significantly advanced the field. Long Short-Term Memory (LSTM) networks were used to capture sequential dependencies in tweets and posts, achieving better sentiment classification than traditional ML approaches [18]. Mahmud H. [8] showed that hybrid models combining CNN and LSTM performed well in real-time topic detection from Twitter streams.

A major breakthrough came with the release of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. [1]. BERT leverages transformer encoders and bidirectional attention to understand context in both directions, outperforming all prior models on multiple NLP tasks, including social media sentiment analysis [9]. Recent studies also demonstrate that when integrated with big data tools such as Apache Spark Streaming, BERT can perform real-time classification with latency under two seconds [3].

Researchers have also focused on misinformation detection. Zimbra et al. [4] conducted a comprehensive review of Twitter sentiment analysis and highlighted the challenges in detecting fake news and bias. Hasan et al. [18] presented approaches for real-time event detection from Twitter, emphasizing the need for scalable and accurate systems.

Although progress is significant, challenges remain. Most prior studies either conduct offline batch processing or face computational limitations when applied to real-time, high-volume streams [6][11]. Handling multilingual and multimodal (text, image, video) data continues to be an open research problem [12].

Methodology

The proposed solution for AI-driven real-time social media analysis integrates **big data frameworks, NLP models, and visualization techniques** into a continuous pipeline. The methodology is divided into several stages:

A. Data Collection:

APIs: Social media networks like Facebook Graph API, Instagram API, and Twitter API v2 [19] offer access through APIs. These make it possible to gather metadata, mentions, posts, hashtags, and comments.

Streaming Tools: Apache Kafka and Apache Flume are used for intake in order to handle massive amounts of real-time data. High-velocity streams can benefit from Kafka's distributed, fault-tolerant message queues [3].

Filtering: To cut down on noise in the original information, only pertinent data is retrieved using location tags, hashtags, or keyword-based filters.

B. Data Preprocessing:

Raw social media data is frequently multilingual, noisy, and unstructured. High-quality inputs for modeling are guaranteed by the preprocessing step. Among the steps are:

Text Cleaning: Eliminating stop words, URLs, punctuation, and repetitive letters is known as text cleaning [6].

Tokenization: Using NLP libraries such as NLTK and SpaCy, sentences are divided into words or subwords.

Normalization: Normalization is the process of reducing words to their most basic form by using lemmatization and stemming.

Emoji/Hashtag Processing: The processing of emojis and hashtags involves mapping them to sentiment labels or keywords, such as " ," which indicates a positive sentiment.

Language Detection: Posts on social media frequently use multiple languages. When necessary, machine translation into English is performed after the language has been identified using tools such as LangDetect [12].

Managing Bots and Spam: Automated posts and duplicate spam tweets are eliminated using heuristic and anomaly-based filters [18].

C. Feature Extraction:

Transforming unstructured text into structured numerical representations is crucial for machine learning models.

- **Traditional Approaches:** Bag-of-Words (BoW) and TF-IDF are used as baselines for feature extraction [16].
- **Word Embeddings:** Word2Vec and GloVe capture semantic similarity and word context in a dense vector space [12].
- **Contextual Embeddings:** Transformer-based models like BERT and RoBERTa provide contextual embeddings, understanding words based on surrounding text [1].
- **Hashtag & Emoji Features:** Hashtags and emojis are encoded as additional features, improving performance in sentiment classification.

D. Real-Time Processing Framework:

- **Apache Spark Streaming:** Processes data in near real-time micro-batches, achieving low latency (<2s) [3].
- **Apache Flink:** Alternative framework for continuous, event-driven stream processing.
- **Parallelization:** Data is partitioned across nodes to scale horizontally, handling millions of posts per hour.

E. Visualization & Decision Support:

- **Dashboards:** Real-time analytics are visualized using Kibana, Power BI, or Tableau, enabling easy interpretation of results.
- **Use Cases:** Decision-makers can track trending hashtags, brand

mentions, or crisis-related keywords in real time.

- **Alert Systems:** Threshold-based triggers can generate alerts when unusual activity (e.g., sudden spike in negative sentiment) is detected [7].

Results and Discussion

According to the research, BERT-based models achieve superior accuracy (>85%) in Twitter sentiment analysis than traditional machine learning techniques [3]. With processing latencies of less than two seconds, Spark Streaming pipelines exhibit performance that is almost real-time [1]. CNN+LSTM hybrids improve event detection accuracy by 10–15%, according to Mahmud H. [8].

Using the Sentiment140 dataset, a prototype experiment showed an 87% sentiment categorization accuracy [9]. Within 10 seconds of the peak posting activity, trending hashtags were identified [13]. Additionally, anomaly detection techniques were used to screen spam and bot accounts with approximately 80% accuracy [18].

Conclusion

This study shows that real-time social media data analysis driven by AI is both possible and extremely effective. Large-scale social media streams can be analyzed with low latency and high accuracy by merging sophisticated NLP models (BERT, LSTM, GPT) with big data frameworks (Kafka, Spark Streaming) [1][3]. These systems facilitate stronger crisis response, better governance, and quicker decision-making. Future developments in edge and multimodal AI will increase the technology's potential and range of uses [12][15].

References

1. [1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). Attention is all you need. Retrieved from https://arxiv.org/abs/1706.03762
3. Ayush Kumar & Kamalraj R (2024). Real-time sentiment analysis on social media streams using BERT and Apache Spark. Journal of Big Data.Retrieved from https://www.researchgate.net/publication/380707012_Real-Time_Sentiment_Analysis_System_Using_the_BERT_Model
4. Zimbra D., Abbasi A., Zeng D., & Chen, H. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. ACM TMIS.Retrieved from https://dl.acm.org/doi/10.1145/3185045
5. Saif, H., He, Y., Fernandez, M., & Alani, H. (2014). Semantic sentiment analysis of Twitter. Retrieved from https://doi.org/10.1007/978-3-31 9-11964-9_32

6. Md Shah Alam (2025). A comparative study of NLP models for real-time event detection. Procedia Computer Science.Retrieved from https://www.researchgate.net/publication/390335725_SENTIMENT_ANALYSIS_IN_SOCIAL_MEDIA_HOW_DATA_SCIENCE_IMPACTS_PUBLIC_OPINION_KNOWLEDGE_INTEGRATES_NATURAL_LANGUAGE_PROCESSING_NLP_WITH_ARTIFICIAL_INTELLIGENCE_AI
7. Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about COVID-19. JMIR Public Health and Surveillance. Retrieved from https://publichealth.jmir.org/2020/2/e19273
8. Mahmud H. (2018). Real-time topic detection from Twitter streams: A deep learning approach. Journal of Intelligent & Fuzzy Systems. Retrieved from https://www.researchgate.net/publication/323743194_Real-time_event_detection_from_the_Twitter_data_stream_using_the_TwitterNews_Framework
9. Muhammad, S., Muhammad, A., & Usman, M. (2022). Comparative analysis of sentiment analysis techniques for Twitter data. PeerJ Computer Science. Retrieved from https://peerj.com/articles/cs-999/
10. Liu, B. (2020). Sentiment analysis and opinion mining. Synthesis Lectures on HLT. Retrieved from https://doi.org/10.2200/S00416ED1V01Y201204HLT016
11. McCreadie, R., Macdonald, C., & Ounis, I. (2019). Incremental update strategies for real-time event detection. Information Retrieval Journal. Retrieved from https://doi.org/10.1007/s10791-018-09341-3
12. Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal Twitter data. Multimedia Tools and Applications. Retrieved from https://doi.org/10.1007/s11042-019-7593-9
13. Saha, P., Mathews, M., & Goyal, P. (2017). Predicting the real-time impact of events using social media. WWW Companion. Retrieved from https://doi.org/10.1145/3041021.3054171
14. Rizoiu, M. A., Lee, Y., Conover, M., & Xie, L. (2018). A data-driven approach to modeling viral content. WSDM.
15. Al-Smadi, M., Talafha, B., & Jararweh, Y. (2019). Aspect-based sentiment analysis of Arabic tweets. Retrieved from https://doi.org/10.1016/j.future.2019.02.012
16. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Retrieved from https://doi.org/10.1145/1014052.1014073
17. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter

sentiment analysis: The good, the bad and the OMG! ICWSM. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14107

18. Hasan, M., Orgun, M. A., & Schwitter, R. (2018). Real-time event detection from Twitter streams. Information Processing & Management. Retrieved from https://doi.org/10.1016/j.ipm.2018.05.001

19. Twitter Developer Documentation. (2024). Twitter API v2 Documentation. Retrieved from https://developer.twitter.com/en/ docs

20. Pawar, R., Patil, M., & Joshi, P. (2023). Survey on text summarization techniques in social media analytics. Sustainability.