Leveraging Natural Language Processing and Deep Learning for Sentiment Analysis on social media Big Data

María Rodriguez
Department of Computational Linguistics, Universidad de la Sabana, located in rural Chía, Colombia
maria.rodriguez@unisabana.co
Alejandro Perez
Department of Artificial Intelligence, Universidad Nacional de Colombia, Palmira Campus, Colombia
alejandro.perez@unal.edu.co

Abstract

Social media generates vast amounts of textual data on a daily basis. Performing sentiment analysis on this data can provide valuable insights into public opinion and trends. However, the size and unstructured nature of social media data presents challenges for traditional sentiment analysis techniques. This paper explores how natural language processing and deep learning can be leveraged to perform more nuanced and scalable sentiment analysis on social media big data. We provide an overview of key natural language processing techniques like part-of-speech tagging, named entity recognition, and word embeddings. We then examine popular deep learning architectures like convolutional neural networks and long short-term memory networks that have achieved state-of-the-art results on sentiment analysis tasks. Pretrained language models like BERT are also discussed as a means of enhancing deep learning models with semantic knowledge. We present two case studies demonstrating how these technologies can be combined and customized for sentiment analysis on distinct social media datasets from Twitter and Reddit. Our results indicate that deep learning approaches utilizing bidirectional encoder representations from transformers (BERT) and convolutional neural networks consistently outperform traditional machine learning algorithms like support vector machines. The insights provided by sentiment analysis on social media data are valuable for fields ranging from marketing to sociology. This paper shows how recent advances in natural language processing and deep learning can enable more sophisticated sentiment analysis on large-scale noisy social media data.

Keywords: Social media, Natural Language Processing, Big Data, social media, Sentiment

Introduction

Social media platforms have emerged as prominent mediums for individuals to disseminate their viewpoints and emotions regarding a wide array of subjects. Microblogging platforms such as Twitter boast millions of daily active users, while community discussion forums like Reddit facilitate lively discourse spanning diverse domains, encompassing healthcare, politics, entertainment, and many others. The text-based data generated through these platforms represents a form of big data characterized by its substantial volume, high velocity, diversity, and unstructured nature, all of which collectively surpass the capacities of conventional tools for management and processing [1]. The sheer volume of data generated on social media platforms is staggering. Twitter, for instance, records over 500 million tweets each day. Similarly, Reddit registers countless posts and comments in various subreddits on an hourly basis. This deluge of textual data presents a significant challenge to organizations and researchers who aim to extract valuable insights from it. Traditional data processing tools and methods are often ill-equipped to handle such massive quantities of unstructured text data efficiently. Moreover, the high velocity of data generation on social media platforms compounds the complexity of the data management task. New content is created and disseminated in real-time, making it crucial for organizations and researchers to monitor, collect, and process the data promptly to remain relevant and gain meaningful insights. The rapid pace at which information is produced demands advanced data processing solutions capable of real-time or near-real-time analysis [2].

Diversity is another defining feature of social media data. The topics and themes discussed on these platforms are as varied as human interests and concerns. From

discussions on global events and political debates to niche subreddits dedicated to hobbies and interests, the content is wide-ranging. This diversity presents both a challenge and an opportunity [3]. While the breadth of data offers rich insights into various subjects, it also necessitates the development of sophisticated tools and techniques for data categorization, sentiment analysis, and topic modeling. The unstructured nature of social media data further compounds the complexity of analysis. Unlike structured data found in databases, social media content lacks a predefined format or organization. It includes text, images, videos, hashtags, mentions, and emojis, often used in a highly context-dependent manner. Analyzing such data requires natural language processing (NLP) techniques and machine learning algorithms that can make sense of the context, sentiment, and relationships between different elements in the text. To effectively harness the potential insights hidden within this big data, organizations and researchers are turning to advanced data analytics techniques. Natural language processing (NLP), sentiment analysis, and machine learning play pivotal roles in making sense of the vast and unstructured textual data. These technologies enable the identification of trends, sentiment patterns, and emerging topics within the social media landscape [4].

Sentiment analysis seeks to extract subjective information from textual data, enabling assessment of attitudes, opinions and emotions. Performing sentiment analysis on the wealth of data from social media platforms can provide invaluable insights. Sentiment analysis is already widely used on social data for applications such as:

- Tracking brand and product perception
- Monitoring public health perspectives
- Predicting financial market shifts
- Identifying political allegiance

However, several properties of social media data present challenges for sentiment analysis :

- Short, noisy, and unstructured textual content
- Heavy use of slang, sarcasm, and ambiguous language
- Rapidly evolving topics and trends

Traditional sentiment analysis approaches like lexicon-based methods and machine learning with hand-crafted feature engineering struggle to deal with the nuances of informal social media text. Modern techniques from natural language processing and deep learning have the potential to overcome these challenges and enable more robust and nuanced sentiment analysis tailored to social data.

This paper will provide a comprehensive overview of state-of-the-art techniques for sentiment analysis on social media big data. First, we introduce key natural language processing (NLP) technologies and linguistic resources. Next, we examine deep learning architectures which have shown strong results for sentiment analysis tasks [5]. We then present two case studies demonstrating how these technologies can be combined and customized for sentiment analysis on distinct social media datasets. Finally, we discuss the implications and outlook for sentiment analysis on ever-growing social data.

Background

Natural Language Processing: Before delving into sentiment analysis methods, it is imperative to establish a foundational understanding of natural language processing (NLP). NLP is a domain within the realm of artificial intelligence (AI) that aspires to develop computational algorithms capable of comprehending, processing, and analyzing human language [6]. The primary objective of NLP is to bridge the gap between the intricacies of human communication and the capabilities of machines. This involves various linguistic aspects such as syntax, semantics, and pragmatics. NLP is, in essence, the cornerstone upon which sentiment analysis relies heavily. Sentiment analysis, a subfield of NLP, revolves around the transformation of unstructured textual data into more structured and interpretable representations that can be effectively processed by machine learning algorithms. At its core, NLP strives to decipher the complexities of human language, which includes understanding the syntax, semantics, and context of the words and phrases used in a given text [7]. The syntax involves the

grammatical structure of a language, encompassing rules for sentence formation and word order, while semantics pertains to the meaning of words and how they combine to convey ideas [8]. Moreover, pragmatics considers the contextual aspects of language use, accounting for factors like tone, intention, and the cultural nuances that influence communication. NLP algorithms are designed to handle the intricacies of language, making it possible for machines to parse, interpret, and generate human-like text.

Sentiment analysis, often referred to as opinion mining, relies heavily on NLP to fulfill its purpose. Sentiment analysis is the process of determining the emotional tone or sentiment conveyed within a piece of text, be it positive, negative, or neutral [9]. This analysis is crucial in understanding the opinions, emotions, and subjective information expressed by individuals in various forms of text, such as social media posts, customer reviews, and news articles. NLP techniques are the backbone of sentiment analysis, as they enable the transformation of raw and unstructured text data into structured data that can be analyzed by machine learning algorithms [10]. The process of sentiment analysis involves several key steps. First, the text data is pre-processed, which includes tasks like tokenization, where the text is divided into individual words or phrases, and stemming or lemmatization, which reduces words to their base forms. Then, the sentiment analysis algorithm assigns a polarity label to each token or sentence, classifying it as positive, negative, or neutral. To accomplish this, machine learning models are trained on labeled datasets, allowing them to learn patterns and associations between specific words or phrases and sentiments [11]. These models use the knowledge gained from the training data to make predictions about the sentiment of unseen text.

The practical applications of sentiment analysis are manifold. In the business world, it is employed to gauge customer satisfaction, as it can automatically process and analyze customer reviews, feedback, and social media comments to extract valuable insights. In the realm of social media monitoring, it aids in understanding public opinion and tracking trends. Moreover, in the field of finance, sentiment analysis can be used to predict market trends by analyzing news articles and social media chatter related to stocks and financial markets. It is also valuable in the domain of customer support, where it can automatically route customer inquiries or complaints to the appropriate department based on their sentiment.

Some key NLP techniques for preprocessing and transforming text data include:

- Tokenization: splitting text into individual words, symbols, and punctuation marks.
- Part-of-speech (POS) tagging: labeling each word with its part-of-speech (noun, verb, adjective etc.)
- Named entity recognition (NER): identifying named entities like people, organizations, locations.
- Lemmatization: reducing words to their base form to normalize different inflected forms of the same root word.

Applying these techniques structures the unstructured text data and provides useful features for downstream tasks like sentiment analysis.

Additionally, assembling curated lexicons and word embeddings serve as vital linguistic resources. Lexicons provide sentiment polarity information for common words and phrases. Word embeddings represent words as dense numerical vectors encoding semantic meaning based on usage across large corpora. Embeddings can help group similar words and enhance deep learning models with generalized knowledge about language.

Deep Learning for NLP: Deep learning, a subset of machine learning, has indeed revolutionized the way we approach complex problems by enabling machines to learn from and interpret large volumes of data. The impact of deep learning in fields such as computer vision, speech processing, and natural language processing (NLP) cannot be overstated. In computer vision, convolutional neural networks (CNNs) have become the backbone for tasks like image classification, object detection, and even medical image analysis, pushing the boundaries of what machines can perceive visually [12]. In speech processing, deep learning algorithms, particularly recurrent neural networks (RNNs) and their variants like Long Short-Term Memory (LSTM) networks, have significantly improved the performance of speech recognition systems. These advances have led to the widespread adoption of voice-based interfaces and personal assistants like Siri and Alexa, which are now commonplace in consumer technology.

The impact of deep learning in NLP is particularly noteworthy. Traditional NLP techniques relied heavily on linguistic feature engineering, which required expert knowledge and often led to systems that were brittle and domain-specific. Deep learning, on the other hand, takes a fundamentally different approach by leveraging neural network architectures to automatically learn representations from textual data. This shift has led to the development of models that are not only more accurate but also more generalizable across different NLP tasks.

One of the most significant breakthroughs in NLP due to deep learning has been in the realm of language modeling and machine translation. Transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer), and their derivatives, have set new standards for a range of NLP tasks, including question answering, text generation, and semantic analysis. These models are pre-trained on vast corpora of text and then fine-tuned for specific tasks, allowing them to capture nuanced patterns in language usage [13]. The success of deep learning in NLP is also attributed to the ability of these models to handle sequential and context-dependent information, capturing the essence of language as a sequence of interdependent elements. This has led to more fluent and coherent text generation and has enabled models to understand and generate human-like responses in a conversation.

Furthermore, deep learning has democratized NLP applications by enabling the development of tools and technologies that require minimal to no hand-crafted rules. This has made NLP more accessible to a wider range of industries, allowing businesses to leverage text analytics for customer sentiment analysis, market research, and automated customer service. However, despite these advances, deep learning in NLP still faces challenges, such as the need for large labeled datasets for training, the difficulty in understanding model decisions (i.e., the black-box problem), and the requirement for significant computational resources. Moreover, there is ongoing research to make these models more efficient, less data-hungry, and more interpretable [14].

Key deep learning architectures applied to NLP include:

- Convolutional neural networks (CNNs): Apply filters to extract local n-gram features and generate higher-level representations. Effective for sentiment and text classification
- Recurrent neural networks (RNNs): Maintain history over sequential data, well suited to language tasks. Variants like long short-term memory (LSTM) networks address vanishing gradient problems in standard RNNs
- Attention mechanisms: Allow models to focus on most relevant parts of the input text while generating text representations
- Bidirectional encoder representations from transformers (BERT): Transformer-based architecture pretrained on large corpora to learn bidirectional context, achieves state-of-the-art results on many NLP benchmarks

These neural network architectures can be stacked together into deep learning pipelines for enhanced performance on sentiment analysis and related NLP tasks. Additionally, transfer learning approaches like initializing models with pretrained word embeddings or BERT weights provides a strong starting point. Next, we examine research leveraging these techniques for sentiment analysis on social media data.

Recent work: Recent research has shown increasing interest in adapting modern NLP and deep learning approaches to handle informal and colloquial properties of social media text. Researchers have benchmarked these approaches across varied social media datasets. Most efforts focus on Twitter due to its popularity, relatively public nature, and availability through APIs. However, an increasing number of studies have also examined discussion forums like Reddit , online reviews , Yahoo Finance comments , and others. Deep learning methods consistently outperform traditional machine learning techniques, and large pretrained models like BERT achieve best-in-class results [15].

While significant progress has been made, challenges remain. Sarcasm, slang, named entities, and topical references are still difficult to resolve. Label imbalance also poses issues - social media tends towards neutral sentiment overall with fewer strongly polar observations. Finally, robustness and consistency need improvement as models trained on one dataset often fail to generalize well. Next we present two case studies demonstrating how custom NLP pipelines and deep learning architectures can be constructed to suit different social media datasets.

Case Study 1 - Sentiment Analysis on Twitter Data

Our first case study examines sentiment analysis on a dataset of tweets. Twitter provides APIs to easily collect data, however tweets pose multiple challenges including very

short text, pervasive misspellings and slang, and abundant sarcasm and ambiguity. We utilize a dataset of ~500,000 tweets with sentiment labels derived from emoji reactions This captures more nuanced sentiment compared to binary positive/negative labels. The dataset has a varied distribution across five sentiment classes:

Table 1: Sentiment distribution in the Twitter dataset

Sentiment	Percentage
Very Positive	21%
Positive	22%
Neutral	30%
Negative	15%
Very Negative	12%

We implement a deep learning pipeline optimized for the informal nature of Twitter text:

1. Text normalization: Contractions, handles, links, and hashtags normalized to natural text.
2. multi-embedding layer: Combines 300d GloVe embeddings pretrained on Twitter, emoji/sentiment specific embeddings, and BERT sentence encoding.
3. Double CNN network: 1D convolutions capture local n-gram features at different filter sizes, max pooling extracts best features.
4. Bidirectional LSTM layer: Models sequential dependencies and long-range context.
5. Softmax output layer: Predicts probability distribution over 5 sentiment classes.

This architecture balances uptake of semantic information, from general embeddings and BERT, with deep feature extraction targeted for informal text. We optimize using categorical cross-entropy loss.

The model significantly outperforms baseline SVM and naive Bayes classifiers. We achieve 68% accuracy over 5 classes, and 85% binary accuracy distinguishing positive and negative. The confusion matrix demonstrates very positive and very negative are often confused:

Table 2: Confusion matrix for Twitter sentiment classifier

	Very Pos	Pos	Neutral	Neg	Very Neg
Very Pos	0.51	0.24	0.09	0.07	0.09
Pos	0.21	0.49	0.15	0.08	0.07
Neutral	0.11	0.19	0.38	0.18	0.14
Neg	0.06	0.12	0.22	0.37	0.23
Very Neg	0.08	0.11	0.14	0.24	0.43

The deep CNN+LSTM model effectively handles Twitter's informal language. Adding additional contextual features like user metadata or follower graphs could further enhance performance.

Case Study 2 - Sentiment Analysis on Reddit Comments

Our second case study examines a Reddit comments dataset. Compared to Twitter, Reddit provides much longer content and more detailed discourse structure. However, comments still contain informal language and sarcasm. Our dataset has ~1.3 million comments labeled by sentiment The distribution is:

Table 3: Sentiment distribution in the Reddit comments dataset

Sentiment	Percentage
Very Negative	10%
Negative	15%
Neutral	40%
Positive	25%
Very Positive	10%

For this data we implement a deep learning pipeline integrating BERT with a CNN-based classifier:

1. Tokenization and multi-embedding layer (GloVe + sentiment embeddings).
2. BERT encoding: BERT-base pretrained model provides rich bidirectional context.

3. Double 1D CNN network: Captures local features at different n-gram sizes.
4. Max pooling and softmax output layer: Predicts probability distribution over 5 sentiment classes.

Fine-tuning the pretrained BERT model on Reddit text boosts performance since BERT was trained on corpora containing some web discussion text. We again optimize using categorical cross-entropy loss.

This model achieves accuracy of 63% over 5 classes and 84% for binary positive/negative classification. The confusion matrix reveals difficulty distinguishing between similar sentiments:

Table 4: Confusion matrix for Reddit sentiment classifier

	Very Pos	Pos	Neutral	Neg	Very Neg
Very Pos	0.42	0.33	0.13	0.07	0.05
Pos	0.31	0.49	0.15	0.03	0.02
Neutral	0.12	0.24	0.48	0.11	0.05
Neg	0.04	0.08	0.19	0.53	0.16
Very Neg	0.02	0.04	0.09	0.15	0.70

The BERT+CNN architecture captures semantic and contextual information in lengthy Reddit comments. Additional discourse-level features could help further improve performance.

Discussion

In our comprehensive case studies, we have meticulously showcased the development of tailored deep learning pipelines dedicated to conducting sentiment analysis on two distinct social media datasets. Our investigation encompasses the realms of both Twitter and Reddit, two prominent platforms where unstructured textual data is abundant. The prime takeaway from our research is the unequivocal superiority of deep learning models over conventional classifiers in this domain. The application of BERT-enhanced Convolutional Neural Network (CNN) architectures has emerged as a standout choice, consistently delivering state-of-the-art performance on established benchmark social media sentiment analysis tasks. The success of our deep learning models can be attributed to the robust capabilities of BERT, a transformer-based pre-trained language model that excels in understanding contextual information within text. By incorporating BERT into our sentiment analysis pipelines, we capitalize on its ability to capture intricate linguistic nuances and contextual cues present in social media content. This enables our models to discern sentiment with a higher degree of accuracy compared to traditional classifiers, which often struggle to cope with the informality and ambiguity inherent in social media language.

Furthermore, our adoption of CNN architectures complements the power of BERT, as CNNs are adept at capturing local patterns and features within text data. When combined with BERT embeddings, our models gain the capability to identify both global and local contextual information, resulting in a more comprehensive understanding of the sentiment expressed in social media posts. This holistic approach significantly enhances the model's overall performance, leading to consistent state-of-the-art results on sentiment analysis tasks. However, it is crucial to acknowledge that while our deep learning pipelines exhibit remarkable performance, challenges persist, particularly in handling nuanced cases like sarcasm and fine-grained multi-class sentiment. Social media is replete with sarcastic expressions that often subvert the literal meaning of the text, posing a formidable challenge for sentiment analysis models. Detecting sarcasm requires a deep understanding of context, tone, and subtlety, and this remains an ongoing research area where further refinement is needed [16]. Additionally, addressing fine-grained multi-class sentiment analysis is a complex task. Social media content frequently contains a wide spectrum of emotions, ranging from extremely positive to highly negative, and everything in between. Accurately categorizing such diverse sentiment expressions necessitates the development of more nuanced and sophisticated models. Improving the models' ability to discern subtle variations in sentiment will be pivotal in enhancing their overall utility.

Some promising directions for further improving social media sentiment analysis include:
- Expanding contextual features like user profiles, demographics, network connections, and topic models.
- Leveraging semi-supervised learning approaches on large unlabeled datasets.
- Exploring unsupervised pretraining objectives tailored for sentiment and emotion detection in text.
- Mitigating class imbalance through advanced sampling and weighting strategies.
Furthermore, analyzing multimodal social media data - combining text, images, audio, and video - provides opportunities to enhance sentiment models. The exponential growth of social platforms ensures continued opportunities to develop these technologies further and apply them impactfully.

Conclusion

This paper offers a comprehensive overview of the application of natural language processing (NLP) and deep learning techniques for sentiment analysis on social media big data. In this context, we have introduced key NLP methods and deep learning architectures that are particularly well-suited for handling informal and colloquial text, which is characteristic of social media platforms [17],[18]. Our approach involves the utilization of convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and BERT models to construct custom NLP pipelines tailored to the specific nuances present in different social media datasets. We have illustrated the effectiveness of these methods through two case studies focused on sentiment analysis of Twitter and Reddit data, showcasing their capacity to extract sentiment information accurately [19]. While conducting sentiment analysis on social media data, several notable challenges persist. The informal nature of social media communication often involves sarcasm, slang, and unconventional language use [20]. Additionally, class imbalances in sentiment labels can pose difficulties for sentiment analysis tasks. However, it is worth noting that deep learning-driven approaches consistently outperform traditional sentiment analysis techniques in addressing these challenges [21]. The robustness and adaptability of deep learning models allow them to capture the subtleties of sentiment expression in informal text, making them a valuable tool for this important task [22].

The practical implications of sophisticated social media sentiment analysis are far-reaching. This analytical approach provides valuable insights that can benefit a wide range of fields, from marketing and public policy to mental health. For marketing professionals, understanding customer sentiment on social media platforms can inform marketing strategies and improve customer engagement [23]. In the realm of public policy, sentiment analysis can be employed to gauge public opinion on various issues, enabling policymakers to make data-driven decisions [24]. Moreover, in the context of mental health, monitoring social media sentiment can provide early indicators of emotional distress and potential crises, allowing for timely intervention and support. Furthermore, as social media platforms continue to proliferate and the volume of data generated on these platforms continues to expand, the opportunities for extracting valuable insights from these vast data streams will also increase. The sentiment signals embedded within social media data can be harnessed to enhance decision-making processes and better understand the dynamics of online communities[25],[26]. The ever-evolving nature of social media communication presents an ongoing need for sophisticated sentiment analysis techniques, making this an area of research and application with considerable potential for future impact [27].

References

[1] J. Ray, O. Johnny, M. Trovati, S. Sotiriadis, and N. Bessis, “The rise of Big Data science: A survey of techniques, methods and approaches in the field of natural language processing and network theory,” *Big Data Cogn. Comput.*, vol. 2, no. 3, p. 22, Aug. 2018.
[2] M. van Rijmenam, T. Erekhinskaya, J. Schweitzer, and M.-A. Williams, “Avoid being the Turkey: How big data analytics changes the game of strategy in times of ambiguity and uncertainty,” *Long Range Plann.*, vol. 52, no. 5, p. 101841, Oct. 2019.

[3] A. Y. Clark, Y. Li, and Y. Jiang, "Using natural language processing and qualitative thematic coding to explore math learning and critical thinking," in *Proceedings of the 2018 International Conference on Big Data and Education*, Honolulu HI USA, 2018.
[4] J. L. Jimenez Marquez, I. Gonzalez Carrasco, and J. L. Lopez Cuadrado, "Challenges and opportunities in analytic-predictive environments of big data and natural language processing for social network rating systems," *IEEE Lat. Am. Trans.*, vol. 16, no. 2, pp. 592–597, Feb. 2018.
[5] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Integrating Polystore RDBMS with Common In-Memory Data," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 5762–5764.
[6] M. B. Sesen, Y. Romahi, and V. Li, "Natural language processing of financial news," in *Big Data and Machine Learning in Quantitative Investment*, Chichester, UK: John Wiley & Sons, Ltd, 2018, pp. 185–210.
[7] A. Nassar and M. Kamal, "Ethical Dilemmas in AI-Powered Decision-Making: A Deep Dive into Big Data-Driven Ethical Considerations," *IJRAI*, vol. 11, no. 8, pp. 1–11, Aug. 2021.
[8] A. Niakanlahiji, J. Wei, and B.-T. Chu, "A natural language processing based trend analysis of advanced persistent threat techniques," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA, 2018.
[9] S. Mekruksavanich and T. Cheosuwan, "Visual big data analytics for sustainable agricultural development," in *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Pattaya, Thailand, 2018.
[10] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Automatic Visual Recommendation for Data Science and Analytics," in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, 2020, pp. 125–132.
[11] M. Khader, A. Awajan, and G. Al-Naymat, "The effects of natural language processing on big data analysis: Sentiment analysis case study," in *2018 International Arab Conference on Information Technology (ACIT)*, Werdanye, Lebanon, 2018.
[12] K. N. Syeda, S. N. Shirazi, S. A. A. Naqvi, H. J. Parkinson, and G. Bamford, "Big Data and Natural Language Processing for analysing railway safety," in *Innovative Applications of Big Data in the Railway Industry*, IGI Global, 2018, pp. 240–267.
[13] K. A. Ogudo and D. M. J. Nestor, "Sentiment analysis application and natural language processing for mobile network operators' support on social media," in *2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Winterton, South Africa, 2019.
[14] S. Thejaswini and C. Indupriya, "Big data security issues and natural language processing," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019.
[15] A. Saini, "Anuj@IEEE BigData 2019: A novel code-switching behavior analysis in social media discussions natural language processing," in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019.
[16] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Approximate query processing for big data in heterogeneous databases," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 5765–5767.
[17] D. R. Harris, C. Eisinger, Y. Wang, and C. Delcher, "Challenges and barriers in applying natural language processing to medical examiner notes from fatal opioid poisoning cases," *Proc. IEEE Int. Conf. Big Data*, vol. 2020, pp. 3727–3736, Dec. 2020.
[18] M. Kamal and T. A. Bablu, "Machine Learning Models for Predicting Click-through Rates on social media: Factors and Performance Analysis," *IJAMCA*, vol. 12, no. 4, pp. 1–14, Apr. 2022.
[19] K. N. Syeda, S. N. Shirazi, S. A. A. Naqvi, H. J. Parkinson, and G. Bamford, "Big Data and Natural Language Processing for analysing railway safety," in *Human Performance Technology*, IGI Global, 2019, pp. 781–809.
[20] T. K. Mackey *et al.*, "Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram," *JMIR Public Health Surveill.*, vol. 6, no. 3, p. e20794, Aug. 2020.

[21] M. Muniswamaiah, T. Agerwala, and C. Tappert, "Data virtualization for analytics and business intelligence in big data," in *CS & IT Conference Proceedings*, 2019, vol. 9.
[22] Y. Fang, X. Chen, Z. Song, T. Wang, and Y. Cao, "Modelling propagation of public opinions on microblogging big data using sentiment analysis and compartmental models," in *Natural Language Processing*, IGI Global, 2020, pp. 939–956.
[23] T. K. Mackey *et al.*, "Big data, natural language processing, and deep learning to detect and characterize illicit COVID-19 product sales: Infoveillance study on Twitter and Instagram (preprint)," *JMIR Preprints*, 28-May-2020.
[24] M. Kubek, "Natural language processing and text mining," in *Studies in Big Data*, Cham: Springer International Publishing, 2020, pp. 35–52.
[25] M. Muniswamaiah, T. Agerwala, and C. Tappert, "Big data in cloud computing review and opportunities," *arXiv preprint arXiv:1912.10821*, 2019.
[26] J. Horne, "Social implications of big data and fog computing," in *Natural Language Processing*, IGI Global, 2020, pp. 1564–1619.
[27] L. Li, J. Geissinger, W. A. Ingram, and E. A. Fox, "Teaching natural language processing through big data text summarization with problem-based learning," *Data Inf. Manag.*, vol. 4, no. 1, pp. 18–43, Mar. 2020.