Sentiment Analysis in Analysing Monkeypox-Related Tweets Based on Deep Learning

Yidan Wang[a]
International College, Jinan University, Guangdong, China

Keywords: Sentiment Analysis, Deep Learning, Social Media, CNN-LSTM, Natural Language Processing.

Abstract: As the internet and social media platforms have rapidly expanded, sentiment analysis has emerged as a significant branch of natural language processing, focused on understanding individuals' emotions and attitudes toward specific topics. This article provides a comprehensive review of sentiment analysis evolution, from early dictionary-based methods to modern deep learning techniques. The focus is on a comparative evaluation of model outcomes from particular research, underscoring the effective performance of a combined Convolutional Neural Networks- Long Short-Term Memory (CNN-LSTM) deep learning model in analyzing sentiment within Monkeypox-related tweets. This model harnesses the local feature recognition of CNNs and the sequential data processing of LSTMs for accurate sentiment detection. Extensive experiments have demonstrated that this model outperforms standalone CNN-LSTM models in terms of stability and generalization capabilities. Future research will focus on utilizing more sophisticated sentiment analysis techniques, such as hierarchical attention networks, and cross-domain models, to enhance precision and applicability in various practical applications.

1 INTRODUCTION

Sentiment analysis also referred to as opinion mining, falls under the domain of natural language processing. It aims to analyze and understand the feelings, emotions, opinions, and attitudes that people express regarding a specific topic or subject (Medhat, 2014). It has garnered significant attention in recent years, particularly in fields such as customer satisfaction analysis and monitoring patients' mental health (Wankhade, 2022). This surge in interest can be attributed to the diverse array of data sources available, as well as advancements in technologies such as blockchain, cloud computing, and the Internet of Things (IoT). The effectiveness of sentiment analysis is further bolstered by the availability of sentiment dictionaries and corpora, which offer rich but varied resources for analysis. Utilizing sentiment analysis allows companies to discern customer emotions and opinions shared on social media, enhancing their understanding of consumer preferences and behaviors. Consequently, this insight supports strategic decision-making, influencing marketing tactics, product improvement, and swift resolution of emerging issues. Ultimately, sentiment analysis contributes to improving brand competitiveness and enhancing consumer satisfaction (Yang, 2020).

Sentiment dictionaries were the initial approach to sentiment analysis that relied on analyzing the quantity and polarity of emotional words in a text. These dictionaries were created either manually or through machine learning, and analyses that used rule-based or statistical techniques such as Naïve Bayes and Support Vector Machines were performed (Turney, 2002; Pang, 2002). While sentiment dictionary-based approaches efficiently capture unstructured features of texts, and are easy to analyze and understand, these methods need consistent updating since idioms and internet-specific language have continuously evolved due to the fast evolution of the internet and rapid information updates, plus they face polysemy issues. Traditional machine learning-based sentiment classification methods focus on extracting emotional features and selecting an appropriate combination of classifiers; thus making a significant impact on the analysis results. Despite this, these methods frequently overlook the

[a] https://orcid.org/0009-0008-9273-9756

Wang, Y.
Sentiment Analysis in Analyzing Monkeypox-Related Tweets Based on Deep Learning.
DOI: 10.5220/0012923200004508
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 215-220
ISBN: 978-989-758-713-9
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

contextual semantics of the text, which has a detrimental effect on classification accuracy. Sentiment analysis leveraging deep learning utilizes methods like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks. The study categorizes deep learning for sentiment analysis into four types: analysis using a single neural network, combined neural network models, analysis incorporating attention mechanisms, and the application of pre-trained models (Zhang, 2018; Yadav, 2020). Furthermore, it suggests that forthcoming sentiment analysis research should emphasize the integration of multimodal data, employing deep learning for cross-domain and real-time sentiment analysis, and conducting cross-cultural studies to improve the depth, precision, and practicality of sentiment interpretation across various sectors (Tan, 2023; Kaur, 2022; Zadeh, 2017).

The objective of this research is to deliver an exhaustive examination of sentiment analysis, spanning from the initial acquisition of text data through to the various phases of its processing. It will delve into the mainstream sentiment analysis methods throughout different historical periods, summarizing and categorizing these approaches. Additionally, it explores the various domains and applications of sentiment analysis, while also addressing the current research bottlenecks and future directions. The primary aim of this detailed review is to furnish insights for experts in the domain and to equip novices with essential knowledge and guidance.

The layout of the paper is divided into four parts. The introduction outlines the importance and scope of emotional analysis, laying the foundation for subsequent discussions. Next, this paper elaborates on key terms and concepts related to emotional analysis. It delves into the various sources of text data used in sentiment analysis, as well as the preprocessing steps required for effective analysis. The third section provides readers with a temporal perspective on the evolution of this field. This section explores different fields of practicality in sentiment analysis, showcasing the practical performance and significance of its core methods. Discuss the existing challenges in emotional analysis research and propose potential avenues for future exploration and development in this field. Finally, the summary will highlight the main discoveries and insights from the analysis, underlining the significance of sentiment analysis and its anticipated progression in both academic and practical arenas. Through these structured sections, readers will gain a comprehensive understanding of emotional analysis, its methods, applications, and future prospects.

2 METHODOLOGIES

2.1 Dataset Description and Preprocessing

Sentiment analysis research relies on key datasets like the Stanford Sentiment Treebank (SST), the Large Movie Review Dataset, and the Quora Question Pairs (QQP) (Zadeh, 2017). These datasets are pivotal for developing and testing sentiment analysis methods, offering diverse text samples, and enabling the exploration of different strategies. Two primary approaches are dictionary-based and machine learning-based methods. Dictionary-based techniques use predefined word lists with sentiment values to assess text sentiment across various contexts, while machine learning strategies involve training on these datasets to achieve deeper, context-sensitive sentiment understanding. These datasets provide critical platforms for advancing sentiment analysis techniques, accommodating various textual analyses, and facilitating continuous improvement of methodologies.

2.2 Proposed Approach

This study offers a comprehensive overview of sentiment analysis, charting its evolutionary path, examining its methodologies, exploring diverse applications, and pinpointing future research directions. To ensure clarity and focus, the study emphasizes specific research objectives and methodological frameworks. The author introduces the fundamental technology underpinning sentiment analysis, elucidating its core concepts and pivotal modules. A crucial aspect of this approach involves visually depicting the sentiment analysis pipeline, illustrating the sequential steps from data acquisition and preprocessing to the application of sentiment analysis techniques and the interpretation of findings. By delineating this systematic process, the study not only elucidates the study's aims but also provides a structured roadmap for conducting sentiment analysis research, fostering a deeper comprehension of its applications and potential advancements. Furthermore, the paper highlights three prominent methods: CNNs, RNNs, and LSTM networks. Figure 1 illustrates these concepts.

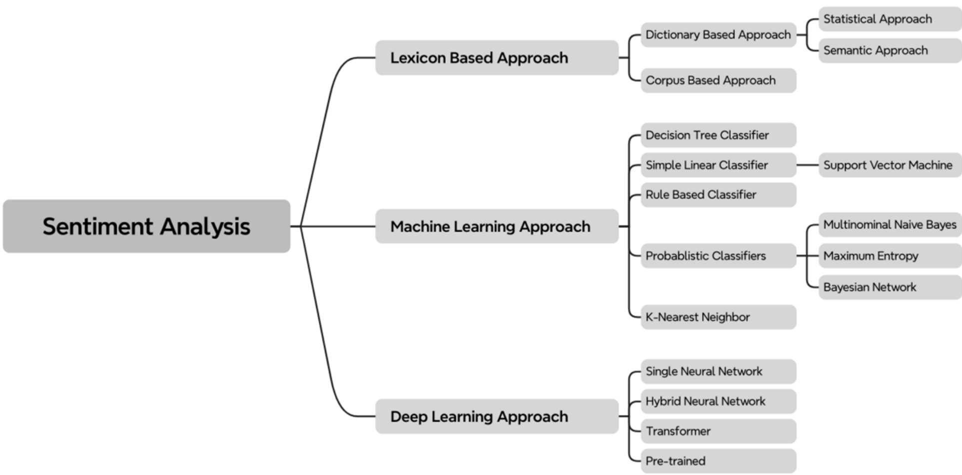

Figure 1: The pipeline of the model (Photo/Picture credit: Original).

2.2.1 Sentiment Lexicon

The analysis method based on sentiment dictionaries is one approach to sentiment mining and analysis. The general procedure involves: first, matching the text with sentiment words; then, aggregating and scoring these sentiment words; and finally, determining the text's sentiment orientation. Currently, the most widely used sentiment dictionaries are mainly of two types: one is the BosonNLP sentiment dictionary, and the other is a sentiment dictionary introduced by CNKI. Sentiment analysis via the BosonNLP sentiment lexicon follows a clear-cut process. Initially, the text undergoes segmentation into sentences and individual words, utilizing tools like jieba for division. Subsequently, this segmented word array is cross-referenced with the BosonNLP lexicon to log the sentiment values of corresponding words. The summation of these values follows, determining the overall sentiment: a total score above zero indicates positivity, whereas a score below zero suggests negativity. This procedure is depicted in Figure 2.

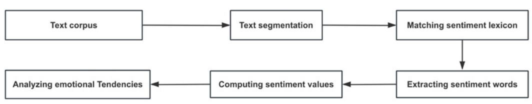

Figure 2: The process based on the Sentiment lexicon (Photo/Picture credit: Original).

Machine learning-based analytical techniques train algorithms on provided datasets to forecast results, a practice that has proven highly effective. For sentiment analysis, these methods harness extensive labeled or unlabeled text corpora, apply statistical algorithms for feature extraction, and perform analysis to generate insights. Sentiment classification leveraging machine learning is broadly segmented into supervised, semi-supervised, and unsupervised techniques. Supervised classification relies on data samples annotated with emotional polarity to identify sentiment categories, though it requires extensive manual labeling and processing. Typical supervised algorithms include k-nearest neighbor (KNN), Naive Bayes, and Support vector machine (SVM). Semi-supervised techniques augment sentiment classification outcomes by leveraging features from unlabeled texts and help mitigate the lack of labeled data. Unsupervised classification assigns sentiments by analyzing textual similarities, albeit it's a less frequent approach in sentiment analysis.

Deep learning analysis methods utilize neural networks for deeper insight extraction. Prominent neural network architectures for learning include CNNs, RNNs, and LSTMs. Deep learning for sentiment analysis is diverse, encompassing: the utilization of singular neural network models, integration of multiple neural networks into hybrid models, incorporation of attention mechanisms for enhanced focus within models, and application of pre-trained models for efficiency gains in analysis.

2.2.2 CNNs

Comprising layers of neurons, CNNs differ from fully connected networks, where each layer's neurons are interlinked with adjacent ones. In CNNs, interconnections occur selectively between layers, forming a three-dimensional structure from the input to convolutional and pooling layers. Distinct from fully connected layers, CNNs include an array of layers: an initial input layer, several convolutional and pooling layers, and a fully connected layer, followed by a SoftMax layer for output in classification tasks. Feature extraction is conducted by the convolutional layer using filters, which move across the input, identifying patterns and local connections. To condense the feature map while preserving essential information, the pooling layer employs techniques like max or average pooling, which also diminish computational load and help avert model overfitting. The fully connected layer synthesizes these features for output analysis, whereas the SoftMax layer transforms the fully connected layer's outputs into a probabilistic distribution for multi-class classification challenges.

In the field of text sentiment analysis, each component of CNNs offers distinct advantages. The convolutional layer can effectively identify local patterns in text, such as fixed collocations or semantic features between words, crucial for capturing emotional expressions. The pooling layer, by reducing the number of parameters and computational complexity, not only speeds up the training process but also helps the model abstract key information from broader text regions, enhancing the model's adaptability to different text lengths and structures. The fully connected layer integrates these features to form a comprehensive judgment of the text's emotional tendency. The SoftMax layer then translates this judgment into specific emotional category probabilities, facilitating accurate classification. Although CNNs have demonstrated strong performance in processing visual data and have shown unique advantages in text sentiment analysis, they still have limitations. Additionally, the performance of CNN models largely depends on the design of the convolutional kernels and parameter selection, requiring extensive experimentation and tuning for optimization. Thus, despite CNNs' significant potential application in text sentiment analysis, their limitations and applicable scenarios must be carefully considered in practical applications.

2.2.3 RNN and LSTM

RNNs, resembling typical neural networks, excel in processing sequential data by retaining contextual information across time steps. In translation tasks, maintaining coherence between successive words is vital, as translating word by word often leads to inaccuracies. RNNs address this by allowing hidden layer neurons to interact, preserving output from prior steps to influence subsequent word translations, thereby enhancing accuracy and cohesion. However, RNNs encounter challenges with lengthy sequences due to vanishing or exploding gradients, impeding their ability to learn distant dependencies. To mitigate this, LSTMs were introduced, featuring specialized gating mechanisms to manage long-term dependencies effectively. The forget gate is responsible for deciding which information from the past should be thrown away, the input gate refreshes the internal states with pertinent data, and the output gate controls the flow of important information to be outputted. LSTMs outperform standard RNNs in tasks necessitating long-term context consideration. Additionally, preprocessing techniques like Word2vec enhance model performance by providing richer input, enabling LSTMs to process and translate text data more accurately.

Word2vec learns semantic information by embedding words into vectors, grouping similar meanings in a multidimensional space. It maps words from their original space to a new one, predicting surrounding words based on initial word vectors and adjusting them through backpropagation to match target words. Similar words sharing contexts converge in this space. Commonly employed are the CBOW and Skip-gram architectures; CBOW forecasts a target word based on context words, while Skip-gram anticipates context words starting from a solitary target word. The Gensim package facilitates training, where the corpus is tokenized, stopwords removed, and words without vectors eliminated. LSTM, an enhanced RNN (Wang, 2016), integrates additional gates to control information flow. Figure 3 depicts the entire process.

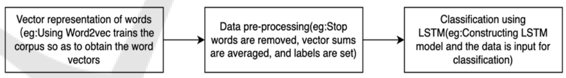

Figure 3: The whole process of LSTM (Photo/Picture credit: Original).

3 RESULTS AND DISCUSSION

This section is dedicated to examining the trends in loss and accuracy for CNN, LSTM, and their combined CNN-LSTM frameworks. It seeks to discern the effectiveness, as well as the potential advantages and limitations, of each model within the scope of identifying sentiments in text.

3.1 CNN

In this part, the author assesses the CNN model's performance in sentiment analysis using a Monkeypox tweets dataset. Figure 4 displays the model's (Mohbey, 2023) loss and accuracy curves on both training and test sets. Initially, the training loss sharply declines and stabilizes after the first 10 epochs. Conversely, the test loss exhibits fluctuations, particularly between the 30th and 60th epochs, suggesting potential issues with generalization or overfitting. While training accuracy improves rapidly, it eventually plateaus, whereas test accuracy, albeit increasing, remains consistently lower and displays minor fluctuations. This discrepancy implies potential overfitting of the model to the training data. Given that CNNs are primarily designed for image data, their application to textual data presents limitations, possibly overlooking essential text features and hindering generalization. Overall, while the CNN model fits well with the training data, it struggles to generalize to the test data. To address this, adjustments such as regularization techniques to mitigate overfitting and data enrichment may be necessary. Additionally, tailoring CNN architectures specifically for text data processing could enhance performance.

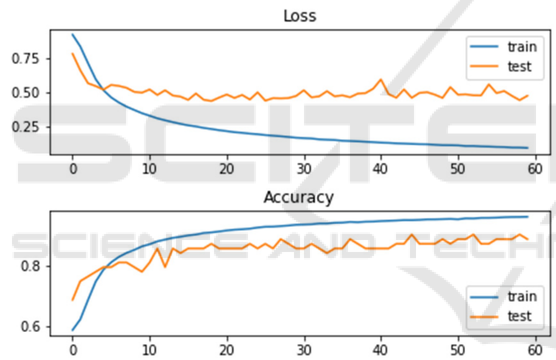

Figure 4: Training and Validation Performance of CNN Model (Mohbey, 2023).

3.2 LSTM

Figure 5 illustrates the training and test loss and accuracy of the LSTM model. From a loss perspective, the training loss initially drops sharply and then stabilizes, while the test loss follows a similar downward trend but with more pronounced fluctuations, particularly in later training stages. This fluctuation indicates a possible lack of stability in the model's capacity to generalize from new, unobserved data. The rapid initial decrease in training loss indicates efficient early learning, but subsequent spikes may indicate challenges in model generalization, possibly stemming from imbalanced data distribution, inadequate parameter tuning, or inappropriate learning rate settings leading to gradient explosion.

Regarding accuracy, the training accuracy steadily increases after an initial rise in the early epochs. Similarly, the test accuracy mirrors this pattern in the first 10 epochs but experiences significant fluctuations thereafter. This inconsistency suggests a suboptimal generalization of the model, possibly influenced by parameter settings and the dynamic learning environment. Therefore, while LSTM is theoretically well-suited for textual data, practical application necessitates meticulous model design, parameter tuning, and validation tailored to the specific application scenario and dataset.

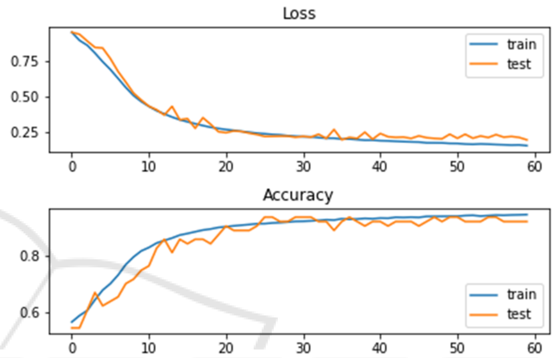

Figure 5: Training and Validation Performance of LSTM Model (Mohbey, 2023).

3.3 CNN-LSTM

The study demonstrates that, when evaluating Monkeypox-related tweet sentiments, the integrated CNN-LSTM model outperforms its individual CNN and LSTM counterparts in both features and effectiveness. Specifically, the CNN-LSTM model demonstrated greater stability in loss rates (show in Figure 6), with smoother curves for both training and test losses, indicating its effectiveness in avoiding overfitting and showcasing strong generalization capabilities when encountering unseen data. In terms of accuracy, the hybrid model not only achieved higher accuracy during training and testing but also maintained a smaller gap between the two, reflecting its consistent predictive power across different types of data. This enhancement likely stems from the effective combination of CNN local feature extraction and LSTM sequential dependency processing, with each complementing the other to ensure accuracy and robustness in text sentiment analysis.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

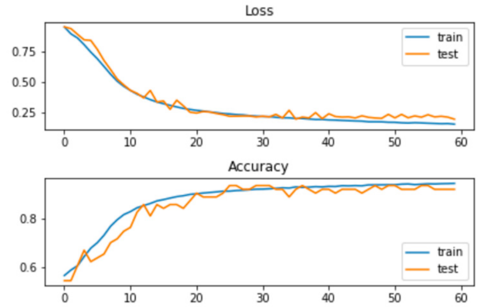

Figure 6: Training and Validation Performance of CNN-LSTM Model (Mohbey, 2023).

3.4 Future Extends

Navigating the intricacies of sentiment analysis requires overcoming several obstacles, including the deciphering of unstructured or ironic expressions, the need for more nuanced sentiment categorization, reliance on data with annotations, limitations inherent to present word embedding techniques, and biases embedded in the datasets used for training. It is imperative that future studies focus on elevating precision and broadening the scope of application. This involves the creation of models that adeptly discern subtle emotional nuances in text, assess the sentiment variance pertaining to different topics, adeptly handle texts with ambiguity and sarcasm, broaden sentiment analysis to encompass a wider range of languages, and enhance the efficacy of sentiment analysis on various social media platforms (Tan, 2023).

Potential approaches may involve employing hierarchical attention networks for nuanced sentiment analysis, integrating topic modeling with sentiment analysis to analyze sentiment distribution, leveraging reinforcement learning techniques to address sarcasm and ambiguity, creating cross-lingual models using transfer learning methods, and generating domain-specific embeddings tailored for social media texts. By surmounting these challenges and advancing model capabilities, sentiment analysis can broaden its practical applications and provide deeper insights into textual sentiments.

4 CONCLUSIONS

This paper provides a sample but comprehensive review of sentiment analysis, covering the entire spectrum from initial data sourcing to subsequent processing phases. Through an in-depth examination of a hybrid CNN-LSTM model, it has been established that such an approach enhances both accuracy in sentiment detection and robustness in sentiment expression. The experimental evaluation revealed that the CNN-LSTM hybrid model exhibits superior stability and generalization compared to standalone CNN or LSTM models. Moving forward, future research endeavors will focus on refining sentiment analysis techniques, integrating cross-linguistic models, and enhancing sentiment analysis effectiveness on social media platforms.

REFERENCES

Kaur, R., & Kautish, S. 2022. Multimodal sentiment analysis: A survey and comparison. Research anthology on implementing sentiment analysis across multiple disciplines, pp: 1846-1870.

Medhat, W., Hassan, A., & Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, vol. 5(4), pp: 1093-1113.

Mohbey, K. K., Meena, G., Kumar, S., & Lokesh, K. 2023. A CNN-LSTM-based hybrid deep learning approach for sentiment analysis on Monkeypox tweets. New Generation Computing, pp: 1-19.

Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. arXiv:0205070.

Tan, K. L., Lee, C. P., & Lim, K. M. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. Applied Sciences, vol. 13(7), p: 4550.

Turney, P. D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. arXiv: 0212032.

Wang, J., Yu, L. C., Lai, K. R., & Zhang, X. 2016. Dimensional sentiment analysis using a regional CNN-LSTM model. In Proceedings of the 54th annual meeting of the association for computational linguistics, vol. 2, pp: 225-230.

Wankhade, M., Rao, A. C. S., & Kulkarni, C. 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, vol. 55(7), pp: 5731-5780.

Yadav, A., & Vishwakarma, D. K. 2020. Sentiment analysis using deep learning architectures: a review. Artificial

Yang, L., Li, Y., Wang, J., & Sherratt, R. S. 2020. Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. vol. 8, pp: 23522-23530.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. 2017. Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250.

Zhang, L., Wang, S., & Liu, B. 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8(4), p: e1253.