Comparative Analysis of Machine Learning Models for Emotion Classification in Textual Data

Gregorius Airlangga
Universitas Katolik Indonesia Atma Jaya, Indonesia
E-mail: gregorius.airlangga@atmajaya.ac.id

Abstract

This research presents a comprehensive comparative analysis of various machine learning models for emotion classification within textual data, aiming to identify the most effective architectures for understanding and interpreting emotional undertones. With the increasing prevalence of digital communications, the ability to accurately classify emotions in text has significant implications across numerous domains, including social media analysis, customer service, and mental health monitoring. This study evaluates traditional algorithms, such as Logistic Regression, and advanced deep learning models, including Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Convolutional Neural Networks combined with Recurrent Neural Networks (CNN-RNN), Autoencoders, and Transformers. Through meticulous cross-validation, hyperparameter tuning, and performance evaluation based on accuracy, precision, recall, and F1 scores, the research elucidates the strengths and weaknesses of each model. LSTM and GRU models demonstrated superior performance, highlighting the importance of sequential data processing capabilities. In contrast, the Autoencoder model underperformed, underscoring the necessity for careful model selection tailored to the task's specifics. Surprisingly, Logistic Regression showed notable efficacy, advocating for its potential utility in scenarios prioritizing computational efficiency. This study enhances the understanding of affective computing within natural language processing, offering insights into the strategic deployment of machine learning models for emotion recognition and paving the way for future advancements in the field.

Keywords: *Emotion Classification, Machine Learning, Natural Language Processing, Deep Learning Models, Affective Computing*

1. Introduction

In the burgeoning field of emotion analysis using textual data, the interplay between machine learning techniques and natural language processing (NLP) has been a subject of extensive research [1]–[3]. This domain's significance is underscored by the diverse applications spanning customer service enhancement, mental health assessment, and the development of empathetic artificial intelligence systems [4]–[6]. The presented research delves into the nuances of emotional undertones in text, leveraging an expansive dataset to train various deep learning models aimed at accurately classifying text into distinct emotional categories: sadness, joy, love, anger, fear, and surprise [7]–[9]. The urgency of this research is rooted in the increasing prevalence of digital communication and the consequential imperative to comprehend the emotional context in vast textual data streams [10]–[12]. The advent of social media, online forums, and digital correspondence has furnished researchers with copious textual material, making it imperative to develop robust methodologies for deciphering embedded emotional sentiments [13]–[15].

A comprehensive literature survey reveals that while traditional machine learning methods have laid the groundwork, the state-of-the-art has progressively shifted towards more sophisticated deep learning models [16]. These include Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and, more recently, Transformer models, which have shown superior capabilities in capturing the semantic nuances of language [17]. However, a notable gap in the existing literature is the comparative analysis of these models' effectiveness in emotion classification within texts of varying contexts and lengths [18]–[20]. Additionally, many studies have predominantly focused on binary sentiment analysis (positive vs. negative), leaving a detailed exploration of nuanced emotional categories relatively untapped [20]–[22]. This research aims to fill these gaps by conducting an exhaustive analysis of logistic regression, different neural network architectures, including LSTM, GRU, Capsule Networks, Autoencoders, CNN-RNN and Transformer-based models, to determine their efficacy in classifying a diverse range of emotional states. The contribution of this study is twofold: first, it offers a comprehensive comparative analysis of various deep learning models in emotion classification, and second, it enhances the understanding of how these models can be optimized and implemented in real-world applications. Following the introduction, the remainder of this article is structured as follows: The next section elaborates on the methodology, encompassing the data preprocessing steps, model training, and evaluation criteria. Subsequent sections present a detailed analysis of the experimental results, followed by a discussion that interprets these findings in the broader context of current NLP challenges and applications. The final sections propose potential future research directions and conclude with a summary of the study's key insights and contributions.

2. Research Methodology

2.1. Data Acquisition and Preprocessing

The Data Acquisition and Preprocessing phase of our research into emotion classification using textual data begins with the careful selection of a dataset. We sourced a rich and diverse dataset from publicly available repositories, containing text snippets labeled with six distinct emotions: sadness, joy, love, anger, fear, and surprise[23]. This diversity is crucial to develop models capable of understanding and classifying a broad spectrum of emotional expressions within texts. Upon acquiring the dataset, we proceeded to load it into a Python environment using the Pandas library, which is renowned for its robust data manipulation capabilities. This step was essential for the preliminary inspection of the dataset, allowing us to gauge its structure, assess the volume of data available, and identify any immediate anomalies or missing values that could impact the analysis. The data loading process is foundational, as it sets the stage for the subsequent cleaning and preprocessing tasks.

In the data cleaning stage, our primary focus was on enhancing the quality and uniformity of the dataset to facilitate more accurate model training and analysis. We meticulously removed any rows with null values, particularly in the text or emotion label columns, to prevent any distortion in the model's learning process. Further, we standardized the text by converting it to lowercase, which is a common practice in text processing to avoid discrepancies in word recognition due to case differences. Additionally, we stripped the text of unnecessary characters such as punctuation, special symbols, and numbers, aiming to streamline the dataset to contain only meaningful textual content that would contribute to emotion classification. Tokenization was the next critical step in our preprocessing pipeline. Through this process, the text snippets were broken down into sequences of integers using the Keras Tokenizer, where each integer represents a unique word in the dataset. This conversion from raw text to a numerical format is vital, as it prepares the data for ingestion into the deep learning models, enabling them to process and analyze the textual content effectively.

Given the variability in the length of text entries, we employed sequence padding to ensure uniformity across all input sequences. This involved extending shorter sequences to match the length of the longest sequence in the dataset, thereby standardizing the input size for the models. Padding is essential for batch processing during model training, as it allows for consistent handling of text snippets, regardless of their original length. Finally, we split the dataset into training and testing sets using an 80/20 ratio. This division is strategic, ensuring that a significant portion of the data is used for training the models, while still reserving a substantial subset for testing. The testing set, comprising unseen data, is instrumental in evaluating the models' ability to generalize and accurately classify emotions in new texts, providing a measure of the models' performance and effectiveness in real-world applications.

2.2. Model Architecture and Implementation

In the realm of emotion classification through textual analysis, our research embarked on exploring various neural network architectures to construct a comprehensive comparative framework. This exploration was rooted in the hypothesis that the nuanced complexities of emotional expression in text necessitate a multifaceted approach to model architecture and implementation. The foundation of our comparative analysis was laid using basic sequential models, incorporating layers such as LSTM, Gated Recurrent Units (GRU), and dense layers. These models were pivotal in establishing a baseline for performance, serving as a benchmark against which the more advanced and specialized models could be evaluated. Sequential models, with their inherent ability to process input data in a linear fashion, were deemed suitable for capturing the temporal dependencies characteristic of textual data, where the sequence of words plays a crucial role in conveying emotion.

Moving beyond the basics, we delved into more complex neural network architectures to better capture the contextual dependencies and nuances within the text. Bidirectional LSTM (Bi-LSTM) models were employed to analyze the text data from both forward and backward directions, aiming to glean a more comprehensive understanding of the context surrounding each word. CNNs, traditionally celebrated for their image processing prowess, were adapted to our textual analysis needs, exploiting their capability to identify salient features and patterns within text snippets. Transformer-based models, renowned for their efficiency in handling long-range dependencies, were another cornerstone of our advanced modeling approach. These models, leveraging mechanisms like attention and self-attention, were particularly adept at discerning the intricate contextual cues embedded within textual data, which are essential for accurate emotion classification. Our exploration extended to specialized architectures like capsule networks and autoencoder models, venturing into relatively uncharted territories of neural network design for emotion classification. Capsule networks, with their dynamic routing capabilities, were hypothesized to excel in capturing hierarchical relationships and various features within text, potentially offering a nuanced understanding of emotional expressions. Autoencoders, known for their proficiency in data compression and feature extraction, were explored to ascertain their efficacy in distilling the essential emotional features from text, thereby facilitating a more refined classification process. In architecting these models, considerable attention was dedicated to the design of embedding layers, which play a critical role in transforming the tokenized text into meaningful vector representations. These vectors serve as the input to subsequent layers, whether recurrent, convolutional, or otherwise, each contributing to the model's ability to decode the linguistic and semantic patterns indicative of different emotions. Through this diverse array of models, our research aimed to not only identify the most effective architecture for emotion classification but also to enrich the broader discourse on how different neural network paradigms can be harnessed to enhance the understanding and interpretation of emotions in textual data.

2.3. Evaluation

In our quest to develop highly reliable models for emotion classification in textual data, we implemented rigorous cross-validation and hyperparameter tuning methodologies alongside a detailed evaluation framework to assess model performance. Ensuring the robustness of our models was paramount, as this directly impacts their utility in real-world applications where the accuracy and sensitivity of emotion detection can have significant implications. To achieve a robust evaluation of model performance, we employed a cross-validation approach, specifically opting for a five-fold strategy. This method involves partitioning the dataset into five subsets, where each subset serves as a test set while the remaining subsets are used for training, in a rotational manner. This technique is particularly effective in maximizing the utility of the training data and minimizing the evaluation bias that might arise from a single random split of data into training and testing sets. By leveraging cross-validation, we aimed to ensure that each model's performance was thoroughly vetted across different segments of the data, providing a more reliable estimate of its ability to generalize to unseen data. Hyperparameter tuning emerged as a critical component of our methodology, given its substantial impact on model performance. This process involved iteratively adjusting the models' hyperparameters—such as the number of layers, the size of each layer, and the learning rate—to identify the configurations that yielded the optimal balance between model complexity and performance. This trial and error process was guided by the performance metrics obtained from cross-validation, allowing us to refine each model's architecture and training process to enhance its emotion classification capabilities.

The evaluation of model performance was grounded in a set of well-established metrics: accuracy, precision, recall, and F1-score. These metrics collectively provided a comprehensive view of each model's effectiveness in classifying emotions within text. Accuracy served as a straightforward measure of overall performance, indicating the proportion of correct predictions out of all predictions made. Precision and recall offered more nuanced insights, with precision focusing on the model's ability to correctly identify texts as belonging to a specific emotion, and recall reflecting the model's success in capturing all relevant instances of each emotion. The F1-score, providing a harmonic mean of precision and recall, offered a balanced metric that considered both the precision and the recall of the models, enabling a more holistic assessment of their performance. Additionally, we generated confusion matrices for each model, which were instrumental in identifying any biases or tendencies in classification across different emotional categories. These matrices highlighted the models' strengths and weaknesses in distinguishing between various emotions, revealing potential areas for improvement in model training and architecture. Through this meticulous approach to cross-validation, hyperparameter tuning, and performance evaluation, our research endeavors to shed light on the capabilities and limitations of various deep learning models in the context of emotion classification. This comprehensive methodology not only enhances our understanding of the models' performance but also contributes valuable insights to the fields of natural language processing and affective computing, ultimately advancing the state-of-the-art in emotion detection within textual data.

3. Results and Discussion

In this section, we delve into the results obtained from the application of various deep learning models and a traditional machine learning model to the task of emotion classification in textual data. The analysis covers Long Short-Term Memory networks (LSTM), Multilayer Perceptrons (MLP), a combination of CNN-RNN, Autoencoders, Transformers, GRU, and Logistic Regression. The models were evaluated based on four primary metrics: mean accuracy, precision, recall, and F1 score. The LSTM model showcased a high level of accuracy in classifying emotions, with a mean accuracy of 93.48%. Its precision and recall were similarly high, indicating a strong capability to correctly classify texts into the appropriate emotional categories and a balanced performance across different emotions. The mean F1 score further confirms the model's consistent performance in precision and recall.

Multilayer Perceptrons (MLP) displayed solid performance with a mean accuracy of 89.97%. While slightly lower than the LSTM, MLP's precision, recall, and F1 score suggest a robust model with a good balance between identifying relevant instances and minimizing incorrect classifications. The CNN-RNN hybrid model achieved a mean accuracy of 92.57%, situating it between the LSTM and MLP models in terms of performance. Its precision and recall metrics underscore its effectiveness in emotion classification, benefiting from the strengths of both CNNs in feature extraction and RNNs in understanding sequence dependencies. Contrastingly, the Autoencoder model underperformed significantly with a mean accuracy of only 33.84%, accompanied by low precision and recall rates. This suggests that, despite its potential in feature extraction and representation learning, the autoencoder's architecture might not be optimal for direct emotion classification tasks without further refinement or integration with other model architectures.

Table 1. Performance Comparison of Deep Learning Models

Models	**Accuracy**	**Precision**	**Recall**	**F1 Score**
LSTM	0.934	0.936	0.934	0.934
MLP	0.899	0.900	0.899	0.899
CNN-RNN	0.925	0.927	0.925	0.925
Autoencoder	0.338	0.114	0.338	0.171
Transformer	0.904	0.906	0.904	0.904
GRU	0.935	0.938	0.935	0.935
Logistic Regression	0.932	0.932	0.932	0.932

The Transformer model, known for its efficiency in handling sequential data and long-range dependencies, yielded a mean accuracy of 90.43%. Its performance metrics indicate a competitive edge, likely due to its advanced attention mechanisms that provide a nuanced understanding of textual context. GRU, akin to LSTM in handling sequential information but with a simpler structure, achieved the highest mean accuracy among the tested models at 93.53%. Its precision, recall, and F1 score were comparably high, highlighting its efficiency and effectiveness in emotion classification tasks. Interestingly, the traditional machine learning model, Logistic Regression, showed commendable performance with a mean accuracy of 93.25%, closely rivaling the more advanced deep learning models. This highlights the potential of simpler models in certain NLP tasks, especially when computational efficiency and model interpretability are considered.

The comparative analysis of these models reveals several insights into the task of emotion classification in textual data. The superior performance of LSTM and GRU models underscores the importance of sequential data processing capabilities in understanding the contextual and temporal nuances of language. The competitive performance of the Transformer model further affirms the value of attention mechanisms in capturing the complexities of text-based emotion recognition. The underwhelming performance of the Autoencoder model suggests that while deep learning models offer substantial benefits for emotion classification, the choice of architecture and its alignment with the task's specific requirements is crucial. Moreover, the notable performance of Logistic Regression indicates that traditional machine learning models remain valuable contenders, particularly in scenarios where simplicity and interpretability are prioritized. These results contribute valuable insights to the ongoing exploration of affective computing and natural language processing. They highlight the diverse capabilities of different models in understanding and classifying emotions in text, offering a foundation for further research and application development in this dynamic field.

4. Conclusion

In conclusion, this research explored the intricate task of emotion classification in textual data through the lens of various machine learning models, spanning from traditional algorithms like Logistic Regression to advanced deep learning architectures such as LSTM, GRU, CNN-RNN, Autoencoder, and Transformer models. The empirical analysis, grounded in a methodical evaluation framework, provided a nuanced understanding of each model's capabilities and limitations in deciphering the emotional undertones of textual content. The LSTM and GRU models emerged as frontrunners, demonstrating exemplary performance with over 93% accuracy, precision, recall, and F1 scores. These results underscore the significance of sequential data processing and the ability of these models to capture the temporal and contextual nuances inherent in natural language. The Transformer model, renowned for its attention mechanism, also showed promising results, reinforcing its stature as a robust tool for natural language processing tasks.

Conversely, the Autoencoder model struggled in direct emotion classification, highlighting the importance of model selection and architecture tuning in alignment with the specific requirements of the task at hand. The MLP and CNN-RNN models showcased commendable performance, indicating their potential utility in emotion classification tasks when optimized effectively. The surprising efficacy of the Logistic Regression model, with performance metrics closely rivaling those of more complex models, serves as a testament to the power of traditional machine learning techniques in certain contexts, particularly where interpretability and computational efficiency are prioritized. This research contributes to the broader field of affective computing and natural language processing by offering a comparative analysis of various models' performance in emotion classification. It provides valuable insights into the design and selection of appropriate machine learning models for emotion recognition tasks, paving the way for future research to explore hybrid models, advanced feature engineering, and the integration of multimodal data sources to further enhance the accuracy and robustness of emotion classification systems.

References

[1] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, “Natural language processing (NLP) in management research: A literature review,” J. Manag. Anal., vol. 7, no. 2, pp. 139–172, 2020.

[2] D. Demner-Fushman, N. Elhadad, and C. Friedman, “Natural language processing for health-related texts,” in Biomedical Informatics: Computer Applications in Health Care and Biomedicine, Springer, 2021, pp. 241–272.

[3] M. Thimmapuram, D. Pal, and G. B. Mohammad, “Sentiment Analysis-Based Extraction of Real-Time Social Media Information From Twitter Using Natural Language Processing,” Soc. Netw. Anal. Theory Appl., pp. 149–173, 2022.

[4] J. Xue et al., “Evaluation of the Current State of Chatbots for Digital Health: Scoping Review,” J. Med. Internet Res., vol. 25, p. e47217, 2023.

[5] F. Khennouche, Y. Elmir, Y. Himeur, N. Djebari, and A. Amira, "Revolutionizing generative pre-traineds: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs," Expert Syst. Appl., vol. 246, p. 123224, 2024.

[6] M. Khaleel and A. Jebrel, "Artificial Intelligence in Computer Science," Int. J. Electr. Eng. Sustain., pp. 1–21, 2024.

[7] E. Troiano, "Where are emotions in text? A human-based and computational investigation of emotion recognition and generation," 2023.

[8] Pratibha, A. Kaur, M. Khurana, and R. Damaševičius, "Multimodal Hinglish Tweet Dataset for Deep Pragmatic Analysis," Data, vol. 9, no. 2, p. 38, 2024.

[9] M. Zaragozá Portolés, "Emotion recognition system through voice for the reproduction of emotionally tuned music," 2023.

[10] S. Lin, J. Zhang, L. Wang, and S. Wang, "Digital Realities: Role Stress, Social Media Burnout, and E-Cigarette Behavior in Post-90 s Urban White-Collar Workers," J. Knowl. Econ., pp. 1–34, 2024.

[11] G. S. Dhanesh and N. Rahman, "Visual communication and public relations: Visual frame building strategies in war and conflict stories," Public Relat. Rev., vol. 47, no. 1, p. 102003, 2021.

[12] E. Troiano, L. Oberländer, and R. Klinger, "Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction," Comput. Linguist., vol. 49, no. 1, pp. 1–72, 2023.

[13] J. Golia, Newspaper confessions: A history of advice columns in a pre-internet age. Oxford University Press, 2021.

[14] A. Whitworth, Mapping information landscapes: new methods for exploring the development and teaching of information literacy. Facet Publishing, 2020.

[15] I. A. Wong, Y. K. P. Wan, and D. Sun, "Understanding hospitality service aesthetics through the lens of aesthetic theory," J. Hosp. Mark. & Manag., vol. 32, no. 3, pp. 410–444, 2023.

[16] A. V Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions," Inf. Fusion, vol. 105, p. 102218, 2024.

[17] M. A. K. Raiaan et al., "A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges," IEEE Access, 2024.

[18] N. C. Dang, M. N. Moreno-Garc'ia, and F. la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics, vol. 9, no. 3, p. 483, 2020.

[19] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," Eng. Reports, vol. 2, no. 7, p. e12189, 2020.

[20] I. A. Farha and W. Magdy, "A comparative study of effective approaches for Arabic sentiment analysis," Inf. Process. & Manag., vol. 58, no. 2, p. 102438, 2021.

[21] G. Nkhata, S. Gauch, U. Anjum, and J. Zhan, "Fine-tuning BERT with Bidirectional LSTM for Fine-grained Movie Reviews Sentiment Analysis."

[22] H. Zhang, Y.-N. Cheah, O. M. Alyasiri, and J. An, "Exploring aspect-based sentiment quadruple extraction with implicit aspects, opinions, and ChatGPT: a comprehensive survey," Artif. Intell. Rev., vol. 57, no. 2, p. 17, 2024.

[23] Kotholo, "Emotions Logistic Regression 93\%." 2020.