INTELLITHINGS
Artificial Intelligence | Internet of Things

Vol. 1 No. 1, 2025 p-ISSN: xxxx-xxxx e-ISSN: xxxx-xxxx https://doi.org/xxxxxxxxx

Sentiment Analysis in the Age of Transformers and Large Language Models: A Comprehensive Review of Recent Advances and Future

Ata Amrullah

Department of Informatics, Darul Ulum Islamic University, East Java, Indonesia

Received: 2 December 2024
Accepted: 17 January 2025
Published: 24 January 2025

Keywords:
Sentiment Analysis;
Large Language Models;
Transformers;
Bias Mitigation;
Explainable AI.

Corresponding author:
Ata Amrullah
ata@unisda.ac.id

Abstract

Sentiment analysis (SA) has undergone a significant transformation with the emergence of Transformer-based models and Large Language Models (LLMs). This review provides a comprehensive overview of recent advances in Transformer-based SA, highlighting the impact of LLMs on accuracy, nuance, and context-awareness. Architectural innovations, domain adaptation techniques, and methods for handling context and long-range dependencies are explored. The review also addresses the critical challenges and limitations associated with LLMs in SA, including bias and fairness, interpretability and explainability, data scarcity, computational cost, and robustness against adversarial attacks. Mitigation strategies and best practices are examined, focusing on data augmentation, adversarial training, bias-aware training objectives, attention visualization, and model distillation. Finally, the review outlines future research directions, emphasizing multimodal sentiment analysis, explainable AI, ethical considerations, low-resource languages, and domain-specific applications. This work concludes that while LLMs offer unprecedented opportunities for advancing SA, addressing the identified challenges is crucial for ensuring responsible and effective deployment in real-world scenarios.

1. Introduction

Sentiment analysis (SA), also known as opinion mining, has become a crucial tool for understanding public attitudes and emotions across various domains. From monitoring brand reputation on social media to gauging customer satisfaction with products and services, SA provides valuable insights that drive decision-making in business, politics, and beyond [1]. Traditionally, SA relied on lexicon-based approaches and machine learning techniques such as Support Vector Machines (SVMs) and Naive Bayes classifiers [2]. However, these methods often struggle to capture the nuanced contextual information and long-range dependencies present in natural language.

The advent of Transformer-based models and Large Language Models (LLMs) has revolutionized the field of SA. Transformers, introduced by Vaswani et al. [3], leverage the self-attention mechanism to weigh the importance of different words in a sentence, enabling them to capture complex relationships and dependencies more effectively. LLMs, such as BERT [4], RoBERTa [5], and the GPT series [6], further enhance this capability by pre-training on massive datasets, allowing them to learn rich contextual representations and achieve state-of-the-art performance on a wide range of NLP tasks, including SA. Recent studies have demonstrated the superior performance of Transformer-based models over

traditional methods in SA across diverse datasets and languages [7], [8], [9].

Despite these advances, several challenges remain in leveraging LLMs for SA. These include addressing biases embedded in pre-trained models [10], improving interpretability and explainability [11], and handling data scarcity for specific languages or domains [12]. Furthermore, the high computational cost associated with training and deploying large Transformer models poses a significant barrier to adoption, particularly in resource-constrained environments. Addressing these challenges is critical for ensuring the responsible and effective deployment of LLMs for SA in real-world applications [13], [14].

This review article aims to provide a comprehensive overview of recent advances in Transformer-based SA, focusing on the impact of LLMs. It will synthesize the latest research on architectural innovations, domain adaptation techniques, bias mitigation strategies, and methods for improving interpretability and efficiency. This review is needed now because the field is rapidly evolving, and a consolidated analysis of the state-of-the-art is essential for guiding future research and development efforts.

The primary objectives of this review article are:

- To synthesize recent advances in Transformer-based architectures and training techniques for sentiment analysis.
- To identify the key challenges and limitations of using Large Language Models (LLMs) for sentiment analysis, including biases, interpretability issues, and computational costs.
- To examine existing strategies for mitigating bias, improving interpretability, and enhancing the efficiency of LLMs for sentiment analysis.
- To suggest promising future research directions in sentiment analysis, with a focus on addressing the identified challenges and leveraging emerging technologies.

This review aims to answer the following key research questions:

- What are the most effective Transformer-based architectures and training techniques for sentiment analysis, considering both accuracy and efficiency?
- What are the main sources of bias in LLMs, and how do these biases impact sentiment analysis results across different demographic groups and contexts?
- What are the most promising techniques for improving the interpretability and explainability of LLMs for sentiment analysis, allowing users to understand why a particular sentiment prediction was made?
- How can we effectively adapt Transformer-based models and LLMs to perform sentiment analysis in low-resource languages and specialized domains?
- What are the key ethical considerations in using LLMs for sentiment analysis, and how can we ensure responsible and fair deployment of these technologies?

The remainder of this review is organized as follows: Section 2 provides background information on sentiment analysis, Transformer architectures, and Large Language Models. Section 3 presents a detailed overview of recent advances in Transformer-based sentiment analysis, including architectural innovations and domain adaptation techniques. Section 4 discusses the challenges and limitations of using LLMs for sentiment analysis, focusing on bias, interpretability, data scarcity, and computational cost. Section 5 examines mitigation strategies and best practices for addressing these challenges. Section 6 explores future research directions in sentiment analysis, including multimodal sentiment analysis, explainable AI, and ethical considerations. Finally, Section 7 concludes the review with a summary of key findings and a perspective on the future of sentiment analysis in the age of Transformers and LLMs.

2. Background on Sentiment Analysis, Transformers, and LLMs

2.1. Sentiment Analysis Fundamentals

Sentiment analysis (SA) is a field of study within Natural Language Processing (NLP) that aims to identify, extract, quantify, and analyze affective states and subjective information [1]. In essence, SA seeks to determine the attitude, emotion, or opinion expressed in a piece of text towards a particular topic, product, service, organization, individual, or event. This analysis

can range from simple binary classification (positive, negative) to more nuanced scales that include neutral sentiment and varying degrees of positivity or negativity [2].

Several types of sentiment analysis exist, each tailored to specific applications and levels of granularity:

a. Fine-grained Sentiment Analysis: This approach goes beyond simple positive, negative, and neutral classifications to identify more subtle sentiment intensities, such as "very positive," "slightly positive," "slightly negative," and "very negative" [15].
b. Aspect-Based Sentiment Analysis (ABSA): ABSA focuses on identifying the specific aspects or features of an entity that are being discussed and determining the sentiment expressed towards each aspect. For example, in a review of a restaurant, ABSA can identify the sentiment towards the food, service, ambiance, and price separately [16].
c. Emotion Detection: This type of SA aims to identify the specific emotions expressed in a text, such as happiness, sadness, anger, fear, surprise, and disgust [17].
d. Intent Analysis: Closely related to SA, intent analysis focuses on understanding the underlying intention or purpose behind a piece of text, such as a customer's intention to purchase a product or a user's intention to perform a specific action [18].

SA has a wide range of applications across various domains:

a. Social Media Monitoring: Analyzing sentiment expressed in social media posts, comments, and tweets to understand public opinion about brands, products, and events [19].
b. Customer Feedback Analysis: Analyzing customer reviews, surveys, and feedback forms to identify areas for improvement and enhance customer satisfaction [20].
c. Market Research: Understanding consumer preferences and trends by analyzing sentiment expressed in online forums, blogs, and news articles [21].
d. Political Analysis: Gauging public sentiment towards political candidates, policies, and events to inform campaign strategies and policy decisions [22].
e. Healthcare: Analyzing patient feedback and medical records to improve healthcare services and patient outcomes [23].

Traditional SA techniques primarily relied on two main approaches:

a. Lexicon-Based Approaches: These methods use pre-defined dictionaries or lexicons of words and phrases associated with positive and negative sentiments. The sentiment of a text is determined by summing the sentiment scores of the words and phrases it contains [24].
b. Machine Learning-Based Approaches: These methods train machine learning models on labeled datasets of text and sentiment to learn patterns and relationships between text and sentiment. Common machine learning algorithms used for SA include Naive Bayes, Support Vector Machines (SVMs), and Maximum Entropy classifiers [25].

While these traditional techniques have been effective in many applications, they often struggle to capture the complexities of natural language, such as sarcasm, irony, and contextual dependencies. The advent of Transformer-based models and LLMs has provided a powerful new approach to SA that can overcome many of these limitations.

2.2. The Rise of Transformers

The Transformer architecture, introduced by Vaswani et al. in their seminal paper "Attention is All You Need" [3], has revolutionized the field of NLP. Unlike previous sequence models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) [26], which process text sequentially, Transformers rely on the self-attention mechanism to process the entire input sequence in parallel. This parallel processing capability enables Transformers to be significantly faster and more efficient than RNNs and LSTMs, particularly for long sequences.

Recently, advancements in deep learning have introduced the Transformer architecture, which has significantly improved sentiment classification performance. Moreover, the architecture has several advantages over previous sequence models [27]:

a. Parallel Processing: Enables faster training and inference, particularly for long sequences.
b. Long-Range Dependencies: Captures long-range dependencies more effectively than RNNs and LSTMs.
c. Contextual Information: Captures rich contextual information by attending to all other words in the sequence.
d. Scalability: Can be scaled to very large models with billions of parameters.

The success of the Transformer architecture has led to its widespread adoption in a wide range of NLP tasks, including machine translation, text summarization, question answering, and sentiment analysis.

2.3. Large Language Models (LLMs)

Large Language Models (LLMs) are neural network models that build upon the Transformer architecture by scaling up the size of the model (number of parameters) and pre-training on massive datasets of text and code [20]. These massive datasets, often consisting of billions or even trillions of tokens, allow LLMs to learn rich statistical patterns and relationships in language, enabling them to generate coherent and fluent text, translate languages, answer questions, and perform a wide variety of other NLP tasks.

Popular LLMs include:

a. BERT (Bidirectional Encoder Representations from Transformers): A masked language model that learns bidirectional representations of text by predicting masked words in a sentence [4].
b. RoBERTa (Robustly Optimized BERT Approach): An improved version of BERT that uses a larger training dataset and a more optimized training procedure [5].
c. GPT Series (Generative Pre-trained Transformer): A series of language models that use a decoder-only Transformer architecture to generate text autoregressively [6].
d. T5 (Text-to-Text Transfer Transformer): A language model that frames all NLP tasks as text-to-text problems, allowing it to be fine-tuned for a wide range of tasks using a single model [21].

LLMs can be fine-tuned for sentiment analysis by training them on labeled datasets of text and sentiment. Fine-tuning involves updating the weights of the pre-trained model to optimize its performance on the specific SA task. This approach typically results in significantly better performance than training a model from scratch, particularly when labeled data is limited [22].

The development of LLMs has had a profound impact on the field of sentiment analysis. LLMs can capture the nuances of language, understand context, and generate accurate sentiment predictions with minimal supervision. However, it's crucial to acknowledge their limitations, including potential biases and the computational resources required for training and deployment, as we explore further in the subsequent sections.

3. Recent Advances in Transformer-Based Sentiment Analysis

The advent of Transformer-based models has spurred significant advancements in sentiment analysis, leading to more accurate, nuanced, and context-aware sentiment predictions. This section delves into recent innovations in Transformer architectures, domain adaptation techniques, methods for handling context and long-range dependencies, and approaches to explainable sentiment analysis.

3.1. Architectural Innovations

Researchers have explored various architectural modifications to the standard Transformer to enhance its performance on sentiment analysis tasks. These innovations often focus on adapting the attention mechanism, incorporating sentiment-specific information, or improving the model's ability to capture subtle sentiment cues. Some notable architectural innovations include:

a. Sentiment-Specific Word Embeddings: Traditional word embeddings may not effectively capture sentiment-related information. Researchers have proposed methods for incorporating sentiment-specific information into word embeddings, such as training embeddings on sentiment-labeled data or using sentiment lexicons to guide the embedding process. For instance, [28] introduced Sentiment Treebank-enhanced embeddings that significantly improve sentiment classification accuracy.

b. Attentive Pooling Mechanisms: Pooling layers are often used in Transformer-based models to aggregate information from different parts of the input sequence. Attentive pooling mechanisms allow the model to selectively attend to the most relevant parts of the sequence when pooling, improving the model's ability to capture important sentiment cues. [29] proposed an attentive pooling mechanism that weighs the importance of different words based on their relevance to the overall sentiment.
c. Sentiment-Aware Attention: This approach modifies the self-attention mechanism to incorporate sentiment information directly into the attention weights. By attending to words that are highly indicative of sentiment, the model can better capture the overall sentiment of the text. For example, [30] proposed a sentiment-aware attention mechanism that uses sentiment lexicons to guide the attention process.
d. Incorporating External Knowledge: Several studies have explored incorporating external knowledge sources, such as sentiment lexicons or knowledge graphs, into Transformer-based models to enhance sentiment analysis performance. [31] showed that incorporating a sentiment lexicon into the Transformer architecture can improve accuracy, especially in cases where labeled data is limited.

3.2. Domain Adaptation and Transfer Learning

Sentiment analysis models often struggle to generalize well across different domains, due to variations in language, style, and sentiment expression. Domain adaptation and transfer learning techniques aim to address this issue by leveraging knowledge learned from one domain (the source domain) to improve performance in another domain (the target domain). Recent advances in this area include:

a. Fine-Tuning Pre-trained Models: A common approach to domain adaptation is to fine-tune a pre-trained Transformer model on the target domain data. This allows the model to leverage the knowledge it has already learned from the source domain while adapting to the specific characteristics of the target domain. [32] demonstrated the effectiveness of fine-tuning BERT on sentiment analysis tasks in various domains.
b. Adversarial Domain Adaptation: This technique uses adversarial training to learn domain-invariant features that are effective across both the source and target domains. By training the model to discriminate between the two domains, adversarial domain adaptation can reduce the domain gap and improve generalization performance. For instance, [33] used adversarial training to adapt a Transformer model for sentiment analysis in the healthcare domain.
c. Multi-Task Learning: This approach trains the model on multiple related tasks simultaneously, such as sentiment analysis and topic classification. By sharing knowledge between the tasks, multi-task learning can improve performance on each individual task and enhance generalization across domains. [34] proposed a multi-task learning framework for sentiment analysis that incorporates both sentiment classification and aspect extraction.
d. Few-Shot Learning: In many real-world scenarios, labeled data in the target domain is scarce. Few-shot learning techniques aim to address this issue by training models that can quickly adapt to new domains with only a few labeled examples. [35] introduced a meta-learning approach for few-shot sentiment analysis that learns to adapt to new domains with minimal supervision.

3.3. Handling Context and Long-Range Dependencies

Effective sentiment analysis requires capturing contextual information and long-range dependencies in text. Transformer-based models are well-suited for this task due to their ability to attend to all parts of the input sequence. Recent advances in this area include:

a. Hierarchical Attention Mechanisms: These mechanisms allow the model to attend to different levels of granularity in the input sequence, such as words, sentences, and paragraphs. By capturing relationships at multiple levels, hierarchical attention can improve the model's ability to understand complex context and long-range dependencies. [36] proposed a hierarchical attention network for sentiment analysis

that captures relationships between words, sentences, and documents.

b. Memory-Augmented Transformers: These models augment the Transformer architecture with external memory modules that allow the model to store and retrieve relevant information from previous inputs. This can be particularly useful for handling long-range dependencies in documents or conversations. [37] used a memory-augmented Transformer for sentiment analysis in online reviews, allowing the model to capture relationships between different reviews from the same user.
c. Graph-Based Approaches: Representing text as a graph, where nodes represent words or sentences and edges represent relationships between them, can be an effective way to capture contextual information and long-range dependencies. Graph-based approaches can be combined with Transformer models to enhance sentiment analysis performance. [38] proposed a graph convolutional network (GCN) combined with a Transformer for sentiment analysis, which models relationships between words in a sentence using a graph structure.

3.4. Explainable Sentiment Analysis

As Transformer-based models become more complex, it is increasingly important to understand why they make particular sentiment predictions. Explainable sentiment analysis (XSA) aims to provide insights into the decision-making process of these models, making them more transparent and trustworthy. Recent advances in XSA include:

a. Attention Visualization: Visualizing the attention weights learned by Transformer models can provide insights into which words or phrases the model is attending to when making sentiment predictions. This can help to understand the model's reasoning process and identify potential biases. For example, [39] used attention visualization to analyze the behavior of BERT for sentiment analysis.
b. Feature Attribution Methods: These methods aim to identify the input features (e.g., words or phrases) that are most responsible for a particular sentiment prediction. Feature attribution methods can help to understand which aspects of the text are driving the model's decision and identify potential areas for improvement. [40] applied feature attribution methods to Transformer-based models for sentiment analysis, identifying the words that have the greatest influence on the sentiment prediction.
c. Rule Extraction: This approach aims to extract human-readable rules from Transformer-based models that explain how they make sentiment predictions. Rule extraction can provide a more transparent and interpretable understanding of the model's decision-making process. [41] proposed a rule extraction method for sentiment analysis that extracts rules based on the attention weights learned by the Transformer model.
d. Knowledge Distillation: This technique involves training a smaller, more interpretable model to mimic the behavior of a larger, more complex Transformer-based model. Knowledge distillation can provide a more efficient and interpretable alternative to directly analyzing the larger model. [42] used knowledge distillation to train a simpler model for sentiment analysis that achieves comparable accuracy to a larger Transformer model while being more interpretable.

The advances discussed in this section highlight the ongoing efforts to improve the accuracy, generalizability, and interpretability of Transformer-based sentiment analysis models. As research in this area continues to evolve, we can expect to see even more sophisticated techniques emerge that push the boundaries of what is possible in sentiment analysis.

4. Challenges and Limitations of LLMs in Sentiment Analysis

While Large Language Models (LLMs) have achieved remarkable success in sentiment analysis, they are not without their limitations. This section explores key challenges related to bias and fairness, interpretability and explainability, data scarcity, computational cost and scalability, and robustness against adversarial attacks.

4.1. Bias and Fairness

One of the most pressing concerns with LLMs is their potential to perpetuate and amplify biases present in the training data. These biases can manifest in various forms, including gender bias, racial bias, political bias, and socioeconomic bias [43]. The impact of bias on sentiment analysis can be significant, leading to inaccurate or unfair sentiment predictions for certain demographic groups or viewpoints.

a. Gender Bias: LLMs may exhibit gender bias in sentiment analysis, associating certain sentiments or emotions with specific genders. For example, a model may be more likely to predict positive sentiment for male-authored text and negative sentiment for female-authored text, even if the content is similar [44].
b. Racial Bias: LLMs can also exhibit racial bias, associating certain sentiments or emotions with specific racial groups. This can lead to discriminatory outcomes in applications such as hiring, loan applications, and criminal justice [45].
c. Political Bias: The training data used for LLMs may contain political biases, leading the model to favor certain political viewpoints over others. This can affect the accuracy of sentiment analysis in political discussions and lead to skewed perceptions of public opinion [46].

Identifying and mitigating bias in LLMs for sentiment analysis is a complex challenge. Strategies for addressing bias include:

a. Data Augmentation: Augmenting the training data with examples that are representative of different demographic groups and viewpoints.
b. Adversarial Training: Training the model to be robust against adversarial examples designed to exploit biases.
c. Bias-Aware Training Objectives: Modifying the training objective to penalize biased predictions.
d. Debiasing Techniques: Applying post-processing techniques to remove or reduce bias in the model's predictions [47].

4.2. Interpretability and Explainability

LLMs are often considered "black boxes" due to their complex architecture and the large number of parameters they contain. Understanding why an LLM makes a particular sentiment prediction can be challenging, hindering trust and accountability. Interpretability and explainability are crucial for ensuring that LLMs are used responsibly and ethically in sentiment analysis.

Challenges related to interpretability and explainability include:

a. Complexity: The complex architecture of LLMs makes it difficult to understand the decision-making process.
b. Non-Linearity: LLMs learn non-linear relationships between input text and sentiment, making it hard to trace the flow of information and identify the key factors influencing the prediction.
c. Lack of Transparency: The internal workings of LLMs are often opaque, making it difficult to assess the model's strengths and weaknesses.

Techniques for improving the interpretability and explainability of LLMs for sentiment analysis include:

a. Attention Visualization: Visualizing the attention weights learned by the model to identify the words or phrases that are most relevant to the sentiment prediction [39].
b. Feature Attribution Methods: Identifying the input features (e.g., words or phrases) that have the greatest influence on the sentiment prediction [41].
c. Rule Extraction: Extracting human-readable rules from the model that explain how it makes sentiment predictions.
d. Knowledge Distillation: Training a simpler, more interpretable model to mimic the behavior of the LLM [42].

4.3. Data Scarcity for Specific Languages/Domains

LLMs typically require massive amounts of training data to achieve high performance. However, labeled data for sentiment analysis is often scarce for certain languages and specialized domains. This data scarcity can limit the accuracy and generalizability of LLMs in these contexts.

Challenges related to data scarcity include:

a. Limited Labeled Data: The lack of labeled data makes it difficult to train accurate sentiment analysis models.
b. Domain Adaptation Issues: Models trained on data from one domain may not generalize well to other domains due to

differences in language, style, and sentiment expression.
c. Language-Specific Challenges: Sentiment analysis in low-resource languages presents unique challenges due to the limited availability of linguistic resources and the complexity of the language [48].

Techniques for addressing data scarcity include:
a. Zero-Shot Learning: Leveraging pre-trained models to perform sentiment analysis in new languages or domains without any labeled data.
b. Few-Shot Learning: Training models that can quickly adapt to new languages or domains with only a few labeled examples.
c. Data Augmentation: Generating synthetic data to increase the size of the training set [49].
d. Transfer Learning: Transferring knowledge learned from other languages or domains to improve performance in the target language or domain.

4.4. Computational Cost and Scalability

LLMs are computationally expensive to train and deploy, requiring significant resources in terms of hardware, software, and energy consumption. The high computational cost can be a barrier to adoption, particularly for organizations with limited resources.

Challenges related to computational cost and scalability include:
a. Training Time: Training LLMs can take days or even weeks, requiring significant computational resources.
b. Inference Time: Deploying LLMs for real-time sentiment analysis can be computationally expensive, leading to slow response times.
c. Memory Requirements: LLMs require large amounts of memory to store the model parameters, which can be a limitation for deployment on resource-constrained devices.

Techniques for reducing the computational cost and improving the scalability of LLMs include:
a. Model Distillation: Training a smaller, more efficient model to mimic the behavior of a larger LLM [42].
b. Quantization: Reducing the precision of the model parameters to reduce memory requirements and improve inference speed.
c. Pruning: Removing less important connections in the model to reduce the number of parameters and improve efficiency.
d. Hardware Acceleration: Utilizing specialized hardware, such as GPUs and TPUs, to accelerate training and inference [50].

4.5. Robustness and Adversarial Attacks

LLMs can be vulnerable to adversarial attacks, where malicious actors craft subtle perturbations to input text that cause the model to make incorrect sentiment predictions. These attacks can have serious consequences in applications such as social media monitoring and fraud detection [51].

Challenges related to robustness and adversarial attacks include:
a. Vulnerability to Perturbations: LLMs can be easily fooled by small changes to the input text.
b. Difficulty in Detecting Attacks: Adversarial attacks can be difficult to detect, as the perturbed text may appear normal to human readers.
c. Lack of Robustness: LLMs may not generalize well to slightly different input text due to their sensitivity to specific word choices and phrasing.

Techniques for improving the robustness of LLMs against adversarial attacks include:
a. Adversarial Training: Training the model on adversarial examples to make it more robust against perturbations.
b. Input Sanitization: Preprocessing the input text to remove or neutralize potential adversarial perturbations.
c. Defense Mechanisms: Implementing defense mechanisms that detect and block adversarial attacks [52].

Addressing these challenges is crucial for ensuring the responsible and effective deployment of LLMs for sentiment analysis in real-world applications. Future research should focus on developing more robust, interpretable, and scalable LLMs that are less susceptible to bias and adversarial attacks.

5. Conclusion

This review has explored the transformative impact of Transformers and Large Language Models (LLMs) on the field of sentiment analysis. We have highlighted the significant advancements achieved through architectural innovations, domain adaptation techniques, and methods for handling context and long-range dependencies. Transformer-based models have demonstrated superior performance compared to traditional sentiment analysis methods, enabling more accurate, nuanced, and context-aware sentiment predictions.

However, this review also acknowledges the significant challenges and limitations associated with leveraging LLMs for sentiment analysis. These include concerns related to bias and fairness, interpretability and explainability, data scarcity for specific languages/domains, computational cost and scalability, and robustness against adversarial attacks. Each of these challenges presents a barrier to the widespread and responsible deployment of LLMs in real-world sentiment analysis applications.

The analysis of mitigation strategies and best practices reveals promising approaches for addressing these challenges. Techniques such as data augmentation, adversarial training, bias-aware training objectives, attention visualization, feature attribution methods, model distillation, quantization, and adversarial defense mechanisms offer potential solutions for improving the robustness, fairness, interpretability, and efficiency of LLMs in sentiment analysis.

References

[1] B. Liu, "Sentiment Analysis and Opinion Mining," 2012, doi: 10.1007/978-3-031-02145-9.

[2] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis*. now, 2008. doi: 10.1561/1500000011.

[3] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Naacl-Hlt 2019*, no. Mlm, pp. 4171–4186, 2018, [Online]. Available: https://aclanthology.org/N19-1423.pdf

[5] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, 2019, [Online]. Available: http://arxiv.org/abs/1907.11692

[6] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[7] A. H. Musawa, Ricky, I. A. Iswanto, and M. F. Hidayat, "Exploring Transformer-Based Model in Sentiment Analysis of Movie Review," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Aug. 2024, pp. 214–219. doi: 10.1109/ICITISEE63424.2024.10730492.

[8] X. Zhu, S. Gardiner, T. Roldán, and D. Rossouw, "The Model Arena for Cross-lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models," pp. 141–152, 2024, [Online]. Available: https://aclanthology.org/2024.wassa-1.12

[9] P. Xue and W. Bai, "A Fine-Grained Sentiment Analysis Method Using Transformer for Weibo Comment Text," *Int. J. Inf. Technol. Syst. Appoach*, vol. 17, no. 1, pp. 1–24, Jul. 2024, doi: 10.4018/IJITSA.345397.

[10] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, in NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 4356–4364.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[12] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach, "Are we really making much progress? A worrying analysis of recent neural recommendation approaches," in *Proceedings of the 13th ACM Conference on Recommender Systems*, in RecSys '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 101–109. doi: 10.1145/3298689.3347058.

[13] R. Stefan, G. Carutasu, and M. Mocan,

"Ethical Considerations in the Implementation and Usage of Large Language Models," in *The 17th International Conference Interdisciplinarity in Engineering*, L. Moldovan and A. Gligor, Eds., Cham: Springer Nature Switzerland, 2024, pp. 131–144.

[14] J. P. Venugopal, A. A. V. Subramanian, G. Sundaram, M. Rivera, and P. Wheeler, "A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data," *Appl. Sci.*, vol. 14, no. 23, 2024, doi: 10.3390/app142311471.

[15] K. L. Revathi, A. R. Satish, and P. S. Rao, "Fine - Grained Sentiment Analysis on Online Reviews," in *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, Jan. 2023, pp. 1–6. doi: 10.1109/ICAECT57570.2023.10118291.

[16] H. Zheng, J. Zhang, Y. Suzuki, F. Fukumoto, and H. Nishizaki, "Semi-Supervised Learning for Aspect-Based Sentiment Analysis," in *2021 International Conference on Cyberworlds (CW)*, 2021, pp. 209–212. doi: 10.1109/CW52790.2021.00042.

[17] A. D. L. Languré and M. Zareei, "Breaking Barriers in Sentiment Analysis and Text Emotion Detection: Toward a Unified Assessment Framework," *IEEE Access*, vol. 11, pp. 125698–125715, 2023, doi: 10.1109/ACCESS.2023.3331323.

[18] S. H. Lye and P. L. Teh, "Customer Intent Prediction using Sentiment Analysis Techniques," in *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2021, pp. 185–190. doi: 10.1109/IDAACS53288.2021.9660391.

[19] J. Y. M. Nip and B. Berthelier, "Social Media Sentiment Analysis," *Encyclopedia*, vol. 4, no. 4, pp. 1590–1598, 2024, doi: 10.3390/encyclopedia4040104.

[20] A. PATEL, P. OZA, and S. AGRAWAL, "Sentiment Analysis of Customer Feedback and Reviews for Airline Services using Language Representation Model," *Procedia Comput. Sci.*, vol. 218, pp. 2459–2467, 2023, doi: https://doi.org/10.1016/j.procs.2023.01.221.

[21] C. P. Gupta and V. V Ravi Kumar, "Sentiment Analysis and its Application in Analysing Consumer Behaviour," in *2023 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*, 2023, pp. 332–337. doi: 10.1109/ICETCI58599.2023.10331537.

[22] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," *Procedia Comput. Sci.*, vol. 167, pp. 1821–1828, 2020, doi: https://doi.org/10.1016/j.procs.2020.03.201.

[23] P. M. Madan, M. R. Madan, and D. P. Thakur, "Analysing The Patient Sentiments in Healthcare Domain Using Machine Learning," *Procedia Comput. Sci.*, vol. 238, pp. 683–690, 2024, doi: https://doi.org/10.1016/j.procs.2024.06.077.

[24] P. Thangavel and R. Lourdusamy, "A lexicon-based approach for sentiment analysis of multimodal content in tweets," *Multimed. Tools Appl.*, vol. 82, no. 16, pp. 24203–24226, 2023, doi: 10.1007/s11042-023-14411-3.

[25] A. Briciu, A.-D. Călin, D.-L. Miholca, C. Moroz-Dubenco, V. Petrașcu, and G. Dascălu, "Machine-Learning-Based Approaches for Multi-Level Sentiment Analysis of Romanian Reviews," *Mathematics*, vol. 12, no. 3, 2024, doi: 10.3390/math12030456.

[26] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, in ICML'15. JMLR.org, 2015, pp. 2342–2350.

[27] I. Perikos and A. Diamantopoulos, "Explainable Aspect-Based Sentiment Analysis Using Transformer Models," *Big Data Cogn. Comput.*, vol. 8, no. 11, 2024, doi: 10.3390/bdcc8110141.

[28] M. Kasri, M. Birjali, M. Nabil, A. Beni-Hssane, A. El-Ansari, and M. El Fissaoui, "Refining Word Embeddings with Sentiment Information for Sentiment Analysis," *J. ICT Stand.*, vol. 10, no. 3, pp. 353–382, 2022, doi: 10.13052/jicts2245-800X.1031.

[29] C. Yue, A. Li, Z. Chen, G. Luan, and S. Guo, "Domain-Aware Neural Network with a Novel Attention-Pooling Technology for Binary Sentiment Classification," *Appl. Sci.*, vol. 14, no. 17, 2024, doi: 10.3390/app14177971.

[30] W. Li, D. Li, H. Yin, L. Zhang, Z. Zhu, and P. Liu, "Lexicon-Enhanced Attention Network Based on Text Representation for Sentiment Classification," *Appl. Sci.*, vol. 9, no. 18, 2019, doi: 10.3390/app9183717.

[31] M. Rizinski, H. Peshov, K. Mishev, M. Jovanovik, and D. Trajanov, "Sentiment Analysis in Finance: From Transformers Back to eXplainable Lexicons (XLex)," *IEEE*

Access, vol. 12, pp. 7170–7198, 2024, doi: 10.1109/ACCESS.2024.3349970.

[32] M. Bilal and A. A. Almazroi, "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews," *Electron. Commer. Res.*, vol. 23, no. 4, pp. 2737–2757, 2023, doi: 10.1007/s10660-022-09560-w.

[33] Z. Li, L. Zhou, X. Yang, H. Jia, W. Li, and J. Zhang, "User Sentiment Analysis of COVID-19 via Adversarial Training Based on the BERT-FGM-BiGRU Model," *Systems*, vol. 11, no. 3, 2023, doi: 10.3390/systems11030129.

[34] G. Zhao, Y. Luo, Q. Chen, and X. Qian, "Aspect-based sentiment analysis via multitask learning for online reviews," *Knowledge-Based Syst.*, vol. 264, p. 110326, 2023, doi: https://doi.org/10.1016/j.knosys.2023.110326.

[35] B. Liang *et al.*, "Few-shot Aspect Category Sentiment Analysis via Meta-learning," *ACM Trans. Inf. Syst.*, vol. 41, no. 1, Jan. 2023, doi: 10.1145/3529954.

[36] P. Wang, J. Li, and J. Hou, "S2SAN: A sentence-to-sentence attention network for sentiment analysis of online reviews," *Decis. Support Syst.*, vol. 149, p. 113603, 2021, doi: https://doi.org/10.1016/j.dss.2021.113603.

[37] D. Lee, C. S. Prakash, J. FitzGerald, and J. Lehmann, "MATTER: Memory-Augmented Transformer Using Heterogeneous Knowledge Sources," 2024, [Online]. Available: http://arxiv.org/abs/2406.04670

[38] B. Liu, W. Guan, C. Yang, Z. Fang, and Z. Lu, "Transformer and Graph Convolutional Network for Text Classification," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, p. 161, 2023, doi: 10.1007/s44196-023-00337-z.

[39] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viegas, and M. Wattenberg, "AttentionViz: A Global View of Transformer Attention," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 262–272, 2024, doi: 10.1109/TVCG.2023.3327163.

[40] J.-S. Pan, G.-L. Wang, S.-C. Chu, D. Yang, and V. Snášel, "New feature attribution method for explainable aspect-based sentiment classification," *Knowledge-Based Syst.*, vol. 304, p. 112550, 2024, doi: https://doi.org/10.1016/j.knosys.2024.112550.

[41] S. Sivakumar and R. Rajalakshmi, "Context-aware sentiment analysis with attention-enhanced features from bidirectional transformers," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 104, 2022, doi: 10.1007/s13278-022-00910-y.

[42] A. Shirgaonkar, N. Pandey, N. C. Abay, T. Aktas, and V. Aski, "Knowledge Distillation Using Frontier Open-source LLMs: Generalizability and the Role of Synthetic Data," 2024, [Online]. Available: http://arxiv.org/abs/2410.18588

[43] D. You and D. Chon, "Trust & Safety of LLMs and LLMs in Trust & Safety," 2024, [Online]. Available: http://arxiv.org/abs/2412.02113

[44] Y. Guo *et al.*, "Bias in Large Language Models: Origin, Evaluation, and Mitigation," pp. 1–47, 2024, [Online]. Available: http://arxiv.org/abs/2411.10915

[45] X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, and X. Zhao, "Bias of AI-generated content: an examination of news produced by large language models.," *Sci. Rep.*, vol. 14, no. 1, p. 5224, Mar. 2024, doi: 10.1038/s41598-024-55686-2.

[46] L. H. X. Ng, I. Cruickshank, and R. K.-W. Lee, "Examining the Influence of Political Bias on Large Language Model Performance in Stance Classification," 2024, [Online]. Available: http://arxiv.org/abs/2407.17688

[47] G. Kuzmin, N. Yadav, I. Smirnov, T. Baldwin, and A. Shelmanov, "Inference-Time Selective Debiasing to Enhance Fairness in Text Classification Models," 2024, [Online]. Available: http://arxiv.org/abs/2407.19345

[48] N. Raghunathan and K. Saravanakumar, "Challenges and Issues in Sentiment Analysis: A Comprehensive Survey," *IEEE Access*, vol. 11, no. June, pp. 69626–69642, 2023, doi: 10.1109/ACCESS.2023.3293041.

[49] F. Ronchini, L. Comanducci, and F. Antonacci, "Synthetic training set generation using text-to-audio models for environmental sound classification," 2024, [Online]. Available: http://arxiv.org/abs/2403.17864

[50] D. Sanmartín and V. Prohaska, "Exploring Tpus for Ai Applications," 2023.

[51] N. Wang, K. Walter, Y. Gao, and A. Abuadbba, "Through the Lens of Attack Objectives," pp. 1–15.

[52] H. Waghela, "Robust Image Classification : Defensive Strategies against FGSM and PGD Adversarial Attacks".