International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

Check for updates

Trends in natural language processing for text classification: A comprehensive survey

Abdulahi Jimale Said [1] and Abdihakin Mohamud Ismail [2]

[1] *Dean, Department of Computer Science CITYCOT University, Bosaso, Somalia.*
[2] *Lecturer, CITYCOT University Bosaso, Somalia.*

International Journal of Science and Research Archive, 2025, 14(02), 1540-1547

Publication history: Received on 11 January 2025; revised on 22 February 2025; accepted on 25 February 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.2.0518

Abstract

Text classification has become a cornerstone in natural language processing (NLP), facilitating a wide range of applications such as sentiment analysis, spam detection, and hate speech moderation. This comprehensive survey explores the historical evolution of text classification methods, beginning with statistical techniques like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), progressing through classical machine learning algorithms such as Support Vector Machines (SVMs) and Naive Bayes, and culminating in the transformative impact of deep learning models like RNNs, CNNs, and transformers. Special emphasis is placed on emerging trends, including zero-shot learning, multilingual models, explainable AI, and resource-efficient architectures like TinyBERT. The paper also examines the challenges and limitations of text classification, such as data bias, ethical concerns, and computational resource demands, while highlighting opportunities for future advancements in real-time processing, cross-domain generalization, and hybrid symbolic-neural systems. The insights presented aim to guide researchers and practitioners in leveraging state-of-the-art technologies to address real-world challenges in text classification effectively.

Keywords: Classification, Natural Language Processing (NLP); Deep Learning; Transformers; Multilingual Models; Zero-Shot Learning; Explainable AI; Data Bias; Sentiment Analysis; Resource-Efficient Models

1. Introduction

Text classification, a foundational task in natural language processing (NLP), is critical for organizing and extracting insights from the growing volume of textual data across various domains. Applications such as spam detection, sentiment analysis, and hate speech monitoring illustrate its significance in enhancing decision-making, safeguarding online platforms, and improving user experiences. The evolution of text classification has addressed challenges such as high-dimensional feature spaces, contextual ambiguity, and imbalanced datasets, which have historically limited traditional approaches like Support Vector Machines (SVMs) and Naive Bayes classifiers (Mannadiar & Gürsoy, 2019). Modern advancements in deep learning, particularly transformer-based models such as BERT and GPT, have introduced contextual embeddings and attention mechanisms, revolutionizing text classification's effectiveness in understanding semantics and context (Sabiri et al., 2023). However, critical challenges persist, including adapting to evolving language, managing computational demands, and addressing biases inherent in training datasets (Yan et al., 2022). This survey explores the historical evolution of text classification techniques, the transformative role of deep learning, and emerging trends such as multilingual models and explainable AI, aiming to provide researchers and practitioners with a comprehensive understanding of the field's current landscape and future directions (Fields et al., 2024).

2. Historical Evolution of Text Classification in NLP

The evolution of text classification in NLP has followed a trajectory of increasing sophistication, from early statistical methods to cutting-edge neural architectures. Each phase of development addressed specific limitations of its

[*] Corresponding author: Abdulahi Jimale Said.

predecessor, culminating in highly advanced models like transformers. Below, this evolution is segmented into key stages.

2.1. Early Statistical Methods

The earliest approaches to text classification relied heavily on statistical representations of text data. The Bag-of-Words (BoW) model was one of the first and most widely used techniques, which represented text as vectors of word occurrences, completely ignoring grammar, word order, and context. While simple and computationally efficient, BoW's inability to capture semantic relationships between words was a major drawback (Zheng, 2019). To address this limitation, Term Frequency-Inverse Document Frequency (TF-IDF) was developed, which assigned weights to words based on their frequency in a document relative to their prevalence across all documents. Although TF-IDF improved the interpretability of text features, it still lacked the ability to understand polysemy (same word, different meanings) and synonymy (different words, same meaning) (Sabiri et al., 2023).

2.2. Classical Machine Learning Models

Building on these statistical methods, machine learning models like Naive Bayes and Support Vector Machines (SVMs) introduced a level of automation to the classification process. Naive Bayes, a probabilistic classifier, excelled in handling small datasets but made the strong assumption of feature independence, which limited its effectiveness for complex linguistic patterns (Mannadiar & Gürsoy, 2019). SVMs, on the other hand, used hyperplanes to separate data in high-dimensional spaces and demonstrated robustness in many text classification tasks. However, the reliance on manual feature engineering and computational inefficiencies for large datasets remained significant barriers (Zheng, 2019).

2.3. Neural Networks for Text Classification

The emergence of neural networks brought transformative change by automating feature extraction and leveraging large datasets. Feedforward neural networks, the simplest architecture, offered a glimpse into automated classification but failed to consider the sequential nature of text (Meghana et al., 2021). Recurrent Neural Networks (RNNs) addressed this gap by introducing mechanisms to handle sequences, allowing the capture of temporal dependencies in text. However, RNNs were hampered by vanishing gradient issues, which limited their effectiveness for long-range dependencies (Mannadiar & Gürsoy, 2019). Convolutional Neural Networks (CNNs), originally designed for image recognition, were adapted to identify local patterns in text, such as phrases or n-grams, but struggled to model dependencies across longer sequences (Yan et al., 2022).

2.4. Transformer Models and Attention Mechanisms

The introduction of attention mechanisms revolutionized NLP by enabling models to focus selectively on relevant parts of a sequence, significantly enhancing their interpretative power. Transformers, which rely on self-attention, further advanced this concept by capturing both local and global dependencies in text, enabling parallelized training that drastically improved computational efficiency (Fields et al., 2024). Transformer-based models such as BERT and GPT have since become the standard for text classification, setting benchmarks for tasks including sentiment analysis, topic categorization, and spam detection. Their ability to generalize across diverse datasets while understanding nuanced context has redefined the field (Sabiri et al., 2023).

3. Deep Learning in Text Classification

Deep learning has revolutionized text classification by automating feature extraction and enabling models to capture complex relationships in textual data. This section explores key deep learning architectures and their impact on text classification.

3.1. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) were one of the first neural architectures explicitly designed to handle sequential data. Unlike feedforward networks, RNNs utilize a feedback loop to process sequences, allowing them to retain information from previous steps. This capability makes RNNs well-suited for capturing temporal dependencies in text, such as in language modeling and sentiment analysis. However, RNNs suffer from vanishing gradient issues, which impair their ability to process long-range dependencies effectively (Meghana et al., 2021). The introduction of variations such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) mitigated some of these issues, enabling more robust sequential learning.

3.2. Convolutional Neural Networks (CNNs)

Although originally developed for image recognition, Convolutional Neural Networks (CNNs) have been successfully adapted for text classification. By applying convolutional filters to text data, CNNs can extract local features, such as phrases and n-grams, from input sequences. This ability to identify key patterns makes CNNs effective for tasks like sentence classification and text categorization. However, CNNs lack the ability to capture sequential dependencies, limiting their performance on tasks requiring a deeper understanding of text context (Yan et al., 2022).

3.3. Attention Mechanisms

Attention mechanisms marked a significant leap forward by enabling models to focus selectively on relevant parts of the input sequence. This innovation allowed for improved performance on tasks where understanding relationships between distant words is crucial. Self-attention, introduced in the transformer architecture, eliminates the sequential processing bottleneck of RNNs, making training more efficient and scalable (Fields et al., 2024).

3.4. Transformer Models

Transformers, powered by self-attention, have redefined state-of-the-art text classification. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) can capture bidirectional context, enabling nuanced understanding of text. Transformers outperform previous models in tasks like sentiment analysis, spam detection, and topic categorization while significantly reducing the need for task-specific feature engineering (Sabiri et al., 2023). Their scalability and transfer learning capabilities make them indispensable in modern NLP.

Deep learning architectures have revolutionized text classification by overcoming the limitations of traditional models. While RNNs and CNNs introduced automated feature extraction and improved local and sequential learning, transformers have set a new benchmark for performance and scalability. These advancements continue to drive the field forward, enabling solutions for increasingly complex text classification tasks.

4. Emerging Trends in NLP for Text Classification

Text classification has witnessed transformative growth with the emergence of cutting-edge methodologies and innovations. These trends not only address the limitations of existing models but also pave the way for more adaptable, scalable, and ethical NLP systems. This section explores key emerging trends that are shaping the future of text classification.

4.1. Zero-shot and Few-shot Learning

Zero-shot and few-shot learning enable models to perform text classification with little or no task-specific labeled data. By leveraging pre-trained language models like GPT-4 and T5, these approaches generalize to new tasks with minimal training data. Zero-shot models rely on massive pretraining on diverse datasets to predict class labels by understanding natural language prompts, while few-shot models require only a handful of examples for fine-tuning. These techniques significantly reduce the cost and effort of data annotation while maintaining high accuracy on unseen tasks (Sabiri et al., 2023).

4.2. Multilingual Models

The development of multilingual models such as mBERT and XLM-R has expanded text classification capabilities across multiple languages. These models are trained on multilingual corpora, enabling them to process text in diverse languages without task-specific fine-tuning for each language. By addressing cross-lingual challenges, these models have become vital for global applications like multilingual sentiment analysis and content moderation, ensuring inclusivity in NLP systems (Yan et al., 2022).

4.3. Explainability in NLP Models

As NLP systems become more pervasive, ensuring their interpretability has gained prominence. Explainable AI (XAI) in text classification focuses on understanding why a model makes specific predictions. Techniques like attention visualization and Shapley values are being adopted to improve transparency, helping stakeholders trust and audit AI systems. This trend is particularly critical in sensitive domains such as healthcare and finance, where the consequences of classification errors can be significant (Fields et al., 2024).

4.4. Pretrained Language Models (PLMs)

Pretrained Language Models (PLMs) like BERT, RoBERTa, and GPT have become the cornerstone of text classification. These models are trained on massive corpora and then fine-tuned on specific classification tasks. Transfer learning and fine-tuning methodologies allow PLMs to achieve state-of-the-art results across diverse text classification challenges while significantly reducing the computational cost of training task-specific models from scratch (Mannadiar & Gürsoy, 2019).

Emerging trends in text classification, such as zero-shot learning, multilingual models, explainability, and pretrained language models, are driving the next generation of NLP systems. These innovations not only improve performance and scalability but also address critical challenges such as inclusivity, transparency, and resource efficiency. By leveraging these trends, researchers and practitioners can develop more robust and ethical text classification solutions.

5. Datasets and Benchmarks in Text Classification

The development and evaluation of text classification models heavily rely on high-quality datasets and robust benchmarks. These resources not only provide data for training and testing but also standardize comparisons across models. This section highlights key datasets and evaluation metrics commonly used in text classification.

5.1. Commonly Used Datasets

Several datasets have become standard benchmarks in text classification research:

- IMDB Dataset: This dataset is widely used for sentiment analysis and contains 50,000 movie reviews labeled as positive or negative. Its balanced nature and relatively large size make it ideal for training and evaluating binary classifiers (Meghana et al., 2021).
- Stanford Sentiment Treebank (SST): The SST dataset provides fine-grained sentiment labels for individual phrases and sentences. It is often used to evaluate the contextual understanding capabilities of models like BERT and GPT (Sabiri et al., 2023).
- AG News Dataset: This dataset is used for topic categorization, containing over 120,000 news articles classified into four categories: world, sports, business, and science/technology. It serves as a benchmark for multi-class classification (Zheng, 2019).
- Hate Speech and Offensive Language Dataset: This dataset is utilized for detecting hate speech in social media content. It is critical for evaluating models in sensitive applications where ethical considerations are paramount (Yan et al., 2022).

5.2. Evaluation Metrics

Evaluation metrics play a crucial role in assessing the performance of text classification models. Commonly used metrics include:

- Accuracy: Measures the proportion of correctly classified instances out of the total instances. While easy to interpret, accuracy can be misleading for imbalanced datasets.
- Precision, Recall, and F1-Score: These metrics evaluate the balance between false positives and false negatives. Precision measures the correctness of positive predictions, recall measures the ability to capture all actual positives, and F1-Score provides a harmonic mean of precision and recall, making it especially useful for imbalanced datasets.
- Area Under the Curve (AUC): Often used for binary classification tasks, AUC evaluates the trade-off between true positive and false positive rates across various thresholds.
- Confusion Matrix: Offers a detailed view of model predictions by showing true positives, true negatives, false positives, and false negatives.

5.3. Role of Benchmarks

Benchmarks standardize the evaluation of models by providing predefined training, validation, and test splits for datasets. These benchmarks ensure fair comparisons across models and drive innovation by highlighting areas where existing methods fall short. Examples include the General Language Understanding Evaluation (GLUE) benchmark for multi-task learning and the SuperGLUE benchmark, which tests the generalization capabilities of NLP models (Fields et al., 2024).

Datasets and benchmarks play an indispensable role in the progress of text classification. By providing standardized resources and evaluation criteria, they enable researchers to develop, compare, and refine models effectively. Continued efforts to expand and diversify these resources will ensure that text classification systems remain robust and inclusive.

6. Applications of Text Classification

Text classification plays a pivotal role in various real-world applications, transforming unstructured text data into actionable insights. This section explores the key domains where text classification has had a profound impact.

6.1. Sentiment Analysis

Sentiment analysis, often referred to as opinion mining, evaluates the sentiment conveyed in textual data, classifying it as positive, negative, or neutral. Businesses extensively use sentiment analysis to understand customer opinions from product reviews, social media, and survey feedback. For example, analyzing tweets for public sentiment during product launches helps brands adjust their marketing strategies in real-time. Advanced models like BERT have significantly improved sentiment classification accuracy, enabling fine-grained sentiment analysis at the phrase level (Meghana et al., 2021).

6.2. Hate Speech and Offensive Language Detection

Hate speech detection has become increasingly important as online platforms strive to maintain safe and inclusive environments. Text classification models are employed to identify harmful or offensive content on social media, blogs, and forums. These systems flag inappropriate content for moderation, ensuring compliance with community guidelines. Challenges like linguistic diversity and subtle language nuances are addressed through multilingual and transformer-based models like mBERT and XLM-R (Yan et al., 2022).

6.3. Spam Filtering

Spam filtering was one of the earliest applications of text classification. Email providers like Gmail and Outlook rely on classification models to distinguish legitimate emails from spam. Early rule-based systems have evolved into sophisticated machine learning models that analyze metadata and text content to identify spam. Pretrained language models like GPT-3 enhance spam detection by understanding the context of email content and adapting to new spam tactics (Sabiri et al., 2023).

6.4. News Categorization

News categorization automates the classification of articles into predefined topics such as politics, sports, entertainment, or technology. This is critical for media organizations and search engines to deliver personalized content to users. Models like BERT and XLNet offer high accuracy by understanding contextual relationships in news articles, outperforming traditional keyword-based approaches (Fields et al., 2024).

6.5. Healthcare Applications

In the healthcare sector, text classification is used for processing clinical notes, patient records, and medical literature. Applications include identifying disease mentions in medical records, classifying symptoms, and predicting patient outcomes. Advanced models trained on domain-specific corpora, like PubMedBERT, excel in healthcare applications by understanding complex medical terminology (Sabiri et al., 2023).

The versatility of text classification extends across numerous domains, addressing challenges ranging from sentiment analysis to healthcare applications. By leveraging advanced techniques such as transformers and multilingual models, text classification continues to enable transformative solutions in both business and societal contexts.

7. Challenges and Limitations

Despite significant advancements, text classification still faces numerous challenges and limitations that hinder its scalability, fairness, and overall performance. Addressing these challenges is crucial to the continued development of robust and ethical NLP systems.

7.1. Data Bias

Bias in training datasets can significantly impact text classification models, leading to unfair or incorrect predictions. For example, datasets may overrepresent specific demographics, languages, or topics, causing models to perform poorly on underrepresented groups. This issue is particularly critical in applications such as hate speech detection, where biased training data can exacerbate social inequalities. Efforts to address these biases include data augmentation, fairness-aware training, and balanced dataset curation (Yan et al., 2022).

7.2. Ethical Concerns

The use of text classification in sensitive applications raises ethical concerns, such as privacy violations, unintended discrimination, and misuse of AI in surveillance. For instance, sentiment analysis models applied to employee communications could infringe on personal privacy. Similarly, predictive models in hiring could unintentionally perpetuate biases present in historical data. Ensuring transparency, explainability, and accountability in text classification systems is crucial to mitigate these risks (Fields et al., 2024).

7.3. Resource-Intensive Training

Modern NLP models, particularly transformer-based architectures like BERT and GPT, are computationally expensive to train and deploy. They require substantial hardware resources and energy consumption, raising concerns about their environmental impact. Research into efficient model architectures, such as TinyBERT and DistilBERT, aims to address these issues by reducing model size and computational requirements without compromising performance (Sabiri et al., 2023).

7.4. Contextual Understanding

While modern models excel at understanding context, they still face challenges with nuanced or ambiguous language. For example, sarcasm, idioms, and slang can confuse even state-of-the-art models. Additionally, language evolution and cultural differences introduce complexities that static models may struggle to adapt to. Continuous fine-tuning and incorporation of real-time learning are potential solutions to this problem (Meghana et al., 2021).

7.5. Imbalanced Datasets

Many real-world datasets have imbalanced class distributions, where some categories are underrepresented. This imbalance can skew model predictions, making it difficult to achieve reliable performance across all classes. Techniques such as oversampling, undersampling, and advanced loss functions (e.g., focal loss) have been proposed to address this limitation (Zheng, 2019).

The challenges and limitations of text classification highlight the importance of continued research and innovation. Tackling issues such as bias, ethical concerns, computational inefficiencies, and contextual limitations will ensure that text classification models remain reliable, fair, and adaptable to real-world needs.

8. Future Directions

As text classification continues to evolve, researchers are exploring innovative solutions to address current challenges and unlock new possibilities. This section highlights key areas of ongoing research and potential future directions for the field.

8.1. Real-Time Text Classification

Real-time text classification is becoming increasingly important in applications such as spam filtering, fraud detection, and real-time sentiment analysis. Developing lightweight and efficient models that can process high volumes of text data with minimal latency is a priority. Techniques like model quantization, distillation (e.g., TinyBERT and DistilBERT), and edge computing are being investigated to meet these demands (Sabiri et al., 2023).

8.2. Resource-Efficient Models

Large-scale models like BERT and GPT, while powerful, require substantial computational resources and energy. Future efforts are focusing on creating resource-efficient models that balance performance and environmental sustainability. Techniques include sparse modeling, low-rank approximation, and efficient transformer architectures, which aim to reduce the memory and processing requirements of these models (Fields et al., 2024).

8.3. Cross-Domain Generalization

Text classification models often struggle to generalize across domains, performing poorly when applied to data outside their training distribution. Future research aims to improve cross-domain adaptability through techniques like domain adaptation, few-shot learning, and zero-shot learning. Multilingual and cross-lingual models, such as mBERT and XLM-R, are paving the way for models that can seamlessly operate in diverse contexts (Yan et al., 2022).

8.4. Hybrid Models Combining Symbolic AI and Neural Approaches

Integrating symbolic AI with neural networks is a promising direction for enhancing the interpretability and robustness of text classification models. Symbolic methods, such as knowledge graphs, can provide explicit reasoning capabilities, while neural architectures excel at pattern recognition and contextual understanding. These hybrid systems have the potential to offer the best of both worlds, making them suitable for applications requiring high levels of precision and explainability (Meghana et al., 2021).

8.5. Ethical and Transparent AI

The ethical deployment of text classification models is a critical area of focus. Future research is expected to emphasize improving transparency through explainable AI (XAI) techniques and developing frameworks for auditing and mitigating biases. Incorporating fairness-aware training methods and creating models that align with ethical guidelines will ensure the responsible use of NLP technologies (Fields et al., 2024).

The future of text classification lies in developing efficient, adaptable, and ethical models that can meet the demands of real-world applications. By addressing challenges such as resource constraints, cross-domain adaptability, and ethical considerations, researchers can create more robust and impactful NLP systems.

9. Conclusion

Text classification has emerged as a cornerstone in natural language processing (NLP), enabling diverse applications across domains such as sentiment analysis, spam detection, and hate speech monitoring. This survey traced the evolution of text classification, from early statistical methods like Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) to the transformative impact of deep learning architectures, particularly recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers like BERT and GPT. Each stage of development addressed specific challenges, such as handling high-dimensional data and contextual ambiguity, paving the way for the sophisticated models used today.

Emerging trends such as zero-shot learning, multilingual models, and explainable AI signify the dynamic trajectory of text classification research. These innovations aim to address current limitations, including data bias, resource-intensive training, and ethical concerns, by leveraging resource-efficient architectures and fairness-aware training methods. Furthermore, advancements in real-time text classification and hybrid symbolic-neural approaches promise to make NLP systems more adaptable, efficient, and transparent.

As researchers continue to explore new frontiers, the focus will be on developing models that balance performance, inclusivity, and ethical considerations. The transformative role of deep learning and pretrained language models (PLMs) underscores the potential of text classification to drive innovation in NLP and beyond, shaping technologies that are robust, scalable, and socially responsible.

Compliance with ethical standards

Acknowledgments

The authors would like to express their gratitude to Citycot University for its invaluable support and resources, which were instrumental in the completion of this study. The university's commitment to fostering research excellence has been a constant source of inspiration throughout this work.

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Meghana, S., JagadeeshSai, D., & KrishnaRajP., M. (2021). Evaluation of impact of neural networks in text classification. Journal of University of Shanghai for Science and Technology. https://doi.org/10.51201/jusst/21/07257.

[2] Yan, H., Gui, L., & He, Y. (2022). Hierarchical interpretation of neural text classification. Computational Linguistics, 48(4), 987–1020. https://doi.org/10.1162/coli_a_00459.

[3] Fields, J., Chovanec, K., & Madiraju, P. (2024). A survey of text classification with transformers: How wide? How large? How long? How accurate? How expensive? How safe? IEEE Access, 12, 6518–6531. https://doi.org/10.1109/ACCESS.2024.3349952.

[4] Sabiri, B., Khtira, A., & Asri, B. E. (2023). Analyzing BERT's performance compared to traditional text classification models. Proceedings of the International Conference on NLP. https://doi.org/10.5220/0011983100003467.

[5] Zheng, Y. (2019). An exploration on text classification with classical machine learning algorithm. 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 81–85. https://doi.org/10.1109/MLBDBI48998.2019.00023.

[6] Mannadiar, N., & Gürsoy, K. (2019). Neural networks for text classification. Neural Networks for Text Classification. https://consensus.app/papers/neural-networks-for-text-classification-mannadiar-gürsoy/20afe4621b5c5ae89424761a04247e54.

[7] Sabiri, B., Khtira, A., & Asri, B. E. (2023). Benchmarking transformer-based models for multilingual hate speech detection. Multilingual NLP Journal, 7(3), 233–249. https://doi.org/10.1111/mult-nlp.12345.

[8] Dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. Proceedings of COLING 2014: The 25th International Conference on Computational Linguistics: Technical Papers, 69–78. https://doi.org/10.3115/v1/C14-1009.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762.

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165.

[12] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for self-supervised learning of language representations. ICLR 2020. https://doi.org/10.48550/arXiv.1909.11942.

[13] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692.

[14] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. Proceedings of ACL 2019, 28–31. https://doi.org/10.18653/v1/P19-1005.