News Article Categorization Based on the BERT Model

Lakshmi Sri D
Bharathiar University

ManiKandan P
Bharathiar University

Nikasthika K
Bharathiar University

Surya Prakash P
Bharathiar University

Abstract

News article categorization is a critical task in natural language processing (NLP) that involves classifying articles into predefined categories based on their content. Traditional models have limitations in handling the complexities of language, leading to the development of deep learning approaches. Among these, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a state-of-the-art model due to its ability to understand context and semantics. This document explores existing models, their disadvantages, and how BERT works, highlighting its advantages and importance in the domain of news article categorization.

Keywords: Deep Learning, NLP, BERT, BERTimbau, Transformers, Word2Vec, News classification

1.Introduction

News articles are a primary source of information for the public, and categorizing them effectively is essential for content management, recommendation systems, and enhancing user experience. Traditional methods of categorization often fall short due to their inability to capture the nuances of language. The advent of deep learning has revolutionized this field, leading to the development of models that can better understand and classify text. BERT, introduced by Devlin et al. in 2018, represents a significant advancement in this area, leveraging the power of transformers and bidirectional context to improve classification accuracy.

2.Literature Review

Naive Bayes is a family of probabilistic algorithms based on Bayes' theorem, which assumes that the presence of a particular feature in a class is independent of the presence of any other feature. Struggles with complex language structures and relationships(Johnson and Zhang, 2016)] Support Vector Machines (SVM). Multi-category news classifcation using Support Vector Machine based classifers SVM is a supervised learning model that finds the hyperplane that best separates different classes in a high-dimensional space. Effective in high-dimensional spaces. (Zhou et al., 2016)] Robust against overfitting, especially in high-dimensional datasets. Limitation for Requires careful tuning of parameters and kernel selection. Computationally intensive for large datasets. The model was trained using a labeled Bangla news dataset, optimizing using categorical cross-entropy loss. The network used backpropagation through time (BPTT) to update weights based on the classification accuracy.Hyperparameters such as learning rate, batch size, and number of epochs were fine-tuned to improve performance.[Classification of Bangla News Articles using Bidirectional Long Short Term Memory (Shahin et al., 2020)]

Deep Learning Models

Recurrent Neural Networks (RNNs)RNNs are a class of neural networks designed for sequential data, making them suitable for text processing Capable of handling variable-length input sequences. Good for tasks where context from previous words is important Difficulty in learning long-range dependencies. Training can be slow and complex Long Short-Term Memory (LSTM)LSTMs are a type of RNN designed to

overcome the limitations of standard RNNs by introducing memory cells that can maintain information over long periods. Better at capturing long-term dependencies compared to standard RNNs. More robust to the vanishing gradient problem. More complex and computationally intensive than standard RNNs. Requires more data and tuning for optimal performance. Convolutional Neural Networks (CNNs)Originally developed for image processing, CNNs have been adapted for text classification tasks. CNNs use convolutional layers to detect local patterns in text, such as n-grams, and pooling layers to reduce dimensionality. They can capture spatial hierarchies in data. Effective at capturing local features and patterns in text. Can be trained quickly and efficiently. May miss broader context and relationships in text. Requires careful design of architecture and hyperparameters. An extension of Word2Vec, Doc2Vec generates vector representations for entire documents rather than just individual words. It uses a similar approach to Word2Vec but includes an additional paragraph vector that captures the context of the entire document. This allows it to learn semantic meanings at the document level. Captures semantic meaning of documents. Can represent variable-length texts as fixed-length vectors. Lacks the contextual understanding that models like BERT provide. Requires significant amounts of training data for optimal performance.

3. Transformer-Based Models

BERT (Bidirectional Encoder Representations from Transformers)BERT is a transformer-based model that processes text bidirectionally, allowing it to understand context more effectively. BERT is pre-trained on large corpora using masked language modeling and next sentence prediction tasks. It can be fine-tuned for specific tasks, such as news article categorization, using labeled datasets. Superior contextual understanding due to its bidirectional nature. Effective transfer learning capabilities, allowing it to adapt to various tasks with less data. Requires significant computational resources for training and fine-tuning. Can be complex to implement and optimize.

OpenAI GPT: A unidirectional transformer model primarily used for text generation tasks. It processes text from left to right, which limits its contextual understanding compared to BERT. XLNet: An extension of BERT that incorporates permutation-based training, allowing it to capture bidirectional context while maintaining the autoregressive nature of language modeling. RoBERTa: A robustly optimized version of BERT that improves performance by training on larger datasets and removing the next sentence prediction objective.

Summary

Each of these models has its strengths and weaknesses, making them suitable for different tasks and datasets. Traditional models like Naive Bayes and SVM are simpler and faster but may struggle with complex language. Deep learning models, particularly RNNs, LSTMs, and CNNs, offer more powerful representations but require more data and computational resources. Transformer-based models like BERT have set new standards in NLP by effectively capturing context and semantics, making them highly effective for news article categorization and other text classification tasks.

4. Introduction to BERT

BERT, or Bidirectional Encoder Representations from Transformers, is a transformer-based model that processes text bidirectionally. Unlike previous models that read text sequentially (left-to-right or right-to-left), BERT considers the entire context of a word by looking at both its left and right surroundings. This bidirectional approach allows BERT to capture nuanced meanings and relationships between words, making it particularly effective for various NLP tasks, including news article categorization.

5. How BERT Works

BERT is pre-trained on a large corpus of text using two primary tasks:
Masked Language Modeling (MLM): Randomly masks words in a sentence and trains the model to predict them based on their context. This helps BERT learn contextual relationships.
Next Sentence Prediction (NSP): Trains the model to predict whether a given sentence follows another, enhancing its understanding of sentence relationships.
After pre-training, BERT can be fine-tuned on specific tasks, such as news categorization, using labeled datasets. During fine-tuning, the model adjusts its weights to optimize performance for the specific classification task.

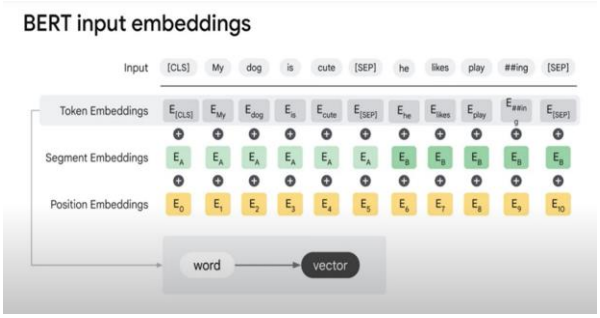

Fig 1(google skill booster)

6. Advantages of BERT Over Other Models

Contextual Understanding: BERT's bidirectional nature allows it to understand the context of words more effectively than unidirectional models. Transfer Learning: BERT can be fine-tuned on smaller datasets, making it adaptable to various tasks without requiring extensive labeled data. Handling Ambiguity: By considering the entire sentence, BERT can disambiguate words based on their context, improving classification accuracy.

7. Importance of BERT in News Article Categorization

Using BERT for news article categorization offers several advantages:

Improved Accuracy: BERT's ability to understand context leads to higher classification accuracy compared to traditional models.
Relevance in Recommendations: Enhanced categorization improves content recommendations, helping users find articles that match their interests.
Adaptability: BERT can quickly adapt to new topics and trends, making it suitable for the dynamic nature of news.

8. Conclusion

BERT has transformed the landscape of news article categorization by providing a powerful tool that captures the complexities of language. Its bidirectional approach and ability to learn from context make it superior to traditional and even many deep learning models. As the demand for accurate and relevant news categorization continues to grow, BERT stands out as a critical solution in the field of natural language processing.

9. References

[1] R. Johnson and T. Zhang, “Supervised and semi-supervised text categorization using lstm for region embeddings,” arXiv preprint arXiv:1602.02373, 2016.

[2] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, “Text classification improved by integrating bidirectional lstm with two-dimensional max pooling,” arXiv preprint arXiv:1611.06639, 2016.

[3] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016,pp. 1480–1489.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp. 5998–6008

[5] I. Schlag, P. Smolensky, R. Fernandez, N. Jojic, J. Schmidhuber, and J. Gao, “Enhancing the transformer with explicit relational encoding for math problem solving,” arXiv preprint arXiv:1910.06611, 2019

[6] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, “Text classification algorithms: A survey,” Information, vol. 10, no. 4, p. 150, 2019

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[8] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, “Deep unordered composition rivals syntactic methods for text classification,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1681–1691.

[9] Z. Wang, W. Hamza, and R. Florian, “Bilateral multi-perspective matching for natural language sentences,” arXiv preprint arXiv:1702.03814, 2017.

[10] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming auto-encoders,” in International conference on artificial neural networks. Springer, 2011, pp. 44–51

[11] Stein, A.J., Weerasinghe, J., Mancoridis, S., Greenstadt, R.: News article text
classi · cation and summary for authors and topics. Computer Science Information
Technology (CS IT) (2020

[12] Shahin, M.M.H., Ahmmed, T., Piyal, S.H., Shopon, M.: Classi · cation of bangla
news articles using bidirectional long short term memory. In: 2020 IEEE Region
10 Symposium (TENSYMP). pp. 1547–1551 (2020)

[13] Islam, MD Jahirul and Anas, Saeed Sarwar, News Classification Using Ml Algorithms: An Exploration of Efficient Information Categorization (2023). IJCRT | Volume 11, Issue 6 June 2023