International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET241126

Multiclass Document Classifier using BERT

Shruti A. Gadewar[1], Prof. P. H. Pawar[2]

[1]ME Aspirant, Department of Computer Science and Engineering, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

[2]Associate Professor, Department of Computer Science and Engineering, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 10 March 2024
Published: 28 March 2024

Publication Issue :
Volume 11, Issue 2
March-April-2024
Page Number :
106-111

ABSTRACT

With the rapid expansion of the internet, there has been an exponential surge in data volume, encompassing a myriad of documents laden with diverse types of information. This vast expanse includes structured and unstructured data, ranging from big data sets to formatted text and unformatted content. However, this abundance of unstructured data poses significant challenges in terms of effective management. Manual classification of this burgeoning data landscape is impractical, necessitating automated solutions. In this paper, we propose leveraging advanced machine learning techniques, particularly the BERT model, to classify documents based on contextual understanding, offering a more efficient and accurate approach to handling the data deluge.

Keywords :- Classification, BERT, Contextual Understanding

I. INTRODUCTION

The exponential growth of data across the internet landscape is undeniable, fueled by the constant generation of information from various sources such as Google Maps and diverse document uploads on numerous websites. This data mosaic comprises a plethora of content formats, including documents, posts, blogs, and videos, contributing to the vast digital repository. In this context, the focal point of this study is document data. Central to the extraction of meaningful insights is the initial step of organizing related data into cohesive classifications.

Classification, a cornerstone of data mining, involves categorizing data based on shared attributes or features. It operates as a supervised learning technique, utilizing labeled data to construct models capable of predicting the class of new, unseen data. Document classification, distinct from sentence classification, poses unique challenges due to the structural complexity of documents comprising multiple sentences with intricate semantic relationships. Managing a growing number of document categories further compounds the complexity. Automated document classification, a supervised machine learning approach, entails discerning whether a document pertains to a specific category by scrutinizing its vocabulary and comparing it against category-associated terms[2]. Moreover, as the number of categories expands, so does the

complexity of decision boundaries, posing challenges, especially in scenarios of imbalanced data[1].

In recent years, deep learning has garnered significant attention, showcasing remarkable achievements in various domains such as image processing, text analysis, and speech recognition. Text classification, a pivotal task in automating text processing, has seen widespread adoption of the long-short term memory (LSTM) mode[3]l, a recurrent neural network (RNN) adept at learning long-term dependencies. However, LSTM models suffer from memory and processing inefficiencies, particularly when confronted with substantial temporal gaps between data points.

To address these limitations, the transformer model emerged as a viable alternative, eschewing the recurrent approach in favor of multi-head self-attention mechanisms. Unlike its predecessors, transformers offer swift and efficient processing capabilities, making them well-suited for handling today's data density. In this project, we leverage BERT (Bidirectional Encoder Representations from Transformers) for multiclass document classification, harnessing its transformative capabilities to navigate the complexities of document analysis with enhanced speed and accuracy.

II. RELATED WORK

Yin et al. conducted a comparative study using convolutional neural network (CNN) and RNN for natural language processing (NLP) [4]. It was observed that the RNN model performed well for NLP and it was found to be more efficient against CNN. However, it was noted that the performance of the RNN model will be degraded when there is a keyword recognition task. They concluded that optimization of hidden size and batch size is very important and affect the performance of the prediction.

Lai et al. proposed a recurrent convolutional neural network (RCNN) model, which is a good method for effectively constructing sentence representations [5]. It was mentioned that the processing of an RNN model would take a lot of time depending on the size of the sentence or document used. Therefore, they proposed RCNN and applied it to the text classification task. Experiments were performed on four different data sets with many models including RNN, CNN and RCNN. It was stated that RCNN methods should be used to obtain more effective and successful results for text classification.\

Several deep learning models for binary sentiment classification were investigated by Ay Karakus et al [6]. Movie reviews on the website www.beyazperde.com were collected to train deep learning models. In this study, CNN and LSTM were used, various variants of these models were created and improvements were made by changing the number of layers and adjusting the hyperparameters. The authors also reported a detailed comparison of the models in terms of accuracy and time performance.

Besides the deep learning models, the quality of word embedding is also an important element in text classification. In [7], the authors targeted to improve word2vec word embeddings by automatically optimizing 5 different hyperparameters. They presented two approaches for tuning hyperparameters to beat grid search and random search. Word embeddings were generated with documents of approximately 300 million words. They conducted experiments to demonstrate how to create quality word embedding using CNN classification models on documents belonging to 10 different classes. It was observed that hyperparameter optimization alone improved classification success by 9%.

Recently, transformer models have attracted attention for text classification and promised superior performance. Transformer has the potential to produce more accurate results than traditional models without disturbing the structure of the text. Gong et al. proposed a new Hierarchical Graph Transformer - based deep learning model for large-scale multi- label

text classification [8]. It is stated that the proposed model can better perceive the features of the text and focus attention in the right direction by using multi-headed and multi-layered structures. Labels and meaningful distances between these labels were added to the model and a hierarchical relationship was established between them. Thus, the model was expected to accurately determine the hierarchy and logic of the text. Extensive experiments on datasets have shown that the model can capture the hierarchy and logic of text and improve performance compared to state-of-the-art methods.

III. TRANSFORMER ARCHITECTURE

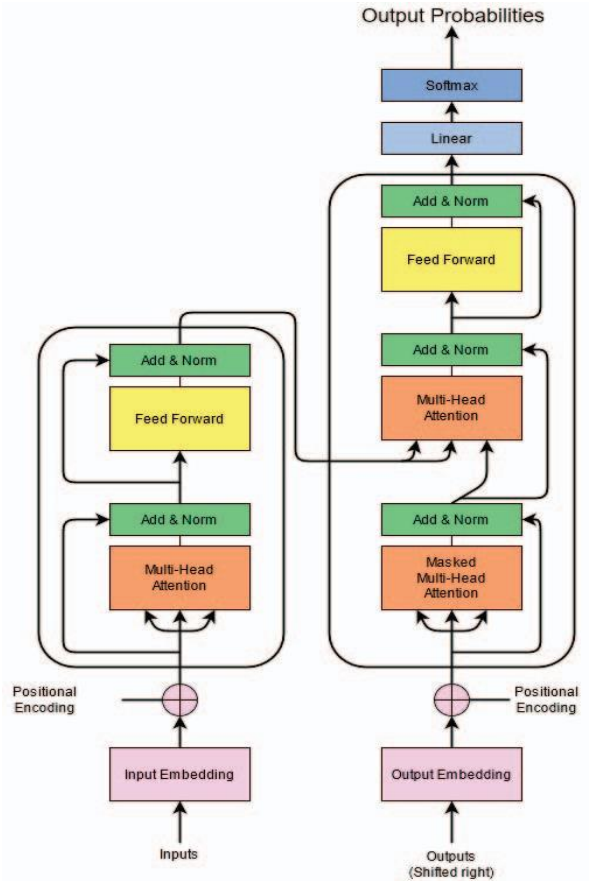

Fig. 1. The Transformer architecture [11]

Sequence-to-sequence models, whose input and output are variable-length arrays, have a two-component architecture. The first component is an encoder that takes a variable-length array as input and converts it to a fixed-shape state. The second component is a decoder that maps the encoded state of a fixed shape to a variable-length array. The encoder takes a symbol representation input sequence and maps it to a continuous output sequence. Given the input array, the decoder returns the output array of symbols, one item at a time. It consumes pre-generated symbols as additional input so that it can produce outputs in other steps. The transformer encoder-decoder architecture is very powerful. It has led to many successful NLP models such as BERT [9] and GPT-3 [10]. The transformer has the encoder-decoder architecture shown in Fig. 1

The encoder part can be used for successful text classification. It has been shown to be a superior approach to the relatively slow RNN models. We see transformer models in Reinforcement Learning [12] as well as in supervised learning such as text classification.

IV. PROBLEM STATEMENT

As we know the large amount of data is generated every day over the internet. This data may also in the form of documents. The very first step of deriving information is to have related data classified and clustered at one place.

Classification is a data mining technique which is a process of classifying or categorizing data on the basis of similar features or attributes. It is a supervised learning technique that uses labelled data to build a model that can predict the class of new, unseen data.

Document classification is structurally different from sentence classification. Documents consist of multiple sentences. Sentences may have ambiguous and complex semantic relationships, which makes it difficult to classify documents. In addition, as the number of document categories increases, their management becomes more difficult.

So this paper is discussing a classifier which is capable of accurately categorizing text documents into multiple predefined classes or categories using BERT

(Bidirectional Encoder Representations from Transformers) model.

V. PROPOSED SYSTEM

Data Collection: Gather a diverse dataset of documents spanning multiple classes/categories.
Preprocessing: Clean and preprocess the documents (e.g., remove stop words, punctuation, special characters).
Tokenization: Tokenize the preprocessed documents into word tokens.
BERT Embedding Extraction: Use a pre-trained BERT model to extract embeddings (dense vector representations) for each token in the documents.
Pooling: Aggregate the token embeddings using pooling techniques (e.g., mean pooling, max pooling) to obtain a fixed-length representation for each document.
Classification Head: Add a classification head on top of the pooled embeddings to perform multiclass classification. This typically consists of one or more dense layers followed by a softmax activation function to produce class probabilities.
Loss Function: Define a suitable loss function (e.g., categorical cross-entropy) to measure the difference between predicted probabilities and ground truth labels.
Training: Train the model using the labeled training data. This involves optimizing the parameters of the entire model using backpropagation and a suitable optimization algorithm (e.g., Adam).
Evaluation: Evaluate the trained model on a separate validation dataset to measure its performance in terms of accuracy, precision, recall, and F1-score.

Inference: Once the model is trained and evaluated, it can be used for inference on new, unseen documents. The model will output predicted class probabilities for each document.

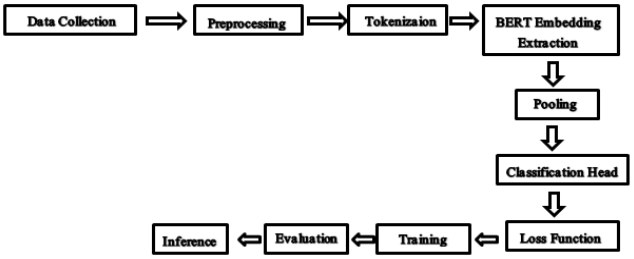

Fig. 2:- Proposed System

VI. CONCLUSION

In this paper, we have explored the challenges associated with the management and classification of the burgeoning volume of unstructured data on the internet, particularly focusing on document classification. Traditional methods of manual classification are impractical given the scale of data, necessitating automated solutions. Leveraging advanced machine learning techniques, particularly the BERT model, we proposed a framework for document classification based on contextual understanding.

By employing BERT, a state-of-the-art transformer model, we addressed the complexities of document analysis more efficiently and accurately. Through a comprehensive review of related works, we highlighted the evolution of deep learning models in text classification and emphasized the superiority of transformer-based architectures over traditional models like LSTM and CNN.

The proposed system involves several key steps, including data collection, preprocessing, tokenization, BERT embedding extraction, pooling, classification head addition, loss function definition, training, evaluation, and inference. By following this systematic approach, we demonstrated the feasibility and effectiveness of utilizing BERT for multiclass document classification tasks.

VII. FUTURE SCOPE

While our proposed system showcases promising results, there are several avenues for future research and development:

Fine Tuning BERT: Further exploration of fine-tuning techniques for BERT to optimize its performance specifically for document classification tasks could enhance model accuracy and efficiency.

Handling Imbalanced Data: Investigating methods to address imbalanced datasets, which are common in real-world scenarios, would be beneficial to improve the robustness of the classification model.

Integration of Domain Specific Knowledge: Incorporating domain-specific knowledge or ontologies into the classification process could improve the model's understanding of document semantics and enhance classification accuracy.

Exploration of Multimodel Approaches: Considering the multimodal nature of documents, exploring approaches that combine textual information with other modalities such as images or metadata could lead to more comprehensive document understanding and classification.

Deployment and Scalability: Streamlining the deployment process of the model and ensuring scalability to handle large volumes of documents in real-time environments would be essential for practical applications.

By addressing these areas of research, we can further advance the capabilities of document classification systems, making them more accurate, efficient, and adaptable to diverse real-world scenarios.

VIII. REFERENCES

[1]. Ilkay Yelmen, Ali Gunes, and Metin Zontul on "Multi-Class Document Classification Using Lexical Ontology-Based Deep Learning" *Appl. Sci.* 2023, *13*(10), 6139; https://doi.org/10.3390/app13106139

[2]. Kadhim, A.I. Survey on supervised machine learning techniques for automatic text classification. *Artif. Intell. Rev.* 2019, *52*, 273–292. [Google Scholar] [CrossRef]

[3]. L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in Fifteenth annual conference of the international speech communication association, 2014

[4]. W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," arXiv preprint arXiv:1702.01923, 2017.

[5]. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in Twenty-ninth AAAI conference on artificial intelligence, 2015.

[6]. B. Ay Karakuş, M. Talo, İ. R. Hallaç, and G. Aydin, "Evaluating deep learning models for sentiment classification," Concurrency and Computation: Practice and Experience, vol. 30, no. 21, p. e4783, 2018.

[7]. B. Yildiz and M. Tezgider, "Improving word embedding quality with innovative automated approaches to hyperparameters," Concurrency and Computation: Practice and Experience, p. e6091, 2021. doi: https://doi.org/10.1002/cpe.6091.

[8]. J. Gong et al., "Hierarchical graph transformer-based deep learning model for large-scale multi-label text classification," IEEE Access, vol. 8, pp. 30885-30896, 2020.

[9]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre- training of deep bidirectional transformers for language

understanding " presented at the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA., 2019.

[10]. T. B. Brown et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020.

[11]. A. Vaswani et al., "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998-6008.

[12]. B. Yildiz, "Reinforcement learning using fully connected, attention, and transformer models in knapsack problem solving," Concurrency and Computation: Practice and Experience, e6509, 2021. doi: https://doi.org/10.1002/cpe.6509