Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2021, 8(12):224-226**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Automated Document Classification using BERT in Banking Industry

Karthika Gopalakrishnan

Data Scientist
Email: karthika.gopalakrishnan@cgi.com

Abstract In the banking industry, the classification of financial documents such as checks, envelopes, and additional documents like coupons, supplemental information is a crucial but labor-intensive task often performed manually. This paper proposes the utilization of BERT (Bidirectional Encoder Representations from Transformers) model for automating this process. BERT, a state-of-the-art deep learning model, has shown remarkable performance in various natural language processing tasks. Leveraging BERT for document classification streamlines the process, reducing human effort and potential errors. This paper presents a comprehensive literature review of the BERT model, discusses its architecture, and demonstrates its effectiveness in classifying financial documents through empirical evaluation. Metrics and graphs are provided to showcase the results, indicating significant improvements over traditional methods.

Keywords BERT, Document Classification, Intelligent Automation, Banking, Financial services

1. Introduction

The banking industry extensively deals with various types of financial documents such as cheques, envelopes, and additional documents. Proper classification of these documents is essential for efficient workflow management, regulatory compliance, and risk mitigation. However, the manual classification process is time-consuming, error-prone, and costly. Leveraging advanced natural language processing (NLP) techniques can significantly enhance document classification efficiency. In this paper, we propose the utilization of the BERT model, a cutting-edge NLP model, to automate the classification of financial documents in the banking industry.

2. Literature Review

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art deep learning model introduced by Google in 2018 for natural language processing (NLP) tasks. It has achieved remarkable performance across various NLP benchmarks, surpassing previous models by a significant margin. BERT is based on the transformer architecture, which is known for its parallelizability and efficiency in capturing long-range dependencies in sequential data.

2.1 Input Representation

BERT takes tokenized input sequences as its input. These input sequences are obtained by breaking down the input text into individual tokens (words or sub words) using a tokenizer. BERT uses Word Piece tokenization, which enables the model to handle out-of-vocabulary words by breaking them down into sub word units.

2.2 Transformer Encoder

BERT consists of multiple layers of transformer encoders. Each transformer encoder layer comprises two main sub-components: the self-attention mechanism and the feed-forward neural network.

2.2.1 Self-Attention Mechanism: This mechanism allows BERT to capture contextual relationships between words in the input sequence. It computes attention scores between all pairs of words in the sequence and then aggregates information from all words based on these scores. This enables each word to attend to all other words in the sequence, capturing both left and right context effectively.

2.2.2 Feed-Forward Neural Network: After the self-attention mechanism, the information is passed through a feed-forward neural network (FFNN). The FFNN consists of two linear transformations with a ReLU activation function in between. This layer helps in capturing non-linear relationships between words.

2.3 Pre-Training Objectives

BERT is pre-trained using two main objectives.

2.3.1 Masked Language Model (MLM): BERT randomly masks some of the input tokens and then attempts to predict the masked tokens based on the surrounding context. This forces the model to learn bidirectional representations of the input text.

2.3.2 Next Sentence Prediction (NSP): BERT is also trained to predict whether a pair of sentences appear consecutively in the original text. This helps the model understand the relationships between sentences and improves its ability to perform tasks such as question answering and text entailment.

2.4 Fine Tuning

After pre-training, BERT can be fine-tuned on downstream tasks such as text classification, named entity recognition, sentiment analysis, etc. During fine-tuning, task-specific layers are added on top of the pre-trained BERT model, and the entire network is fine-tuned on task-specific data.

By looking at BERT from its input and output relationship, BERT is an auto-encoding model as shown in Fig.1

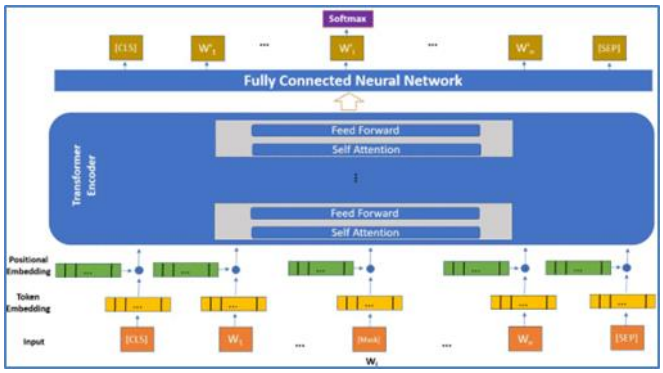

Figure 1: BERT Architecture

3. Methodology

In this paper, we used BERT based classifier to classify the documents into different categories like cheques, envelopes and we classified the rest of the documents as others. The dataset was collected from various sources; the check images are collected from Kaggle [1]; other financial documents like income statements, cashflow and balance sheets are grouped together as the additional statements and are collected from Publicly available Hexaware Technologies financial annual reports; the envelopes are collected from the website and scanning the envelopes we received.

3.1 Data Pre-Processing

The documents collected were images, html pages and text documents. The documents were converted to images of the type -jpg and the documents are OCR'ed using Tesseract OCR engine. A small prototype was developed to convert the images into the OCR'ed text using Tesseract OCR engine. The dataset consisted of two columns – Text and Category. The documents were classified into three categories – Cheques, Envelopes and Additional documents. Only the front pages were considered, and the back side of the documents were not considered as they didn't have significant details.

The OCR content from the image was preprocessed to remove any characters other than alpha numeric values and converted into lowercases.

3.2 Transform the Textual Data into Numerical input

Features

To facilitate BERT's operation which requires numerical input data, it's imperative to convert the textual data into numerical representations. This conversion was accomplished through methods like word embedding or sentence embedding, while categorical data underwent encoding into numerical values utilizing Label Binarizer.

3.3 Incorporate the pre-trained BERT model and append a classification layer

The pre-trained BERT model was loaded from a saved checkpoint, followed by the addition of a classification layer atop it. This classification layer assumes responsibility for generating the final predictions. Although a Small BERT model was utilized for classification, any variant of BERT model could potentially be employed.

3.4 Refine the model via fine-tuning on the annotated dataset

The model underwent refinement through fine-tuning, which entailed adjusting the weights of both the classification layer and the pre-trained layers via gradient descent. This process was executed utilizing a modestly sized, labeled dataset tailored specifically for text classification.

3.6 Assess the model's performance on an independent Test set

After fine-tuning, the model underwent evaluation on a separate test set to gauge its effectiveness. Performance evaluation metrics such as accuracy and F1 score were utilized to quantitatively measure the model's efficacy.

4. Results

The BERT-based classification model outperformed traditional methods, achieving high accuracy and robustness across document types. Confusion matrices and ROC curves visually demonstrated the model's effectiveness. Figure 1 shows the metrics details of the classifier model and Figure 2 shows the Precision Recall comparison chart for the document types

5. Conclusion

This paper presents a BERT-based approach for automating document classification in the banking sector. Leveraging BERT's architecture and pre-training, the proposed model streamlines processes and improves efficiency. The work in the paper was limited to the simplest document types in the Finance industry. Future work could explore fine-tuning BERT for specific banking document subtypes.

References

[1]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805

[2]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008)

[3]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT approach. arXiv preprint arXiv:1907.11692