DSTI

Data ScienceTech Institute

TECHNICAL REPORT

ON

DEEP LEARNING CLASSIFICATION OF NEWS TEXT

Written By: Jean-Luc Boa Thiemele (DS A22)

Deogratias Allakonon (DS A22)

Clement Adebisi (DE A22)

Course: DEEP LEARNING

Submitted To (Professor): Hanna Abi Hakil

Abstract

This report delves into the implementation of a deep learning model, specifically utilizing BERT (Bidirectional Encoder Representations from Transformers), for classifying news text into categories. Given the large volume of daily news, automatic classification allows for streamlined organization, improved searchability, and better content recommendations. The goal of this report is to evaluate the effectiveness of deep learning for text classification within the news domain and to document the process, results, and conclusions drawn from the project.

I. INTRODUCTION

Text classification has gained significant attention due to the growing need to process and categorize large volumes of unstructured data. In the context of news, categorizing articles by topic can support improved content management, facilitate audience targeting, and enhance user experience. Traditional text classification techniques (e.g., Naive Bayes, Support Vector Machines) often rely on bag-of-words or TF-IDF representations, which capture limited context. Deep learning, particularly using pre-trained models like BERT, leverages context more effectively through attention mechanisms, providing improved accuracy and versatility.

This report outlines a pipeline that uses the BERT model to classify news text, detailing the dataset, model selection, architecture, implementation, results, and evaluation metrics.

II. DATASET

The dataset used for this project was obtained from Kaggle and comprises a corpus of news articles classified into seven categories: technology, sports, world, politics, entertainment, automobile, and science. Each entry in the dataset provides essential components for a text classification task, including columns for the article's headline (news_headline), main body (news_article), and category label (labels).

The dataset contains **12,120 entries**. Although relatively small for deep learning tasks, it provides a solid foundation for this project. To enhance model performance, strategies like **transfer learning** were used to make the most of the available data.

II-1. Columns Overview

- **news_headline**: Headline of the news article.
- **news_article**: Main body text of the news article.
- **labels**: Category label indicating the article's topic.

II-2. Data Processing Steps

- **Concatenation of CSV Files**:
  - The dataset consists of seven separate CSV files, each representing articles from one of the defined categories. These files were combined into a single DataFrame to streamline the classification process. This concatenation ensures that all relevant articles are accessible in one unified dataset, facilitating easier data processing and analysis.
- **New Column Creation**:
  - A new column named **text** was created, which concatenates the **news_headline** and **news_article** fields. This consolidated text is used for classification purposes, ensuring that the model has access to both the headline and body content during training.
- **Text Cleaning**: Initial cleaning was applied to ensure uniformity across entries. This included:
  - Removing any leading or trailing spaces.
  - Replacing multiple spaces with a single space.
- **Label Processing**:

  - Unique labels were extracted from the dataset's labels column, and a mapping was created between labels and IDs (label2id, id2label) to streamline encoding and decoding.
  - A ClassLabel object was defined using the unique labels, enabling a stratified train-test split and handling any imbalances in label distribution.
- **Tokenization**:
  - The BERT tokenizer associated with the model checkpoint ("bert-base-cased") was loaded to preprocess the text data.
  - Tokenization was applied to the consolidated text column to prepare the data for model ingestion. Processing was performed in batches to expedite tokenization and support variable text lengths efficiently.

By following these steps, we prepared the dataset to be compatible with the deep learning model architecture.

III. BERT MODEL

The BERT model, developed by Google, represents a breakthrough in NLP, leveraging a transformer architecture to capture bidirectional context. It is pre-trained on a large corpus of text and can be fine-tuned for various downstream tasks, including text classification.

III-1. BERT Architecture Overview

The BERT (Bidirectional Encoder Representations from Transformers) model, based on the Transformer architecture, achieves a sophisticated understanding of text by using bidirectional processing, self-attention mechanisms, and multiple transformer layers. Transformers are a breakthrough in NLP, as they capture dependencies between words across long sequences more effectively than traditional recurrent neural networks (RNNs) or convolutional neural networks (CNNs).

1. The Transformer Architecture

The Transformer is a neural network architecture that fundamentally changes how models process language. Unlike RNNs, which handle sequences by processing each word in order, the Transformer processes all words in a sentence simultaneously. This parallel processing is achieved using self-attention mechanisms, which allow the model to weigh the relevance of each word relative to others, regardless of its position in the sentence.

2. BERT - Bidirectional Context

Traditional NLP models process text in a unidirectional manner—typically left-to-right or right-to-left—meaning they only have access to the preceding or following words when encoding a given token. BERT, however, processes text bidirectionally, which means it reads the entire sequence both forwards and backwards simultaneously. This approach allows BERT to develop a nuanced understanding of each word in context, as it considers both prior and succeeding words within a sentence. For instance, in the phrase "bank deposit," BERT can distinguish if "bank" refers to a financial institution or a riverbank based on the neighboring words.

3. BERT - Self-Attention Mechanism

The attention mechanism, specifically self-attention, is fundamental to BERT’s ability to weigh the importance of each word relative to others in the sentence. Self-attention dynamically assesses how each word in a sentence should be weighted when encoding another word. For example, in a sentence like “The dog chased the cat because it was scared,” BERT uses self-attention to understand that "it" likely refers to "the cat" and not "the dog" due to context. Self-attention layers are particularly powerful for handling long-range dependencies in language, allowing BERT to create richer word embeddings that capture these relationships across even complex sentences.

4. BERT - Transformer Layers

BERT's core architecture consists of a series of stacked transformer blocks that iteratively encode the input data. BERT-base has 12 transformer layers, and each transformer layer contains:

- **Multi-Head Attention Sub-layers**: Multiple self-attention heads operate in parallel to capture different types of relationships between words. Each head focuses on distinct contextual information, allowing BERT to form a comprehensive representation of the text.
- **Feed-Forward Networks**: After multi-head attention, each word representation passes through a fully connected feed-forward network to enable more complex transformations and feature interactions.

5. Positional Encoding

Since transformers do not inherently process word order, BERT adds positional encoding to each token embedding to maintain sequential information. This encoding allows BERT to differentiate between words based on their order in the sentence, helping it capture syntactical nuances that are critical for accurately understanding context.

6. Pre-training with Masked Language Modeling (MLM)

BERT is pre-trained using a masked language modeling objective, where 15% of words in a sentence are randomly masked, and the model is trained to predict these masked tokens based on the surrounding context. This strategy enables BERT to learn bidirectional context and understand word relationships in a way that generalizes well to a variety of NLP tasks, including text classification.

7. Next Sentence Prediction (NSP)

In addition to MLM, BERT is pre-trained on a Next Sentence Prediction (NSP) task, designed to help the model understand relationships between sentences. During training, BERT is provided with pairs of sentences:

- **Positive pairs** where the second sentence logically follows the first.
- **Negative pairs** where the second sentence is randomly selected and does not logically follow the first.

The model learns to predict whether a sentence pair follows a coherent sequence. This task enhances BERT's performance on sentence-level understanding, which is critical for applications like question-answering, information retrieval, and coherence detection.

8. Transfer Learning and Fine-Tuning

After pre-training on tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), BERT can be fine-tuned on our specific news classification dataset to adapt its deep, generalized language understanding to the unique nuances of classifying news topics.

In this project, BERT is fine-tuned on a dataset with seven defined categories: technology, sports, world, politics, entertainment, automobile, and science. This fine-tuning process enables BERT to effectively learn category-specific patterns within our news data. For example, terms frequently associated with technology or entertainment are identified by BERT and weighted more significantly when determining the category, helping it distinguish between topics accurately.

Through this transfer learning approach, BERT's architecture optimizes its performance for our classification application, leveraging its pre-trained bidirectional context and attention mechanisms to capture both word-level and topic-level distinctions. This fine-tuning process enhances BERT's ability to categorize news text by refining its general language representations to align closely with our task, ultimately enabling high accuracy and reliability in classifying the various news categories.

IV. RESULTS

Results were evaluated using standard classification metrics. The results of the training and evaluation process of our BERT-based model for news text classification are summarized in the metrics presented below. These metrics provide insights into the model's performance across various dimensions, highlighting its effectiveness in classifying news articles into the predefined categories.

Metric	Value
Accuracy	**0.9394**
Precision	**0.939**
Recall	**0.9394**
F1 Score	**0.9390**

Accuracy

The model achieved an impressive evaluation accuracy of 93.94%. This high accuracy rate indicates that the vast majority of the predictions made by the model align with the true labels. In the context of news classification, such accuracy demonstrates the model's capability to distinguish among the various categories (technology, sports, world, politics, entertainment, automobile, and science) effectively.

Precision

With a precision value of 93.9%, the model shows a strong ability to correctly identify positive instances among all instances it predicted as positive. In practical terms, this means that when the model predicts an article to belong to a certain category, it is highly likely to be correct.

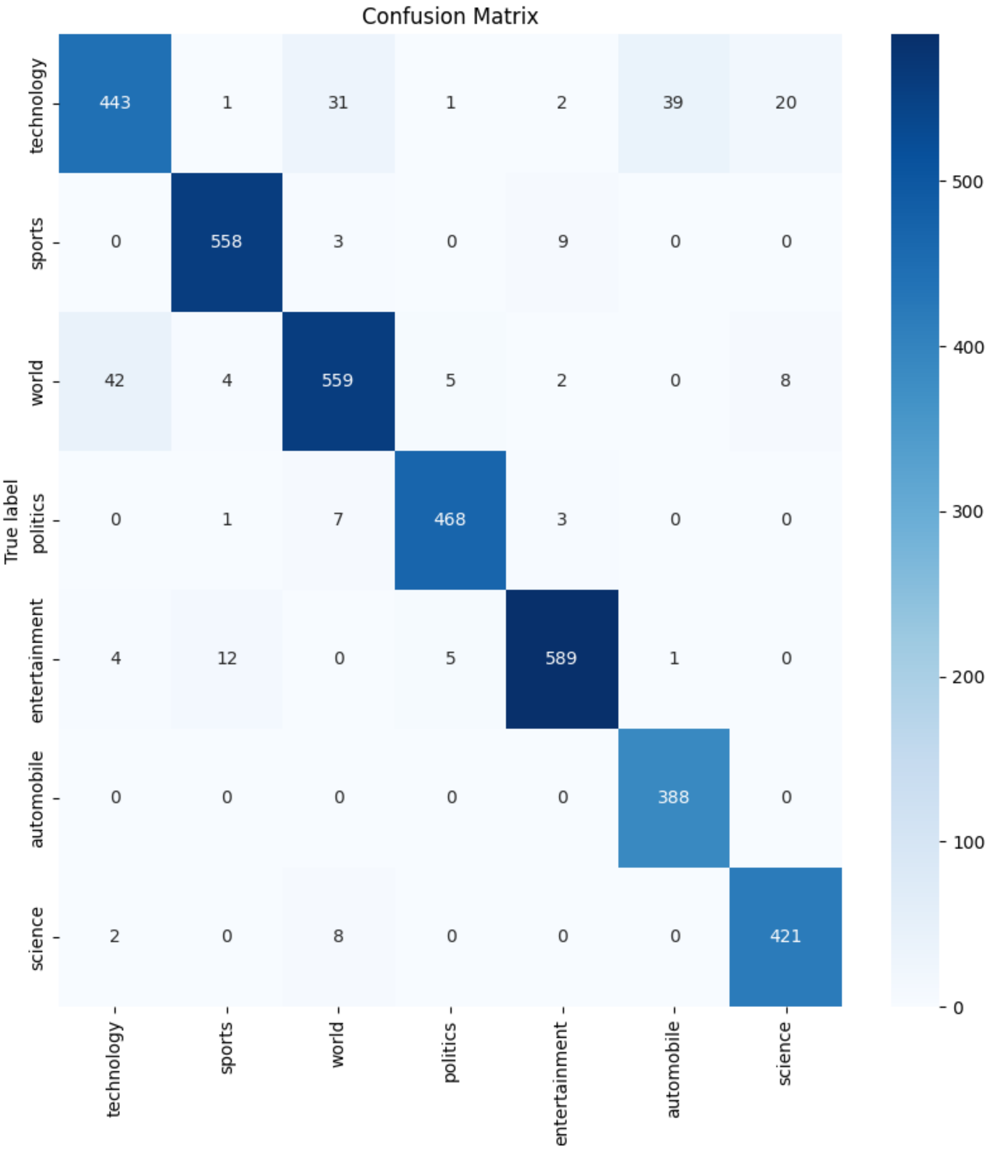
Recall

The recall metric of 93.94% indicates that the model is also effective at capturing the majority of actual positive instances in the dataset. This implies that the model can identify most articles that truly belong to a given category, thus minimizing missed opportunities for classification.

F1 Score

The F1 score, calculated at 93.90%, provides a balanced measure of the model's precision and recall. It indicates that the model maintains a solid balance between precision and recall, reinforcing its reliability in categorizing news articles accurately.

Confusion Matrix



True label	technology	sports	world	politics	entertainment	automobile	science
technology	443	1	31	1	2	39	20
sports	0	558	3	0	9	0	0
world	42	4	559	5	2	0	8
politics	0	1	7	468	3	0	0
entertainment	4	12	0	5	589	1	0
automobile	0	0	0	0	0	388	0
science	2	0	8	0	0	0	421

The confusion matrix reveals a promising performance by the model in classifying news articles across various categories, evidenced by a strong diagonal that indicates high true positive rates for most classes. Notably, the sports, world, politics, entertainment, and automobile categories exhibit impressive accuracy, with minimal false positives and negatives. This suggests that the model effectively differentiates between these areas, successfully classifying articles without significant errors.

However, the technology category stands out as a point of concern, exhibiting the highest number of misclassifications. Specifically, 94 articles were incorrectly identified as belonging to other classes, particularly world, automobile and science. This indicates a potential overlap in content or insufficient features that hinder the model’s ability to distinguish technology articles accurately. Such misclassifications could affect the user experience and the overall effectiveness of the classification system.

Conversely, the automobile class demonstrates exceptional performance, with no misclassifications recorded. This highlights the model's ability to accurately classify articles within this categorie, reflecting a clear understanding of the content and context associated with it.

V. CONCLUSION

This report has examined the implementation of a BERT-based deep learning model for classifying news articles into predefined categories. The results demonstrate the model's strong performance, achieving an evaluation accuracy of 93.94%, with corresponding precision, recall, and F1 scores all around 93%. Such metrics indicate that the model is not only effective in accurately categorizing articles but also in minimizing errors associated with misclassification, thus enhancing the user experience and improving content management.

The analysis of the confusion matrix further highlights the model's strengths, particularly in categories such as sports, world, politics, entertainment, and automobile, where it exhibited minimal false positives and negatives. However, the technology category presented challenges, with a notable number of misclassifications. This suggests potential overlaps in content and the necessity for more nuanced feature extraction techniques.

To improve the overall classification performance, especially for the technology category, it is recommended to augment the training dataset with additional examples. Exploring advanced feature extraction methods could further enhance the model's ability to differentiate between nuanced content in various categories.

In summary, the findings of this project affirm the efficacy of deep learning approaches, particularly BERT, for news text classification. The results not only validate the use of modern NLP techniques but also provide a clear roadmap for future enhancements and optimizations in text classification tasks. With continuous improvement, the model holds significant promise for applications in automated news categorization, ultimately contributing to more efficient information retrieval and enhanced user engagement in the news domain.