UNIVERSITY POLITEHNICA OF BUCHAREST
FACULTY OF AUTOMATIC CONTROL AND COMPUTERS
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

Fake News Detection using a Multi-Modal Dataset

Matei Vlad Cristian

Supervisor:

Dr. Ing. Claudiu Ifrim

BUCHAREST
2023

1 INTRODUCTION

In the fast-changing world of smart computers, scientists are working hard to make machines think and learn like humans. They've created fancy models inspired by the complexities of the human brain. From the foundational perceptron to the transformative capabilities of transformers, the journey through artificial neural networks has been marked by continuous innovation.

This article delves into the realm of multimodal artificial intelligence, exploring the fusion of text and image modalities. In the modern digital era, where information is not confined to text alone, the proposed supervised multimodal bitransformer emerges as a powerful solution. Moreover, in a world of big data, the necessity for robust fake news detection becomes crucial.

In presenting our solution, we focus on the approach of detecting fake news using a dataset centered around Covid-19-related information, which can be classified into six categories: *Satirical news*, *Real news*, *Propagandistic news*, *Plausible news*, *Fictional news*, and *Fake news*.

The two primary models underlying our solution are based on transformers, namely BERT and MultiModal BiTransformers. Through this approach, we aim to explore the effectiveness of these models in detecting fake news, with the central theme revolving around the Covid-19 pandemic context.

2 BACKGROUND

2.1 Transformers

The Transformer [7], introduced by researchers at Google, revolutionized the sequential nature of text processing by adopting an **attention mechanism** proposed in [5]. The most significant advantage brought by Transformers is parallelization, with the model being trained in just 3.5 days on 8 GPUs.

Architecture

Transformers are based on an encoder-decoder structure, with the encoder mapping a series of symbols $(x_1, x_2, ..., x_n)$ into a sequence of representations $(z_1, z_2, ..., z_n)$. This sequence is then processed by the decoder, generating the output sequence step by step $(y_1, ..., y_m)$.

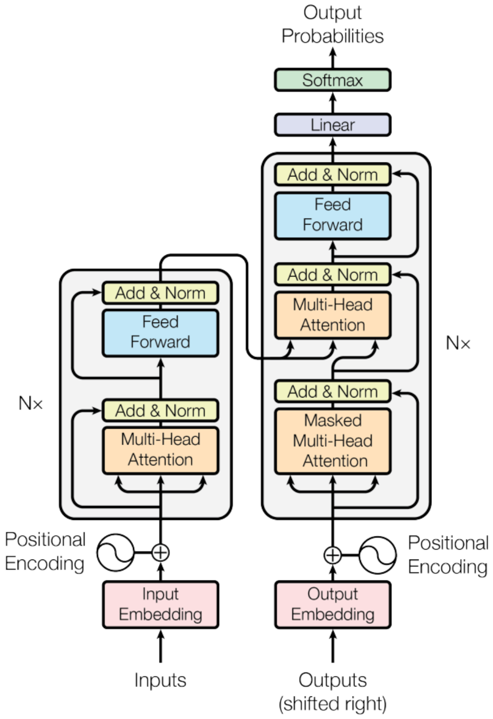

Figura 2.6 Transformers Arhitecture [7]

Encoder: It consists of 6 stacked units, each unit containing two layers—an attention mechanism with multiple heads and a classic feedforward network. Each layer benefits from a residual connection [3] to maintain more stable gradients, and the result of their addition is then normalized.

Decoder: It is also composed of 6 stacked units, each containing the same two layers—an attention mechanism with multiple heads and a classic feedforward network. However, the decoder additionally includes a third layer that applies attention mechanism over the outputs received from the encoder. Residual connections are also present within these units.

Attention: Performs a mapping that can be visualized as a query and a set of key-value pairs. All these elements are vectors, as each word is represented as a vector using word embeddings.

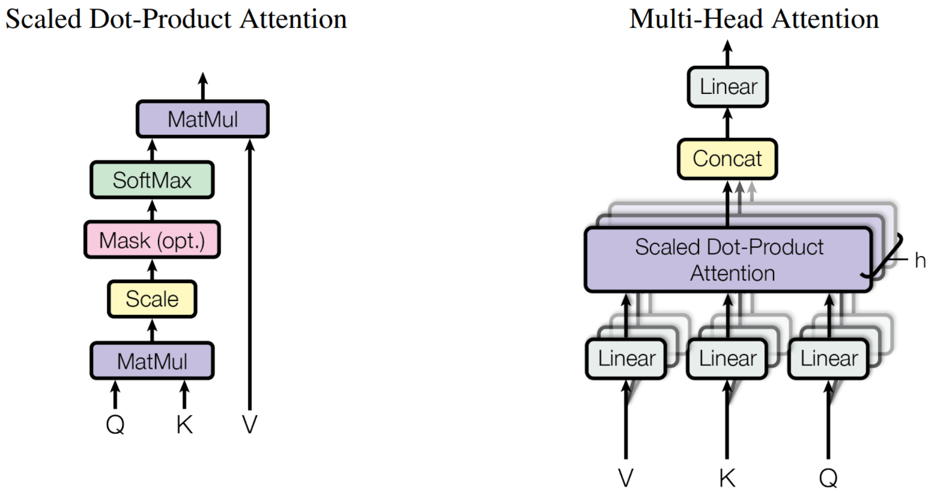

Figura 2.7 Attention Mechanism [7]

Scaled Dot-Product Attention For each word, treated as the query, it explores similarities with all other words in the context through key-value pairs. The score is computed based on the value, initially representing word encodings through word embeddings.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V, \tag{1}$$

where key and query have dimension d_k, and value d_v. The division by $\sqrt{d_k}$ is performed because, at larger numbers, the softmax function can lead gradients to very small values. **Multi-Head Attention Mechanism** involves processing encoded words, splitting into eight attention heads, obtaining Q, K, and V matrices, calculating scores using Formula 13, concatenating results, and multiplying by W_O to produce the final matrix Z. This matrix is then fed into the feedforward network for the ultimate encoding choice.

Positional Encoding Since there is no longer recurrence, the model needs to be able to calculate the original order of words in a sequence. To achieve this, positional information is introduced through a formula for determining the position.

$$PE_{(pos.2i)} = \sin(pos/10000^{2i/d_{model}}) \tag{2}$$

$$PE_{(pos.2i+1)} = \cos(pos/10000^{2i/d_{model}}) \tag{3}$$

2.2 Pre-training BERT

Bidirectional Encoder Representations from Transformers [1] represents a significant breakthrough in NLP, achieving outstanding performance. From recurrent neural networks, the field progressed to LSTM as a better solution for the Vanishing Gradients problem. Then came ELMo, which used bidirectional LSTM for contextualizing encodings. Finally, the development of the attention mechanism led to the emergence of the Transformers architecture, expanding the concept of multi-head attention.

BERT can solve a variety of tasks and can be trained for specific requirements. The two complete implementation phases of BERT are **pre-training** and **fine-tuning**.

2.2.1 Arhitecture

All encoder modules will compute representations and pass the information up to the final layer, which will provide the probability distribution for the next prediction.

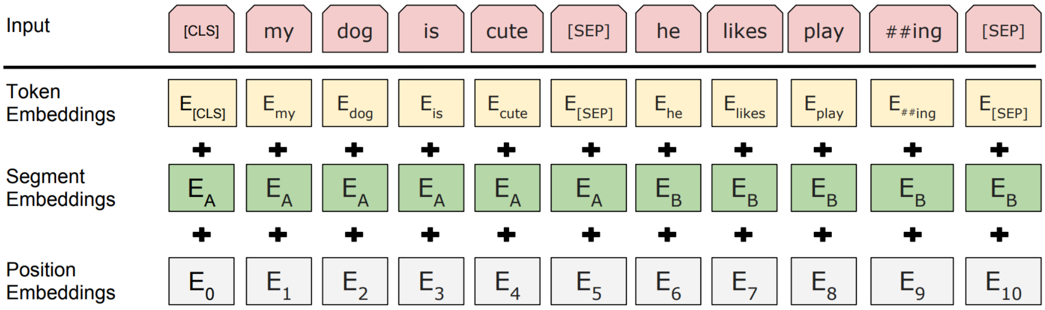

Figura 2.8 BERT Arhitecture [1]

To address various tasks BERT can frame one or two sentences from the text in a symbolic representation. The symbols $[CLS]$ and $[SEP]$ mark the beginning of a sequence and separate encodings for two sentences, respectively. The input representation is defined by the sum of word encodings, segment encodings, and positional encodings, as illustrated in Figure 2.14. The text is indexed, and symbols are assigned to sentences, capturing this information in the encoding process.

BERT is pretrained with the goal of addressing two main tasks:

Masked Language Modeling (MLM): A language modeling system is desirable to benefit from context both from the past and the future. However, achieving bidirectionality in standard language models would allow the network to cheat by reading ahead the word to be predicted. BERT circumvents this issue by employing masks. Hidden words are replaced with the symbol $[MASK]$. During the fine-tuning phase, 15% of words are randomly selected for masking. Out of these, 80% are replaced with $[MASK]$, 10% with a random symbol, and with a 10% chance, the word remains unchanged.

Next Sentence Prediction (NSP): This task aims to better understand relationships between sentences. Two sentences, A and B, are chosen, and with a 50% chance, B is the actual next sentence after A; otherwise, a random sentence from the text is selected. The model's task is to predict whether the next sentence is likely to follow sentence A, implementing binary classification.

3 DATASET

The dataset presented in [6], named Fakerom, encompasses around 14,000 articles, with a focus on COVID-19, and 1,200 of these are labeled for fake news detection.

Annotators classify articles into **six** classes: *real*, *plausible*, *propaganda*, *fabricated*, *satire*, and *fictional*.

Descriptive analysis reveals the distribution of articles across categories, with a notable skew towards real news. The length analysis indicates that over half of the articles are roughly less than 500 words, providing insights into resource requirements and potential correlations between article type and length.

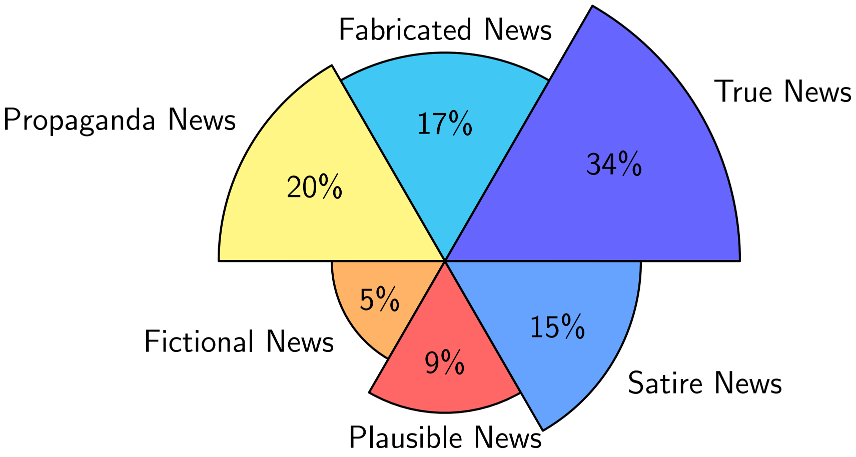

Figura 3.0 The distribution of classes in the FakeRom dataset

The dataset contributes valuable insights to the development of effective fake news detection mechanisms in the Romanian language.

4 SOLUTION

Next, I will present the implemented solution, focusing on two variants. Firstly, I will attempt automatic fake news detection using only the text part, achieved through fine-tuning BERT. Finally, I will implement automatic fake news detection using the multimodal dataset.

4.1 Fine-tuning BERT

The fine-tuning process of BERT [1] for automated fake news detection involves adjusting the weights obtained through pre-trained BERT, aiming to efficiently identify and classify offensive subtypes in the given input data context. Compared to the pre-training phase, where a massive amount of text is used to model language, with the goal of discovering linguistic patterns in context, the fine-tuning phase is relatively computationally inexpensive.

The first step in this process was to establish an annotated dataset, assigning six labels (*Satirical news*, *Real news*, *Propagandistic news*, *Plausible news*, *Fictional news* and *Fake news*) to each text paragraph. These labels enable a more nuanced understanding of romanian language, as opposed to a simple binary classification determining whether the news article is fake or not. Additionally, it's crucial to consider that BERT has a large architecture capable of capturing complex information.

Considering that the fine-tuning process involves creating a classification with six possible classes, the next step is to add a **fully connected layer** with six neurons above the last layer of the BERT architecture, one for each predictible class. Additionally, the **softmax** function is applied over this layer to provide the probability distribution of the classes.

During training, neuron weights were updated, attempting to minimize the results obtained through the *cross-entropy* error function, which measured the dissimilarities between the predicted label and the true one.

$$\mathcal{L}_{CE} = -\sum_{i=1}^{n} t_i \log (p_i) \tag{4}$$

where t_i is the true label, and p_i is the probability for class i.

4.2 MultiModal BiTransformers

The MultiModal Transformers [4] represent a foundational approach to multimodal BERT-like architectures, addressing the increasing prevalence of information presented in both textual and visual formats in the modern digital landscape. The proposed supervised multimodal bitransformer is designed to jointly fine-tune unimodally pretrained text and image encoders by strategically projecting image embeddings into the text token space.

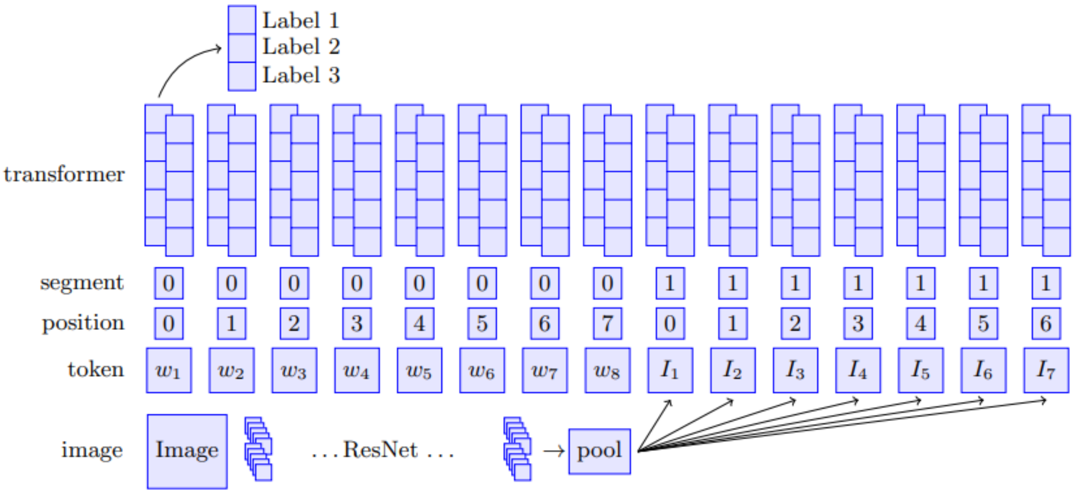

Figure 4.0 Illustration of the multimodal bitransformer architecture.

In the context of multimodal bitransformers, the text structure remains relatively unchanged, while image encodings are generated using a ResNet [2] architecture. The ultimate goal of this process is to obtain numerical representations (encodings) for both text and images, so that they can be concatenated and integrated into a transformers model. The concatenation and integration of these encodings allow the model to effectively understand and work with information from both modalities (text and image).

When addressing image encoding and classification within the domain of computer vision, a standard approach involves transferring the final fully connected layer of a pre-trained convolutional neural network. In this process, the output is commonly derived through a pooling operation over feature maps. Nevertheless, for multimodal bitransformers, the necessity for this pooling step diminishes, given that these models exhibit the capability to manage arbitrary numbers of dense inputs without depending on such pooling operations.

So, the model learn weights $W_n \in R^{P \times D}$ to project each of the N image embeddings to D-dimensional token input embedding space:

$$I_n = W_n \cdot f(\text{img}, n), \quad (1)$$

where $f(\cdot, n)$ represents the n-th output of the image encoder's final pooling operation.

5 CONCLUSION

In the ever-evolving landscape of artificial intelligence, this article has navigated the intricate journey from foundational perceptrons to transformative transformers, with focus on multimodal artificial intelligence.

The proposed supervised multimodal bitransformer stands out as a robust solution, particularly in the context of fake news detection within a Covid-19-themed dataset. Leveraging BERT and MultiModal BiTransformers, the article delves into the complexities of neural networks, showcasing a nuanced understanding of the Romanian language through the Fakerom dataset.

In essence, this article not only contributes to the discourse on fake news detection but also exemplifies the intersection of advanced models and datasets in shaping the future of intelligent information processing.

BIBLIOGRAFIE

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 10 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016.

[4] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text, 2020.

[5] Minh-Thang Luong, Hieu Pham, and Christopher Manning. Effective approaches to attention-based neural machine translation. 08 2015.

[6] Andrei Preda, Stefan Ruseti, Simina-Maria Terian, and Mihai Dascalu. Romanian fake news identification using language models. pages 73–79, 01 2022.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 06 2017.