FAKE NEWS DETECTION

Bar Rushkin (ID. 308574870), Alon Bar Nathan (ID. 308337997)
Department of Computer Science, Reichman University
Bachelor's Student in Computer Science

Submitted as final project report for the NLP course, RUNI, 2023

1 Introduction

The spread of fake news has become a major problem in recent years, with serious consequences for society (Shu et al., 2017; Gelfert, 2018). Natural language processing (NLP) methods can be used to classify fake news articles, but there is no one-size-fits-all approach. This article comprehensively explores diverse NLP techniques and models to address fake news classification. Initially , we explored the BERT (Bidirectional Encoder Representations from Transformers) model, chosen for its adeptness in capturing intricate contextual nuances essential for differentiating authenticity. Enhancements encompass BERT model extension through additional layers (dropout, ReLU activation, two fully connected layers and a SoftMax activation for generating class probabilities), hyperparameter fine-tuning, and varied pre-processing strategies evaluation. After our attempts with the Bert-Architecture yielded less favorable results, we pivoted to our new method, which subsequently led to remarkable outcomes. Our new approach involves leveraging the BertForSequenceClassification model from the transformers library, specifically the "bert-base-uncased" variant. This model excels in sequence classification tasks, making it a potent tool for discerning between real and fake news. By configuring it with "num labels = 2" and optimizing outputs, we tailored the architecture for binary classification and made the Preprocessing with BertTokenizerFast ensures seamless integration. This methodological shift played a pivotal role in achieving impressive results in our pursuit of effective fake news classification. This study holistically advances understanding of fake news classification. Moreover, diverse pre-processing strategies were explored, encompassing both tagged and untagged datasets. The efficacy of these pre-processing methodologies was evaluated to understand their impact on model performance.

1.1 Data sets

ISOT: The dataset comprises real and fake news articles, obtained from reputable sources like Reuters.com and flagged as unreliable sources by Politifact

and Wikipedia. Covering diverse topics with a political and global news focus, it offers "True.csv" with over 12,600 authentic articles and "Fake.csv" with an equal number of fabricated articles. Entries encompass title, text, category, and publication date, mainly spanning 2016 to 2017 and subjected to cleaning while retaining errors for fake news analysis.

News	**Size** (Number of articles)	***Subjects***	
Real-News	21417	**Type**	**Articles size**
		World-News	*10145*
		Politics-News	*11272*
Fake-News	23481	**Type**	**Articles size**
		Government-News	*1570*
		Middle-east	*778*
		US News	*783*
		left-news	*4459*
		politics	*6841*
		News	*9050*

Figure 1: The table provides a breakdown of categories and article counts

LIAR: The LIAR dataset serves as a valuable resource for detecting fake news, spanning a decade with 12.8K manually labeled short statements. Gathered from POLITIFACT.COM, renowned for its analytical reports and linked source documents, this dataset aids both fact-checking and research. Comprising human-labeled statements, each evaluated by POLITIFACT.COM editors, it provides insights into truthfulness levels. The labels follow PolitiFact's "truth-o-meter" definitions, ranging from true and mostly-true to false and pants-fire for outlandish claims. Label Distribution and "Truthiness" (taken from PolitiFact's "truth-o-meter" methodology page):

1. true – The statement is accurate and there's nothing significant missing.
2. mostly-true – The statement is accurate but needs clarification or additional information.
3. half-true – The statement is partially accurate but leaves out important details or takes things out of context.
4. barely-true – The statement contains an element of truth but ignores critical facts that would give a different impression.
5. false – The statement is not accurate.
6. pants-fire – The statement is not accurate and makes a ridiculous claim. a.k.a. "Liar, Liar, Pants on Fire!"

For further references see detailed datasets exploration and preparation

1.2 Related Works

In the realm of fake news classification, two pivotal datasets have garnered significant attention: the ISOT dataset (Ahmed et al., 2018; Ahmed et al., 2017) and the LIAR dataset (Wang, 2017). Classifying fake news presents multifaceted challenges. Defining fake news itself is intricate, with evolving definitions capturing various forms of misinformation. From outright falsehoods to subtle manipulations, the dynamic nature of fake news necessitates adaptable classification strategies (De Oliveira et al., 2021). Linguistic nuances, diffusion patterns, and intent variations compound the complexity. Researchers have explored diverse NLP techniques, including models like BERT, while innovative dataset creation and pre-processing methods further contribute to this pursuit. The amalgamation of ISOT and LIAR datasets provides a comprehensive foundation for effective fake news classification strategies, illuminating pathways to address this intricate concern. The ISOT and LIAR datasets are still under development, but they have already made a significant contribution to research on fake news detection. These datasets will continue to be valuable resources for researchers as they work to develop more effective methods for detecting fake news.

2 APPROACHES

Key challenges in fake news detection, explored through ISOT and LIAR datasets, encompass identifying falsehoods in brief statements, assessing social media's role in dissemination, leveraging NLP methods for detection, and devising human-in-the-loop systems for accurate identification. We observed numerous implementations that exclusively employed the ISOT dataset or the LIAR dataset for their studies. However, we embarked on a different path by combining both datasets to enhance our investigation. Moreover, we have investigated preprocessing of the different columns and combined different columns to figure out the most effective feature representation for our comprehensive analysis (as you can see in the dataset exploration section in our project). To achieve this, we adopted a conversion format inspired by the LIAR dataset's description. In this approach, labels such as "almost-true" and "true" were transformed into 1, while the remaining labels were set to 0, indicating falsehood. Our exploration extended beyond the datasets, encompassing a variety of combinations and models. We meticulously evaluated each approach to discern the most effective methods. Our exploration involved two main categories: tagged and untagged data. Under the tagged category, we considered multiple combinations, including subject, title, and text, each accompanied by relevant tags (as detailed in the code). Additionally, we explored a combination with solely the title and text. In the untagged category, we pursued similar methodologies. Throughout our experimentation, we tested varying dataset sizes: 5000, 15000, 30000, and the entire dataset (further details provided below). Furthermore, we examined different tokenized sentence lengths, ranging from 15 to 512 tokens.

Due to the size of the combined dataset, we split the dataset to 80-10-10 train-validation-test ratio. Initially, we opted for the BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) model due to its exceptional contextual comprehension, crucial for discerning nuanced distinctions between real and fake news. BERT's bidirectional attention mechanism enables a deeper understanding of linguistic subtleties, enhancing its efficacy in fake news detection. Our custom **BERT-based** architecture further optimizes BERT's strengths. By adding layers for dropout, ReLU activation, and fully connected networks, and fine-tuning hyperparameters, **BERT Arch** maximizes the model's capacity to grasp intricate patterns, tailored to the complexities of fake news classification. This approach aligns seamlessly with our objective of bolstering fake news detection accuracy. In our pursuit after the best results possible, we have embraced a new method that yielded significantly improved results. In this new approach, we leveraged the power of the **BertForSequenceClassification** model from the **Hugging Face** transformers library.

2.1 RESULTS

Our pursuit involved an intricate web of diverse model combinations, accounting for variable data sizes and maximum string lengths. Yet, this endeavor was not devoid of challenges, as we grappled with limitations posed by Google Colab's computational constraints. Navigating these constraints, we seized the opportunity to propel our investigation to advanced realms by securing a Pro+ subscription, which unlocked the gates to intricate runs—marked by heightened run times and augmented resource access. Notably, our quest brought us face-to-face with a critical insight: the absence of a marked performance advantage when our models were evaluated on labeled versus unlabeled datasets. However, the trajectory of improvement charted a remarkable ascent when we ventured beyond "off the shelf" models, devising our tailored arsenal and pitting it against a unified dataset. Resplendent in performance, our best model is our latest approach, using the **BertForSequenceClassification** model from Hugging Face with some minor adaptions.

Following the meticulous scrutiny of these models, our journey ventured into the realm of data set consolidation's impact on model performance. Unveiling our revelations, we've curated a repository housing all trained models, supplemented by accompanying run-up notebooks. The forthcoming article is poised to unveil our paramount findings, each encapsulated succinctly:

1. Benchmark - "Bert-Arch", with a maximum sequence length of 120 with no prior training: 53%

Test Evaluation:

	precision	recall	f1-score	support
0	0.53	1.00	0.69	1522
1	1.00	0.00	0.00	1362
accuracy			0.53	2884
macro avg	0.76	0.50	0.35	2884
weighted avg	0.75	0.53	0.37	2884

2. "Bert-Arch", with a maximum sequence length of 256 tokens and training over 2 epochs, achieved an accuracy of 82%

Test Evaluation:

	precision	recall	f1-score	support
0	0.83	0.85	0.84	6272
1	0.81	0.80	0.80	5266
accuracy			0.82	11538
macro avg	0.82	0.82	0.82	11538
weighted avg	0.82	0.82	0.82	11538

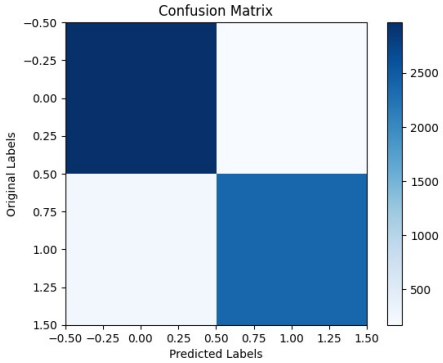
While seeking the best sequence length, we have found that all lengths between 20-256 produce accuracy that ranges between 78-82% with no significant impact on the overall performance

3. **BertForSequenceClassification** model, with a maximum sequence length of 120 tokens and training over 2 epochs, our approach achieved an impressive accuracy of 93%

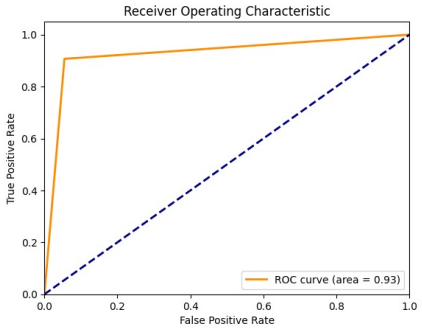
Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	3138
1	0.93	0.91	0.92	2631
accuracy			0.93	5769
macro avg	0.93	0.93	0.93	5769
weighted avg	0.93	0.93	0.93	5769

While seeking the best sequence length, we have found that all lengths between 20-120 produce accuracy that ranges between 91-93% with no significant impact on the overall performance



Confusion matrix



ROC curve

3 Discussion

Our exploration into fake news detection, driven by our roles as computer science undergraduates, illuminated challenges and insights. We honed various pre-processing techniques to optimize data readiness and built and worked with two BERT-based language models for precision.
While our investigations revealed that the maximum sequence length of a text has minimal impact on results, it's intriguing to speculate that the primary idea frequently emerges at the text's outset or even within its title. This speculation suggests that the crux of fake news might often be evident right from the start, potentially guiding us toward more efficient classification techniques.

Moving forward, our focus extends to refining the interface, fostering interactivity, harnessing real-time data, and promoting media literacy.

4 Code

For further references see Our project github repository or go to the project's google drive folder: Google Drive folder

5 BIBLIOGRAPHY

1. Gelfert, A. (2018). Fake News: a definition. Informal Logic, 38(1), 84–117. https://doi.org/10.22329/il.v38i1.5068

2. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media. SIGKDD Explorations, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

3. De Oliveira, N. R., Pisa, P. S., Lopez, M. A., Medeiros, D. S. V., & Mattos, D. M. F. (2021). Identifying fake news on social networks based on natural language Processing: Trends and challenges. Information, 12(1), 38. https://doi.org/10.3390/info12010038

4. Ahmed H, Traore I, Saad S. "Detecting opinion spams and fake news using text classification", Journal of Security and Privacy, Volume 1, Issue 1, Wiley, January/February 2018.

5. Ahmed H, Traore I, Saad S. (2017) "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham (pp. 127- 138).

6. Wang, W. (2017). "Liar, liar Pants on Fire": a new benchmark dataset for fake news detection. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1705.00648