From Hype to Reality: Revealing the Accuracy and Robustness of Transformer-Based Models for Fake News Detection

Dorsaf Sallami[1,*], Ahmed Gueddiche[1] and Esma Aïmeur[1]

[1] *Department of Computer Science and Operations Research (DIRO), University of Montreal, Canada*

Abstract

The prevalence of fake news in today's society is a serious concern, as it can compromise the reliability of information and have detrimental effects on individuals and communities. In this article, we conduct a comprehensive evaluation of six distinct Transformers to investigate their effectiveness in detecting Fake News. First, we examine the performance of these models on four diverse datasets, each representing a distinct language. Second, we investigate the robustness of these models against adversarial attacks to assess their vulnerability and measure the impact of such attacks on their performance. Our findings indicate that while transformers are commonly employed, their performance exhibits significant variability across datasets and languages. Moreover, our analysis reveals their vulnerability to attacks, as demonstrated by a notable drop in accuracy when confronted with deliberate manipulations.

Keywords

Fake News Detection, Transformers, Adversarial attacks, Comparative study

1. Introduction

Misinformation is a significant concern globally, with particular emphasis on the harm it can cause during crises [1]. Whether it is a natural disaster [2], a public health emergency [3], or a political upheaval [4], a crisis creates a great deal of uncertainty and fear among the public [5]. In such circumstances, people tend to seek information and guidance, which may prompt them to rely on sources they typically wouldn't consider trustworthy [6]. Unfortunately, this can lead to the spread of fake news, which can exacerbate the crisis, particularly on social media [7].

In the midst of the ongoing problem posed by fake news, academics are dedicating significant effort to develop efficient methods for identifying and countering misleading information [8, 9, 10]. One approach that has gained widespread popularity is the utilization of transformer models. In fact, Transformer is a highly influential deep learning model that has gained widespread adoption across multiple domains, including natural language processing, computer vision, and speech processing [11].

Figure 1 displays the number of published papers on the subject per year. They were retrieved using the Google Scholar internal search engine[1]. Notice the strongly increasing trend. Nonetheless, a crucial question arises: Can transformers genuinely excel in detecting fake news?

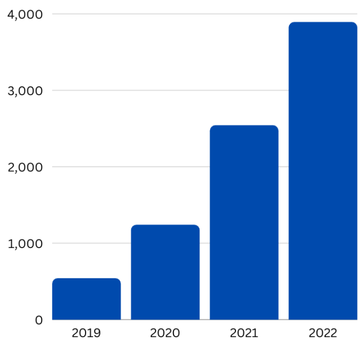


Figure 1: Number of papers retrieved from Google Scholar.

In this article, we conduct a comparative analysis of six recent Transformer-based architectures, namely BERT, RoBERTa, DistilBERT, XLNet, GPT-J, and GPT-2. Our focus is on evaluating their performances in detecting fake news on various datasets. The objective is to assess the performance of transformers across different languages, expanding the scope beyond the conventional English-focused evaluations. By incorporating datasets

AiOfAi'23: 3rd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, Macao, China

*Corresponding author.
dorsaf.sallami@umontreal.ca (D. Sallami); ahmed.gueddiche@umontreal.ca (A. Gueddiche); aimeur@iro.umontreal.ca (E. Aïmeur)
0000-0001-5077-3413 (D. Sallami); 0000-0001-7414-5454 (E. Aïmeur)

[1] We conducted a search on Google Scholar using the query "fake news detection transformers". Last updated: 7 May 2023

in multiple languages, we are able to evaluate the effectiveness and adaptability of transformers in detecting fake news across a wide range of linguistic contexts. Additionally, we investigate the resilience of these models against adversarial attacks. Therefore, the main contributions are:

- Assessing various cutting-edge transformer models to evaluate the impact of various designs on the same dataset.
- Assessing their performance and effectiveness in detecting fake news in different languages.
- Explore the robustness of these transformer models by implementing adversarial attacks to evaluate their resistance to deliberate manipulation and deception techniques commonly used in spreading fake news.
- Developing an interactive interface that allows us to visualize and explore our experimental results.

The remaining sections of this article are structured as follows: Section 2 provides a concise overview of important research on detecting fake news and adversarial attacks. Section 3 outlines the methodology employed for the comparative evaluation. Section 4 provides an examination of the transformer models employed in this research. Our primary objective is to harness the inherent capabilities of transformers to enhance the accuracy and efficiency of fake news detection. In Section 5, we delve into an exploration of cutting-edge techniques for adversarial attacks. Through our extensive testing, we aim to uncover vulnerabilities and assess the robustness of the models under various adversarial scenarios. Section 6 details the four datasets, experimental setup, and preliminary analysis, enabling the replication of our tests. The results are presented and discussed in Section 7. Finally, Section 8 concludes the article.

2. Literature Review

Fake news detection is an ever-expanding research topic that is gaining a lot of attention since there are still a lot of challenges that need to be investigated [12]. There are various studies have been carried out on the detection of fake news [13], which can be classified into four categories: *Knowledge-based* approaches involve fact-checking news content by comparing claims to known facts. *Style-based* approaches analyze the content for patterns. *Propagation-based* approaches examine the spread of fake news, and *credibility-based* approaches assess the credibility of the news source and users who share it.

This section delves into two distinct aspects of fake news research. Firstly, we examine the utilization of transformers for detecting fake news. Secondly, we review recent studies that concentrate on adversarial attacks targeted at fake news detection systems.

2.1. Transformers for Fake News Detection

The proliferation of fake news has become a major concern in the age of social media [14]. To address this issue, researchers have been investigating a range of techniques, among them the use of sophisticated machine learning algorithms such as transformers. These algorithms have gained significant traction in the field of natural language processing [15].

Some researchers [16, 17, 18] have used *pre-trained transformers* such as BERT [19], leveraging their pre-training capabilities by fine-tuning them on fake news detection datasets. Other scholars have opted for *adapted transformers*, modifying existing transformer models to improve their performance on fake news detection tasks [20, 21]. They experiment with different architectures and hyperparameters and evaluate their performance on various datasets. Finally, *domain-specific transformers* are fine-tuned on datasets that are specific to domains like politics, finance, or health [22, 23].

2.2. Adversarial Attacks

Adversarial Machine Learning is a developing field in applied machine learning that aims to comprehend how machine learning classifiers can be targeted and compromised by malicious users.

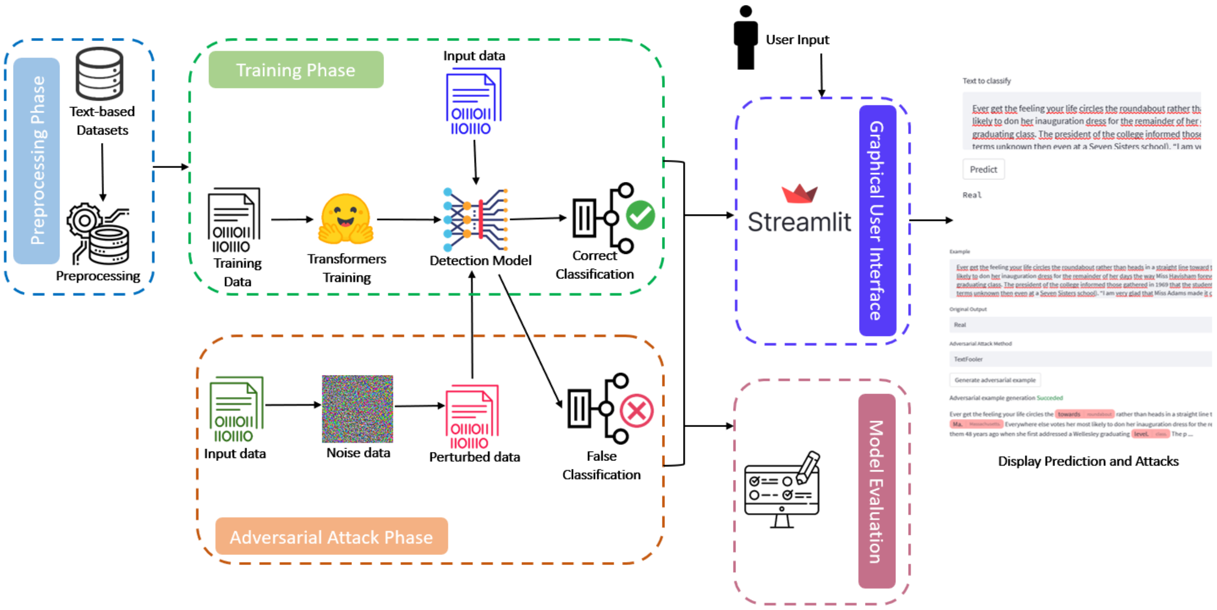
Adversarial training was initially developed to increase model robustness by incorporating small perturbations into the training data, and has since been found to enhance model generalization [24]. In computer vision, different attacks [25] have achieved remarkable results. Meanwhile, in recent years, researchers have proposed various adversarial training techniques for natural language processing tasks [26].

To improve the accuracy of fake news detection models, many studies have incorporated adversarial attacks, particularly in the context of multimodal models that include images [27, 28, 29]. In contrast, there has been comparatively less focus on integrating adversarial attacks into classifiers that rely solely on text inputs for fake news detection, [30, 31].

3. Methodology

In this section, we present the system architecture illustrated in Figure 2. It consists of five major components: 1) data preprocessing component; 2) transformer-based fake news detection component; 3) adversarial attacks component; 4) model evaluation component and 5) Graphical User Interface (GUI) component.

The project is divided into two phases: offline training and online prediction via GUI. During the offline training



Text to classify

Ever get the feeling your life circles the roundabout rather th
likely to don her inauguration dress for the remainder of her
graduating class. The president of the college informed those
terms unknown then even at a Seven Sisters school). "I am ve

Predict

Real

Example

Ever get the feeling your life circles the roundabout rather than heads in a straight line toward t
likely to don her inauguration dress for the remainder of her days the way Miss Havisham forev
graduating class. The president of the college informed those gathered in 1969 that the student
terms unknown then even at a Seven Sisters school). "I am very glad that Miss Adams made it c

Original Output

Real

Adversarial Attack Method

TextFooler

Generate adversarial example

Adversarial example generation Succeded

Ever get the feeling your life circles the towards roundabout rather than heads in a straight line t
Ma. Massachusetts Everywhere else votes her most likely to don her inauguration dress for the re
them 48 years ago when she first addressed a Wellesley graduating level. class. The p ...

Figure 2: System Architecture.

phase, a diverse range of datasets comprising various languages are collected and preprocessed. These datasets are then utilized to train several transformer-based models. In this paper, we cover the auto-regressive models such as GPT-J, GPT-2, and XLNET, as well as the auto-encoder architecture such as BERT and post-BERT models like DistilBERT and RoBERTa. The second part involves subjecting the trained models to adversarial attacks to evaluate their robustness. Once the offline training phase is completed, the trained model is integrated into a graphical user interface (GUI). The GUI serves as an interactive platform where users can input text or news and obtain predictions, such as determining if the content is fake or real, as well as the outcomes of adversarial attacks.

4. Models Architectures

In this section, we provide a concise overview of the architectures employed in our experiments.

4.1. Transformer architecture

The Transformer is based on an encoder-decoder structure [15], where it takes a sequence $X = (x_1, \ldots, x_N)$ and produces a latent representation $Z = (z_1, \ldots, z_N)$. Due to the autoregressive property of this model, the output sequence $Y_M = (y_1, \ldots, y_M)$ is produced one element at a time, i.e., the word y_M uses the latent representation Z and the previously created sequence $Y_{M-1} = (y_1, \ldots, y_{M-1})$ to be generated.

The Encoder and the Decoder are using the same Multi-Head Attention layer. A single Attention layer maps a query Q and keys K to a weighted sum of the values V. For technical reasons, there is a scaling factor of $\sqrt{\frac{1}{d_k}}$.

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

4.2. GPT-J

EleutherAI[2] has developed an open-source artificial intelligence language model called GPT-J, which exhibits comparable performance to OpenAI's GPT-3 across various zero-shot downstream tasks and even surpasses it in code generation tasks[3]. GPT-J-6B, the latest version, is based on The Pile[4], an open-source language modeling dataset comprising 22 smaller datasets, with a total size of 825 gibibytes. While GPT-J shares similar capabilities with ChatGPT, it does not function as a chatbot, but rather as a text predictor[5].

[2] https://gpt3demo.com/apps/gpt-j-6b

[3] https://www.forefront.ai/blog-posts/gpt-j-6b-an-introduction-to-the-largest-open-sourced-gpt-model

[4] https://pile.eleuther.ai/

[5] https://towardsdatascience.com/how-you-can-use-gpt-j-9c4299dd8526

4.3. GPT-2

An open-source artificial intelligence developed by OpenAI in February 2019 [32]. It has the capability to translate text, answer questions, summarize passages, and generate text output at a level that can be indistinguishable from human-generated text. However, it may become repetitive or nonsensical when generating longer passages. GPT-2 is a general-purpose learner that was not trained specifically for any of these tasks; rather, its ability to perform them stems from its general ability to accurately synthesize the next item in a sequence. GPT-2 is a "direct scale-up" of OpenAI's 2018 GPT model, with a ten-fold increase in both its parameter count and the size of its training dataset.

4.4. XLNet

XLNet is a pretraining method proposed in [33] that utilizes generalized autoregressive techniques to enable bidirectional context learning. It achieves this by maximizing the expected likelihood over all possible permutations of the factorization order. XLNet has demonstrated superior performance over BERT in various tasks, including question answering, natural language inference, sentiment analysis, and document ranking, often by a significant margin. However, there is currently no well-established XLNet model pre-trained on a Spanish corpus. Therefore, in our study, we implemented the same pre-trained XLNet model for both datasets and evaluated its performance using a zero-shot cross-lingual transfer technique [34].

4.5. BERT

BERT [19] is a language representation model that employs bidirectional pre-training by conditioning on both left and right context of unlabeled text. The model is pre-trained using two objectives: (1) Masked Language Modeling (MLM), where the model randomly masks 15% of the words in a sentence and predicts the masked words; and (2) Next Sentence Prediction (NSP), where the model concatenates two masked sentences as inputs and predicts whether they follow each other or not. Fine-tuning of the model for specific tasks can be achieved with the addition of just one extra output layer.

4.6. DistilBERT

DistilBERT [35] is a technique for pre-training a universal language representation model that results in a smaller model than BERT. Through a distillation process, DistilBERT reduces the size of a BERT model by 40%, while still retaining 97% of its language understanding abilities and exhibiting a 60% faster processing speed.

4.7. RoBERTa

In [36], the authors present a replication study of BERT pre-training and introduce modifications to enhance its performance. These modifications involve training the model for longer durations with larger batches, eliminating the next sentence prediction objective, training on longer sequences, and dynamically adjusting the masking pattern used in the training data.

5. Adversarial Attacks

To assess the robustness of different transformers, adversarial examples are generated using the following methods:

5.1. TextFooler

TextFooler [37] was introduced as a technique to investigate the robustness of BERT, a pre-trained language model, against natural language attacks on text classification and entailment. The steps for creating new adversarial examples can be described as follow:

1. *Word Importance Ranking*: This step involves ranking the importance of words in a given sentence and choosing the words that significantly influence the final prediction results. The importance score I_{w_i} is therefore calculated as the prediction change before and after deleting the word w_i.
2. *Synonym Extraction*: Gather a candidate set for all possible replacements of the selected word w_i . Candidates is initiated with the N closest synonyms according to the cosine similarity between w_i and every other word in the vocabulary.
3. *POS Checking*: In the set of candidates of the word w_i, we only keep the ones with the same part-of-speech (POS) as w_i. This step is to assure that the grammar of the text is mostly maintained.
4. *Semantic Similarity Checking*: For each remaining word in candidates, it is substituted for w_i in the sentence X, and obtain the adversarial example. The target model F is then used to compute the corresponding prediction scores $F(X_{adv})$. The sentence semantic similarity between the source X and adversarial counterpart X_{adv} is also computed to filter out the best semantically similar sentences.

5.2. Bert-Attack

Bert-Attack (BAE) [38] is a black-box adversarial attack method that generates minimal perturbations to a given sentence to mislead a target model's prediction using the language model Bert. The method is efficient and effective in generating adversarial examples that preserve the original semantics of the sentence while maximizing the risk of wrong predictions.

The main two steps to generate adversarial examples using this technique are as follows:

1. *Finding Vulnerable Words* In this step, the authors aim to find the words in a given sentence that have a high significance influence on the final output logit of a target model. To do this, they define an importance score, Iwi, for each word in the sentence. The importance score is defined as the difference between the logit output by the target model for the original sentence and the logit output for the sentence with the current word replaced with [MASK]. The authors then rank all the words in the sentence according to their importance score in descending order to create a word list. The authors only take a certain percentage of the most important words in the list since they tend to keep perturbations to a minimum.
2. *Word Replacement via BERT* In this step, the authors aim to find perturbations that can mislead the target model by iteratively replacing words in the previously generated word list with semantically consistent perturbations. Previous approaches have used human-crafted rules and strategies to ensure the generated examples are semantically consistent and grammatically correct, but these strategies are insufficient in fluency control and semantic consistency. To overcome these limitations, the authors leverage BERT for word replacement. They use the masked language model to generate perturbations that are relatively fluent, grammar-correct, and preserve most semantic information. Unlike previous methods, the contextualized perturbation generator generates minimal perturbations with only one forward pass, making the process extremely efficient.

6. Experimental Implementation

6.1. Datasets

In our experiments, we utilized four publicly available datasets: Kaggle [39], CHEKED [40], AFND [41], and The Spanish Fake News Corpus (FNCS) [42]. While the majority of datasets used for fake news detection are commonly available in English [43], we were determined to include datasets in Arabic, Chinese, and Spanish languages. The objective was to assess the performance of transformers across different languages, expanding the scope beyond the conventional English-focused evaluations. Table 1 provides a summary of these datasets.

6.2. Experimental Setup

The experimental setup involved utilizing Google Colab Pro and PyTorch for conducting the experiments. To pre-trained models, we used the Hugging Face library[6]. The specific hyperparameters used in our model can be found in Table 2.

For each dataset, we carefully selected the most suitable transformer variation to achieve optimal performance. The choice of transformer variation was based on language-specific considerations of the dataset. For example, with the Kaggle dataset, we utilized the best-base-case transformer variation, which is specifically designed for English language processing. Similarly, for the CHECKED dataset, we employed the bert-base-chinese transformer variation, which is tailored to handle the unique features of the Chinese language. We followed a similar approach for the remaining transformers and datasets.

6.3. Evaluation Metrics

The evaluation of different transformers for fake news detection involves assessing their performance and robustness. Performance evaluation utilizes four commonly used metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's effectiveness in identifying fake news.

For robustness evaluation, various automatic evaluation metrics are employed to assess the quality of generated samples. Key indicators include attack accuracy, which measures the success rate of attacks on the model, and perturbed percentage, which quantifies the extent to which generated samples have been modified. In a black-box setting, the number of queries per sample is limited to 1000 due to computational resource constraints, serving as a significant metric for evaluating robustness.

6.4. GUI Implementation:

The streamlit library[7] is employed to create a user-friendly interface for visualizing and analyzing experimental results during the development process. It integrates with transformers-based machine learning models. The application's front end is responsible for receiving user data, while the back end utilizes the model to process the data provided by the user. Finally, the output is displayed on the screen.

[6] https://huggingface.co/docs/transformers/index
[7] https://streamlit.io/

Table 1
Summary of Datasets Used.

Dataset	Language	Train set	Test set	#Fake news	#Real news
Kaggle	English	35918	8980	23481	21417
CHECKED	Chinese	1683	421	344	1760
AFND	Arabic	64000	16000	40000	40000
FNCS	Spanish	676	572	624	624

Table 2
Hyperparameters used for Training.

Hyperparameters	Experimental value
Number of epochs	5
Batch size	8
Warmup steps	500
Weight decay	0.01
Logging steps	400

7. Results and Discussion

7.1. Assessing Performance

Table 3 offers a comprehensive breakdown of the performance metrics for each model across various datasets. For the Kaggle dataset, all models achieved excellent scores (99 or 100) across all evaluation metrics, indicating exceptional performance on this dataset. On the CHECKED dataset, BERT model had the highest accuracy (99) balanced precision (98), and F1 score (99). In the case of the AFND dataset, the DistilBERT model achieved the highest accuracy (77) and F1 score (77), while BERT had the highest precision (85). Lastly, for the FNCS dataset, the DistilBERT model outperformed others with the highest accuracy (70), precision (82), and F1 score (64).

Examining the performance of transformer models on a specific dataset provides valuable insights into their strengths and weaknesses with respect to that particular data. Figure 3 specifically concentrates on the CHECKED dataset, allowing for a detailed analysis of how different models perform on this specific data.

Among the transformer models evaluated on the CHECKED dataset, BERT performs exceptionally well across all metrics, showing high accuracy, precision, recall, and F1 scores. RoBERTa demonstrates competitive performance with high accuracy and F1 score, but falls slightly behind BERT in terms of precision and recall. DistilBERT achieves strong performance overall, although its precision and recall are slightly lower compared to BERT. XLNet emerges as the top-performing model on the CHECKED dataset, surpassing others in accuracy, precision, recall, and F1 score, indicating its balanced performance in correctly classifying positive samples and capturing true positive instances.

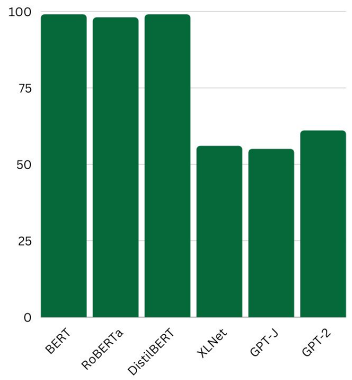


Figure 3: Accuracy of Transformers on the CHECKED Dataset.

7.2. Evaluating Robustness

Given the findings from the previous section, which highlighted BERT and DistilBERT as the top-performing models across different datasets, this section aims to delve into their resilience against adversarial attacks.

Based on the results, as shown in Table 4, it can be inferred that both TextFooler and Bert-Attack are capable of performing successful attacks on the model, as shown by the significant number of successful attacks and the noticeable decrease in accuracy under attack. Nonetheless, Bert-Attack appears to be more proficient in altering the predicted labels of the inputs, having a higher attack success rate and more successful attacks compared to TextFooler.

Table 5 displays an adversarial example that was generated. It shows examples of how different adversarial attacks can modify a piece of text, leading to different levels of modification and different results in terms of whether the modified text is classified as real or fake. From this, we can conclude that the effectiveness of an adversarial attack depends on the level of modification made to the text, as well as the specific machine learning model or classifier being targeted. Additionally, the table highlights the importance of developing robust machine-learning models that are resistant to adversarial attacks,

Table 3
Evaluation Results.

Dataset	Model	Accuracy	Precision	Recall	F1 score
Kaggle	BERT	99	99	99	99
	RoBERTa	99	99	99	99
	DistilBERT	99	99	99	99
	XLNet	100	100	100	100
	GPT-J	100	100	100	100
	GPT-2	100	100	100	100
CHECKED	BERT	99	98	100	99
	RoBERTa	98	98	92	95
	DistilBERT	99	96	98	97
	XLNet	56	55	63	59
	GPT-J	55	56	49	52
	GPT-2	61	62	57	59
AFND	BERT	68	85	44	58
	RoBERTa	69	46	51	66
	DistilBERT	77	79	75	77
	XLNet	57	56	64	60
	GPT-J	55	56	49	52
	GPT-2	61	62	57	59
FNCS	BERT	62	82	31	45
	RoBERTa	66	76	46	58
	DistilBERT	70	82	53	64
	XLNet	57	56	64	60
	GPT-J	50	50	52	51
	GPT-2	55	54	73	62

Table 4
Results of the Kaggle dataset attack utilizing DistilBERT.

Metrics	TextFooler	BAE
Number of successful attacks	1730	2036
Number of failed attacks	896	590
Original accuracy	99%	99%
Accuracy under attack	31.73%	20.66%
Attack success rate	65.88%	77.53%
Average perturbated word	25.34%	27.53%
Average num. words per input	11.72%	11.72%
Avgerage num. queries	76.97	84.75

especially in applications where the integrity of the data is critical.

7.3. Interactive Interface

The interface provides a convenient way to explore the performance of different transformer models for fake news detection and the effects the adversarial attacks.

Figure 4 showcases the interface for fake news detection. Users can input text and generate predictions by simply clicking the "Predict" button. On the other hand, Figure 5 presents the interface designed for exploring adversarial attacks. It allows users to select an attack method and displays the modified words in the original text as the output.

7.4. Discussion

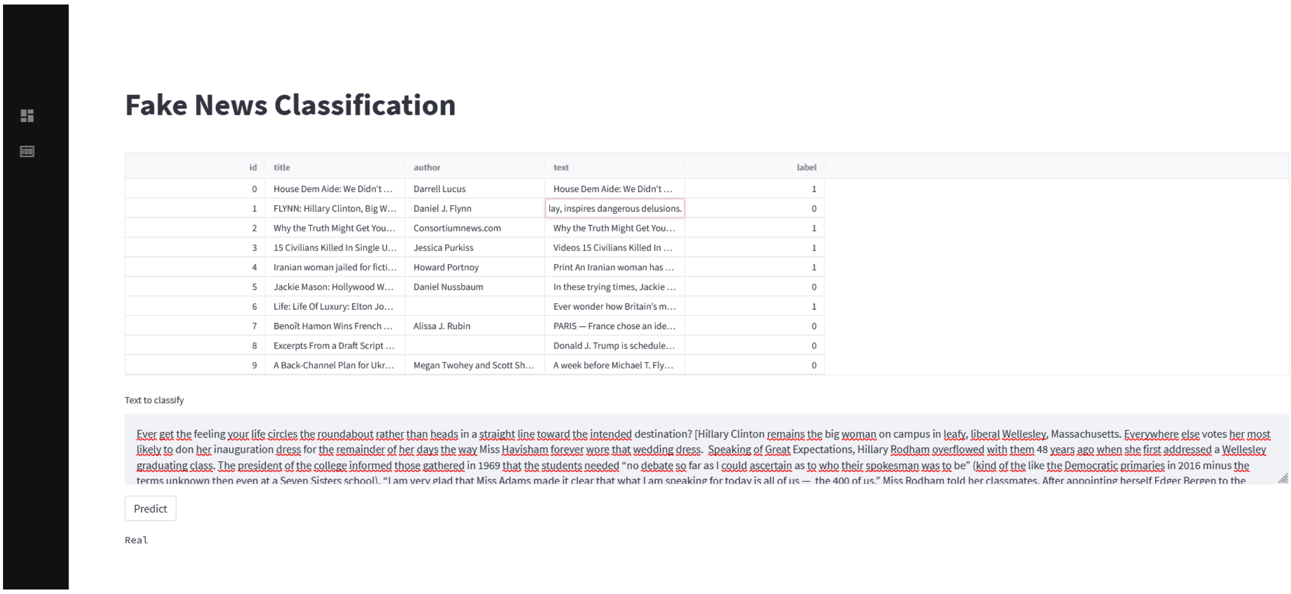
Dataset Variability: The performance variation across different datasets suggests that dataset characteristics have a significant impact on model performance. The language factor significantly influences how well a model generalizes and performs on a given dataset. It is worth noting that the rapid advancements in transformer architectures have primarily focused on testing and reporting performance on high-resource languages such as English [44, 45]. This could potentially explain why the Kaggle dataset yielded the best performance. Another factor worth considering is the dataset size. As evident from the decreased performance of the models on the FNCS dataset, which comprised only 1248 items.

Trade-off between Precision and Recall: The F1 score provides a balance between precision and recall, and the variations observed in F1 scores on the CHECKED dataset indicate potential trade-offs between these metrics. XLNet achieves high precision and recall, while BERT and DistilBERT strike a slightly different balance, resulting in slightly higher F1 scores.

The Impact of Adversarial Attacks: Despite DistilBERT being one of the best-performing models on the

Table 5
Adversarial Examples of different attacks

Original	Professor and Attorney Rahul Manchanda worked for one of the largest law firms in Manhattan where he focused on asbestos litigation. At the United Nations Commission on International Trade Law ("UNCITRAL") in Vienna, Austria, Mr. Manchanda was exposed... He later worked for ... multi-national law firms in Paris France, Coudert Frères, where he ...	Fake
TextFooler	Professor and Attorney Rahul Manchanda worked for one of the largest law ... At the United Nations Commission on International Trade Law ("UNCITRAL") in Vienna, Austria, Mr. Manchanda was exposed... He later worked ... multinational law firms in Paris France, Coudert Frères, where he ...	Real
BAE	Schoolmaster and Attorney Rahul Manchanda worked for one of the grande law firms in Harlem where he focused on asbestos litigation. During the United Nations Commission on International Trade Law ("UNCITRAL") in Vienna, Austria, Mr. Manchanda was displayed ... He again worked for one of the largest multi-national legislature company in ..	Real



Fake News Classification

id	title	author	text	label
0	House Dem Aide: We Didn't …	Darrell Lucus	House Dem Aide: We Didn't …	1
1	FLYNN: Hillary Clinton, Big W…	Daniel J. Flynn	lay, inspires dangerous delusions.	0
2	Why the Truth Might Get You…	Consortiumnews.com	Why the Truth Might Get You…	1
3	15 Civilians Killed In Single U…	Jessica Purkiss	Videos 15 Civilians Killed In …	1
4	Iranian woman jailed for ficti…	Howard Portnoy	Print An Iranian woman has …	1
5	Jackie Mason: Hollywood W…	Daniel Nussbaum	In these trying times, Jackie …	0
6	Life: Life Of Luxury: Elton Jo…		Ever wonder how Britain's m…	1
7	Benoît Hamon Wins French …	Alissa J. Rubin	PARIS — France chose an ide…	0
8	Excerpts From a Draft Script …		Donald J. Trump is schedule…	0
9	A Back-Channel Plan for Ukr…	Megan Twohey and Scott Sh…	A week before Michael T. Fly…	0

Text to classify

Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the intended destination? [Hillary Clinton remains the big woman on campus in leafy, liberal Wellesley, Massachusetts. Everywhere else votes her most likely to don her inauguration dress for the remainder of her days the way Miss Havisham forever wore that wedding dress. Speaking of Great Expectations, Hillary Rodham overflowed with them 48 years ago when she first addressed a Wellesley graduating class. The president of the college informed those gathered in 1969 that the students needed "no debate so far as I could ascertain as to who their spokesman was to be" (kind of the like the Democratic primaries in 2016 minus the terms unknown then even at a Seven Sisters school). "I am very glad that Miss Adams made it clear that what I am speaking for today is all of us — the 400 of us." Miss Rodham told her classmates. After appointing herself Edger Bergen to the

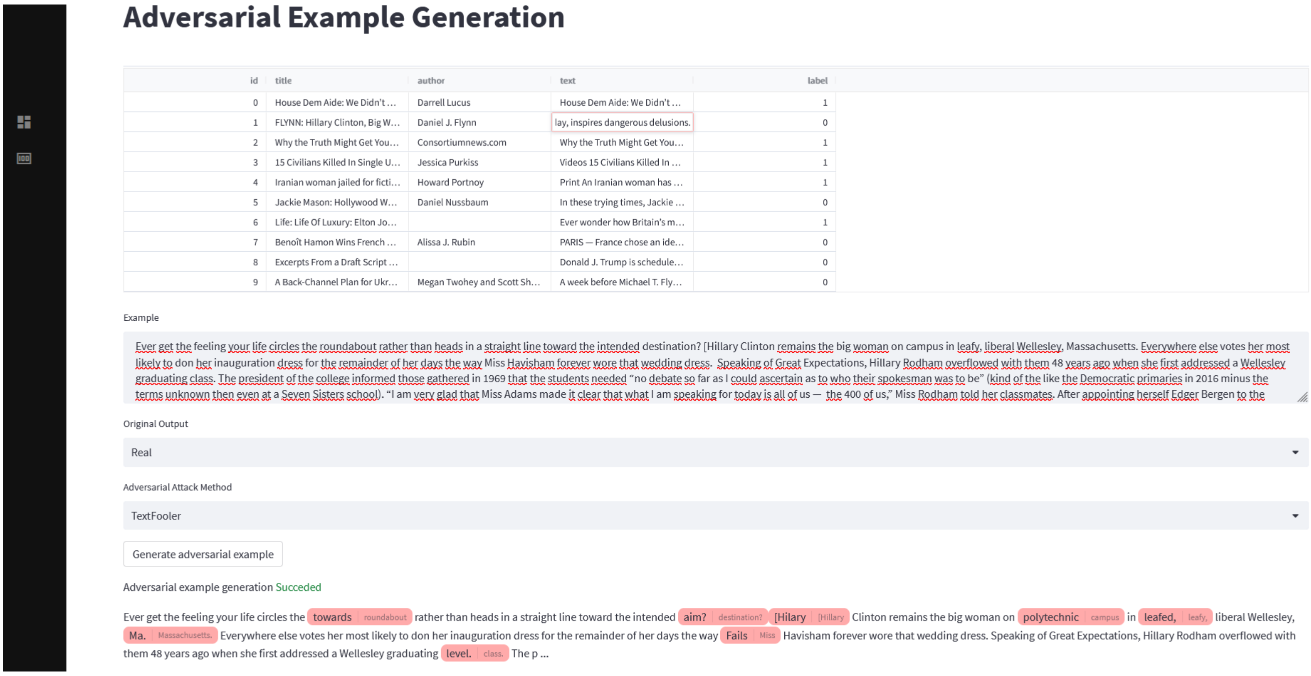
```
Real
```

Figure 4: Fake news detection Interface.

four datasets, the results reveal its vulnerability to adversarial attacks. The high number of successful attacks demonstrates that DistilBERT can be easily manipulated. Additionally, the accuracy of the model significantly decreases when subjected to both the TextFooler and BAE attacks, illustrating the adverse impact of these attacks on its performance. These findings emphasize the need to address the susceptibility of DistilBERT to adversarial examples to ensure its reliability and robustness in real-world scenarios.

8. Conclusion and Future Research

In summary, our findings demonstrate that the accuracy of Transformers is vulnerable to manipulation by adversarial attacks. In addition, we find that the performance and effectiveness of Transformers are affected by the language of the training datasets. This study emphasizes the need for effective and dependable detection methods to combat the problem of fake news. Our approach of utilizing adversarial attacks to test the resilience of Transformer-based models offers a promising solution to enhance the accuracy and robustness of fake news de-



Adversarial Example Generation

id	title	author	text	label
0	House Dem Aide: We Didn't …	Darrell Lucus	House Dem Aide: We Didn't …	1
1	FLYNN: Hillary Clinton, Big W…	Daniel J. Flynn	lay, inspires dangerous delusions.	0
2	Why the Truth Might Get You…	Consortiumnews.com	Why the Truth Might Get You…	1
3	15 Civilians Killed In Single U…	Jessica Purkiss	Videos 15 Civilians Killed In …	1
4	Iranian woman jailed for ficti…	Howard Portnoy	Print An Iranian woman has …	1
5	Jackie Mason: Hollywood W…	Daniel Nussbaum	In these trying times, Jackie …	0
6	Life: Life Of Luxury: Elton Jo…		Ever wonder how Britain's m…	1
7	Benoît Hamon Wins French …	Alissa J. Rubin	PARIS — France chose an ide…	0
8	Excerpts From a Draft Script …		Donald J. Trump is schedule…	0
9	A Back-Channel Plan for Ukr…	Megan Twohey and Scott Sh…	A week before Michael T. Fly…	0

Example

Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the intended destination? [Hillary Clinton remains the big woman on campus in leafy, liberal Wellesley, Massachusetts. Everywhere else votes her most likely to don her inauguration dress for the remainder of her days the way Miss Havisham forever wore that wedding dress. Speaking of Great Expectations, Hillary Rodham overflowed with them 48 years ago when she first addressed a Wellesley graduating class. The president of the college informed those gathered in 1969 that the students needed "no debate so far as I could ascertain as to who their spokesman was to be" (kind of the like the Democratic primaries in 2016 minus the terms unknown then even at a Seven Sisters school). "I am very glad that Miss Adams made it clear that what I am speaking for today is all of us — the 400 of us," Miss Rodham told her classmates. After appointing herself Edger Bergen to the

Original Output

Real

Adversarial Attack Method

TextFooler

Generate adversarial example

Adversarial example generation Succeded

Ever get the feeling your life circles the towards roundabout rather than heads in a straight line toward the intended aim? destination? [Hilary [Hillary Clinton remains the big woman on polytechnic campus in leafed, leafy, liberal Wellesley, Ma. Massachusetts. Everywhere else votes her most likely to don her inauguration dress for the remainder of her days the way Fails Miss Havisham forever wore that wedding dress. Speaking of Great Expectations, Hillary Rodham overflowed with them 48 years ago when she first addressed a Wellesley graduating level. class. The p ...

Figure 5: Adverserial attacks Interface.

tection systems. Our comparative evaluation of different models and exploration of attack techniques showcase the potential for improving the accuracy and robustness of existing models. Furthermore, the interface developed in this study provides an accessible means of visualizing the results of these experiments, making them more comprehensible to a broader audience. In our future research, we intend to investigate the models' resilience to adversarial attacks and identify potential vulnerabilities. To protect against such attacks, we will implement adversarial training and analyze the trade-off between robustness and performance. These efforts will contribute to the development of more reliable and effective methods for fake news detection.

References

[1] S. Waisbord, Truth is what happens to news: On journalism, fake news, and post-truth, Journalism studies 19 (2018) 1866–1878.

[2] S. S. Azim, A. Roy, A. Aich, D. Dey, Fake news in the time of environmental disaster: Preparing framework for covid-19 (2020).

[3] V. Balakrishnan, N. W. Zhen, S. M. Chong, G. J. Han, T. J. Lee, Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review, International Journal of Disaster Risk Reduction (2022) 103144.

[4] S. C. Rhodes, Filter bubbles, echo chambers, and fake news: how social media conditions individuals to be less critical of political misinformation, Political Communication 39 (2022) 1–22.

[5] G. R. Rodriguez, S. Gautam, A. Tapia, Understanding twitters behavior during the pandemic: Fake news and fear, arXiv preprint arXiv:2202.05134 (2022).

[6] R. Das, W. Ahmed, Rethinking fake news: Disinformation and ideology during the time of covid-19 global pandemic, IIM Kozhikode Society & Management Review 11 (2022) 146–159.

[7] C. Melchior, M. Oliveira, Health-related fake news

on social media platforms: A systematic literature review, new media & society 24 (2022) 1500–1522.

[8] S. Amri, D. Sallami, E. Aïmeur, Exmulf: An explainable multimodal content-based fake news detection system, in: International Symposium on Foundations and Practice of Security, Springer, 2022, pp. 177–187.

[9] D. Sallami, R. Ben Salem, E. Aïmeur, Trust-based recommender system for fake news mitigation, in: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, 2023, pp. 104–109.

[10] X. Zhou, K. Shu, V. V. Phoha, H. Liu, R. Zafarani, "this is fake! shared it by mistake": Assessing the intent of fake news spreaders, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3685–3694.

[11] T. Lin, Y. Wang, X. Liu, X. Qiu, A survey of transformers, AI Open (2022).

[12] W. Shahid, B. Jamshidi, S. Hakak, H. Isah, W. Z. Khan, M. K. Khan, K.-K. R. Choo, Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities, IEEE Transactions on Computational Social Systems (2022).

[13] X. Zhou, R. Zafarani, Fake news: A survey of research, detection methods, and opportunities, arXiv preprint arXiv:1812.00315 2 (2018).

[14] D.-R. Obadă, D.-C. Dabija, "in flow"! why do users share fake news about environmentally friendly brands on social media?, International Journal of Environmental Research and Public Health 19 (2022) 4861.

[15] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled, Overview of the transformer-based models for nlp tasks, in: 2020 15th Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2020, pp. 179–183.

[16] A. Hande, K. Puranik, R. Priyadharshini, S. Thavareesan, B. R. Chakravarthi, Evaluating pretrained transformer-based models for covid-19 fake news detection, in: 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2021, pp. 766–772.

[17] D. Mehta, A. Dwivedi, A. Patra, M. Anand Kumar, A transformer-based architecture for fake news classification, Social network analysis and mining 11 (2021) 1–12.

[18] C. Blackledge, A. Atapour-Abarghouei, Transforming fake news: Robust generalisable news classification using transformers, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 3960–3968.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[20] N. Rai, D. Kumar, N. Kaushik, C. Raj, A. Ali, Fake news classification using transformer based enhanced lstm and bert, International Journal of Cognitive Computing in Engineering 3 (2022) 98–105.

[21] A. Aggarwal, A. Chauhan, D. Kumar, S. Verma, M. Mittal, Classification of fake news by fine-tuning deep bidirectional transformers based language model, EAI Endorsed Transactions on Scalable Information Systems 7 (2020) e10–e10.

[22] R. Vijjali, P. Potluri, S. Kumar, S. Teki, Two stage transformer model for covid-19 fake news detection and fact checking, arXiv preprint arXiv:2011.13253 (2020).

[23] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system, arXiv preprint arXiv:2101.00180 (2021).

[24] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, J. Liu, Freelb: Enhanced adversarial training for natural language understanding, arXiv preprint arXiv:1909.11764 (2019).

[25] N. Akhtar, A. Mian, N. Kardan, M. Shah, Advances in adversarial attacks and defenses in computer vision: A survey, IEEE Access 9 (2021) 155161–155196.

[26] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020) 1–41.

[27] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.

[28] J. Chen, C. Jia, H. Zheng, R. Chen, C. Fu, Is multi-modal necessarily better? robustness evaluation of multi-modal fake news detection, IEEE Transactions on Network Science and Engineering (2023).

[29] S. Cresci, M. Petrocchi, A. Spognardi, S. Tognazzi, Adversarial machine learning for protecting against online manipulation, IEEE Internet Computing 26 (2021) 47–52.

[30] C. Koenders, J. Filla, N. Schneider, V. Woloszyn, How vulnerable are automatic fake news detection methods to adversarial attacks?, arXiv preprint arXiv:2107.07970 (2021).

[31] Z. Zhou, H. Guan, M. M. Bhat, J. Hsu, Fake news detection via nlp is vulnerable to adversarial attacks, arXiv preprint arXiv:1901.09657 (2019).

[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[33] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive

pretraining for language understanding, Advances in neural information processing systems 32 (2019).

[34] G. Chen, S. Ma, Y. Chen, L. Dong, D. Zhang, J. Pan, W. Wang, F. Wei, Zero-shot cross-lingual transfer of neural machine translation with multilingual pre-trained encoders, arXiv preprint arXiv:2104.08757 (2021).

[35] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[36] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[37] D. Jin, Z. Jin, J. T. Zhou, P. Szolovits, Is bert really robust? natural language attack on text classification and entailment, arXiv preprint arXiv:1907.11932 2 (2019).

[38] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, Bert-attack: Adversarial attack against bert using bert, 2020. URL: https://arxiv.org/abs/2004.09984. doi:10.48550/ARXIV.2004.09984.

[39] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Security and Privacy 1 (2018) e9.

[40] C. Yang, X. Zhou, R. Zafarani, Checked: Chinese covid-19 fake news dataset, Social Network Analysis and Mining 11 (2021) 58.

[41] A. Khalil, M. Jarrah, M. Aldwairi, M. Jaradat, Afnd: Arabic fake news dataset for the detection and classification of articles credibility, Data in Brief 42 (2022) 108141.

[42] H. Gómez-Adorno, J. P. Posadas-Durán, G. B. Enguix, C. P. Capetillo, Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task, Procesamiento del lenguaje natural 67 (2021) 223–231.

[43] P. H. A. Faustini, T. F. Covoes, Fake news detection in multiple platforms and languages, Expert Systems with Applications 158 (2020) 113503.

[44] K. Jain, A. Deshpande, K. Shridhar, F. Laumann, A. Dash, Indic-transformers: An analysis of transformer language models for indian languages, arXiv preprint arXiv:2011.02323 (2020).

[45] W. Antoun, F. Baly, H. Hajj, Arabert: Transformer-based model for arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).