From Hype to Reality: Transformer-Based Models for Fake News Detection Performance and Robustness Revealed

Dorsaf Sallami, Ahmed Gueddiche and Esma Aïmeur

(University of Montréal, Canada)

AiOfAi: 3rd Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies

at the 32nd International Conference on Artificial Intelligence (**IJCAI-23**)

Table of Contents

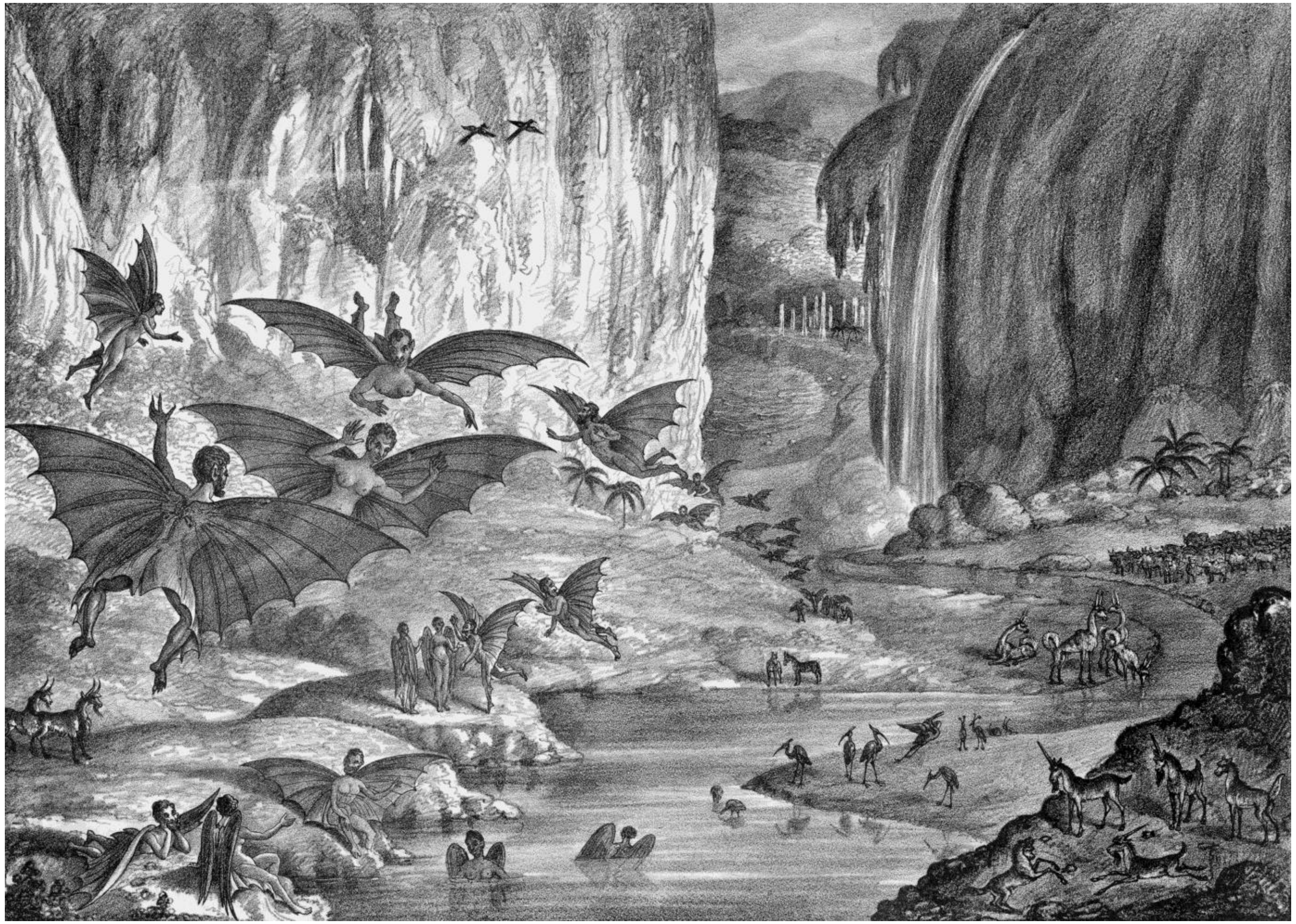
General Context (1/4): Fake News

> “ We live in a world with fake news being put out there. You don't really know what to trust, and it's a real danger to society. ”
>
> \- Austin Aries

Source: https://www.denverpost.com/2016/11/21/cartoons-of-the-day-fake-news/

General Context (2/4): An Old Problem

The Great Moon Hoax by the tabloid The Sun from 1835.

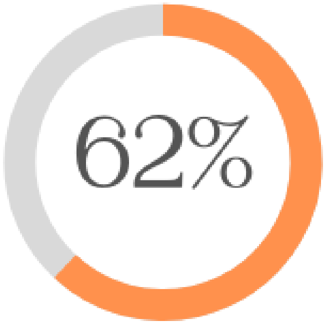


Source: https://www.cits.ucsb.edu/fake-news/brief-history

General Context (3/4): But conditions have changed….

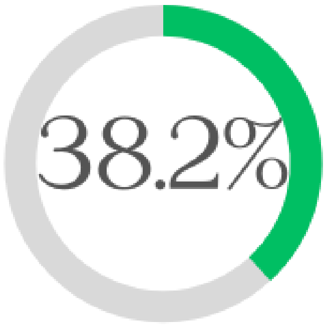
General Context (4/4):

Don't Believe Everything You Read

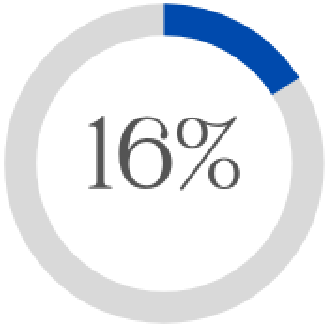


of all Internet information can be fake

of US adults have consumed fake news



of Americans have accidentally shared fake news



say they've shared a story they later realized was fake

Statistics from : https://techjury.net/blog/fake-news-statistics/

Table of Contents

Related Work(1/6): Widespread Adoption of Transformers

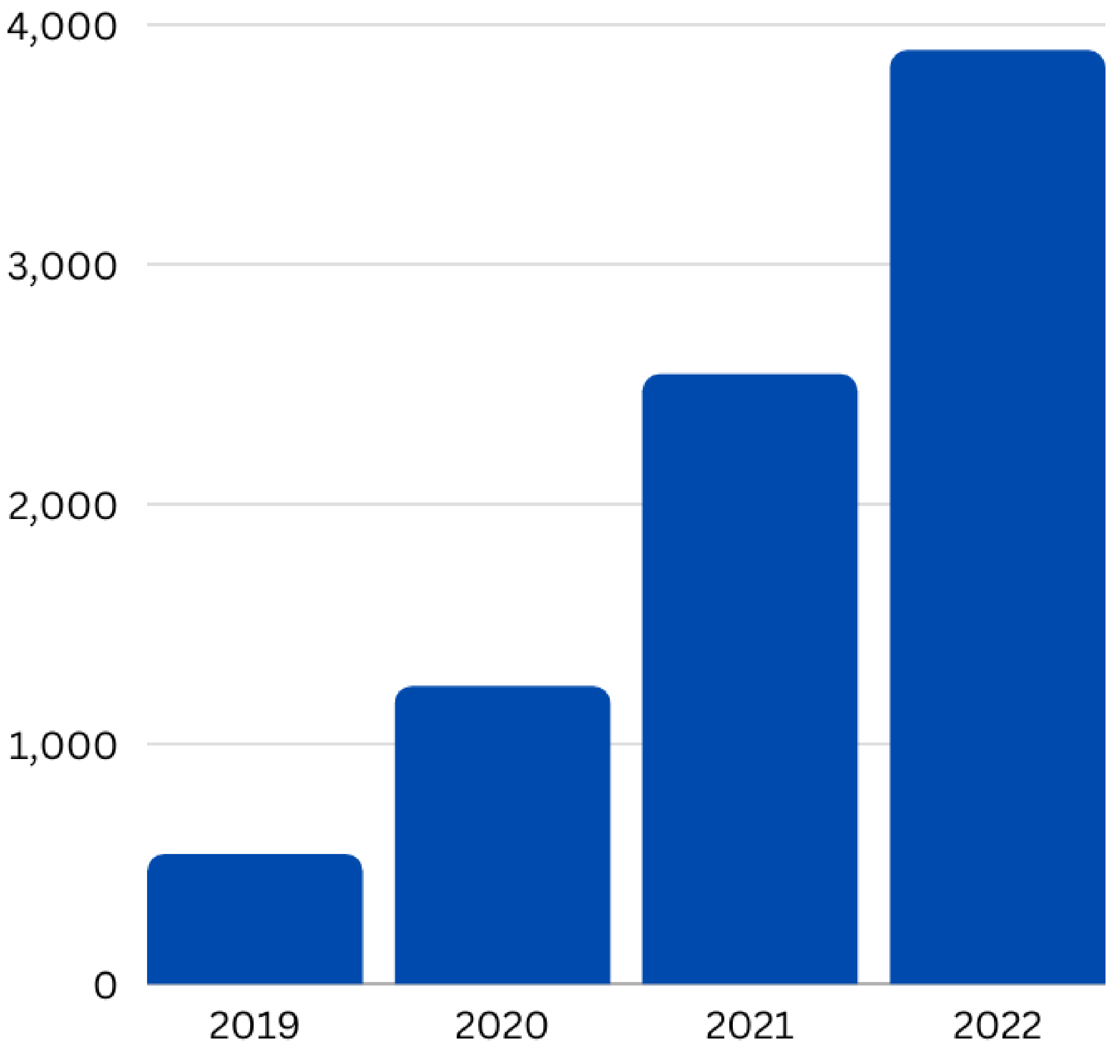


Figure 1: Number of papers retrieved from Google Scholar.

Related Work(2/6): What Makes Transformers So Powerful?

- Understand the relationship between sequential elements that are far from each other.
- Way more accurate.
- Equal **attention** to all the elements in the sequence.
- ....

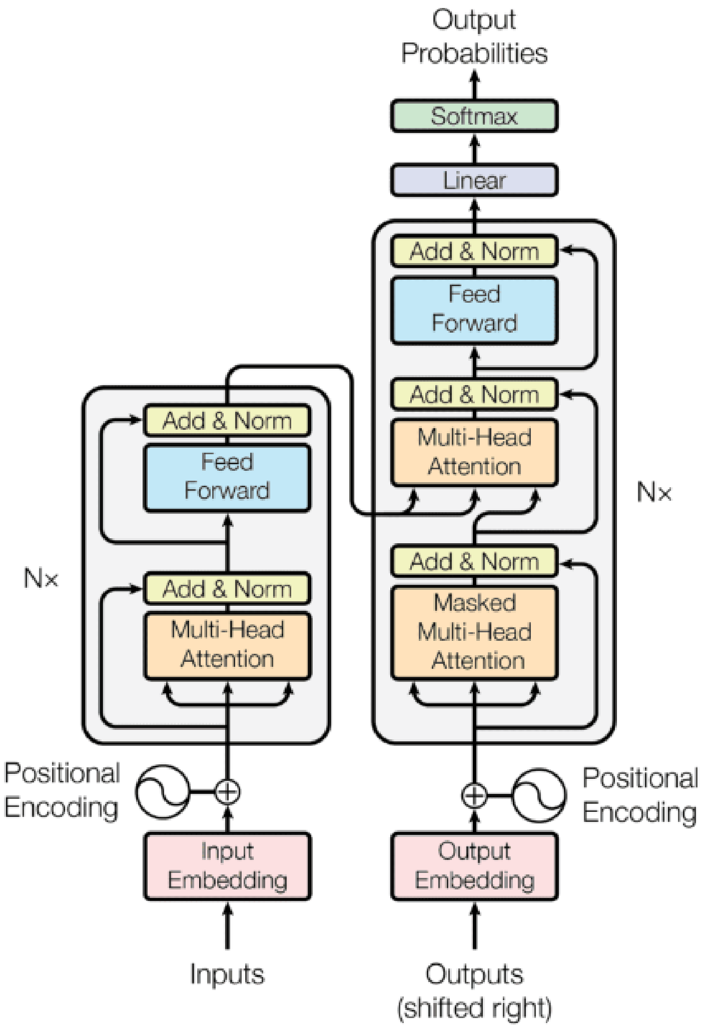
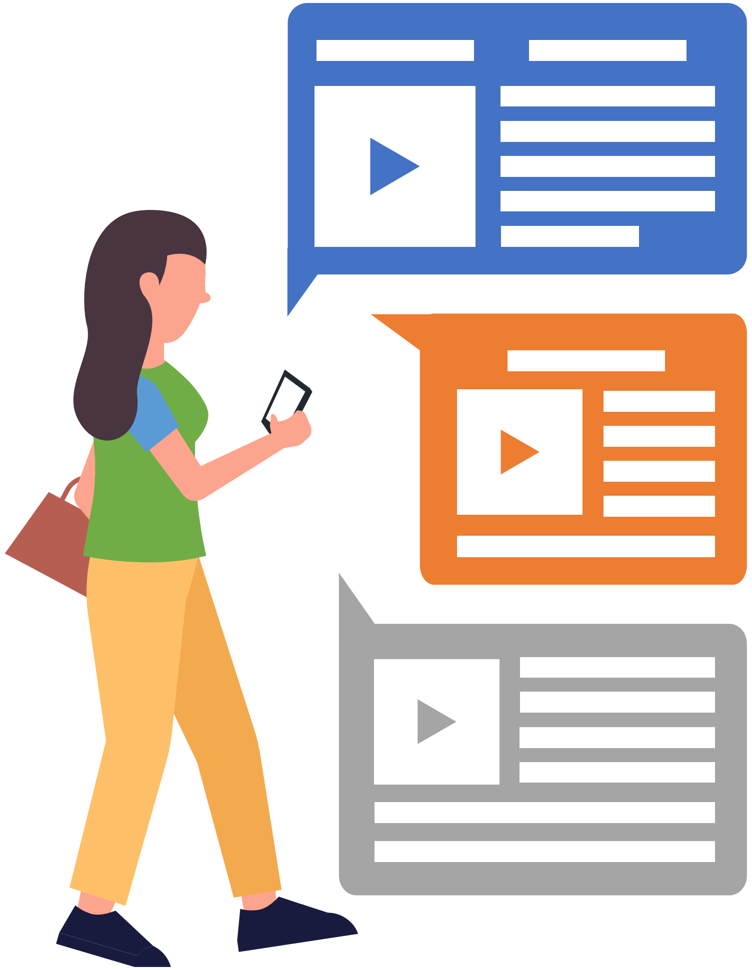


Figure 2: The Basic Architecture (Vaswani et al., 2022).

Related Work(3/6): Diverse Adopted Approaches



Pre-trained transformers (Hande et al., 2021; Mehta et al., 2021; Blackledge and Atapour-Abarghouei, 2021).

Adapted transformers (Rai et al., 2022; Aggarwal et al., 2020).

Domain-specific transformers (Vijjali et al., 2020; Gundapu and Mamidi, 2021).

Related Work(4/6): Trade-off Between Robustness and Accuracy

- Transformers have shown remarkable **accuracy** in identifying fake news.
- Choosing accuracy over **robustness** might miss tricky fake news, and concentrating too much on robustness could reduce overall accuracy.
- Achieving an optimal **equilibrium** between robustness and accuracy ensures effective fake news detection and contributes to the fight against misinformation.

Related Work(5/6): Adversarial Attacks

- Any attempt to fool a deep learning model with a **deceptive** input.
- First seminal work : “Intriguing properties of neural networks” by (Szegedy et al. 2013).
- Especially reaserched in image recognition, but can also be applied to audio, text or tabular data.
- Less focus on textual data (Koenders et al., 2021).

Source: https://www.analyticsvidhya.com/blog/2022/09/machine-learning-adversarial-attacks-and-defense/

Related Work(6/6): Why should we care about Robustness ?

- When building models, we mostly focus on classification effectiveness / minimizing the error.
- Little work on model security and robustness.
- Some of these attacks are 100% effective in fooling normal neural networks.
- E.g : The classification accuracy of GoogLeNet on MNIST under adversarial attacks drops from **98% to 18%** (for ProjGrad attack) or 1% (DeepFool attack).

Attack	Lenet				
	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
Noise	MNIST	0.984	1.0	0.9858	1.0
	ILSVRC2012	NA	NA	NA	NA
	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
Semantic	MNIST	0.233	0.645	0.986	1.0
	ILSVRC2012	NA	NA	NA	NA
	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
Fast Gradient Sign Method	MNIST	0.509	0.993	0.986	1.0
	ILSVRC2012	NA	NA	NA	NA
	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
Projected Gradient Descent	MNIST	0.187	0.982	0.986	1.0
	ILSVRC2012	NA	NA	NA	NA
	Dataset	Acc@1 w/	Acc@5 w/	Acc@1 w/o	Acc@5 w/o
DeepFool	MNIST	0.012	1.0	0.9858	1.0
	ILSVRC2012	NA	NA	NA	NA

Table of Contents

Methodology (1/8): Contributions

1. Assessing various cutting-edge transformer models to evaluate the impact of various designs on the same dataset.
2. Assessing their performance and effectiveness in detecting fake news in different languages.
3. Exploring the robustness of these transformer models by implementing adversarial attacks.
4. Developing an interactive interface that allows us to visualize and explore our experimental results.

Methodology (2/8): End-to-End Architecture

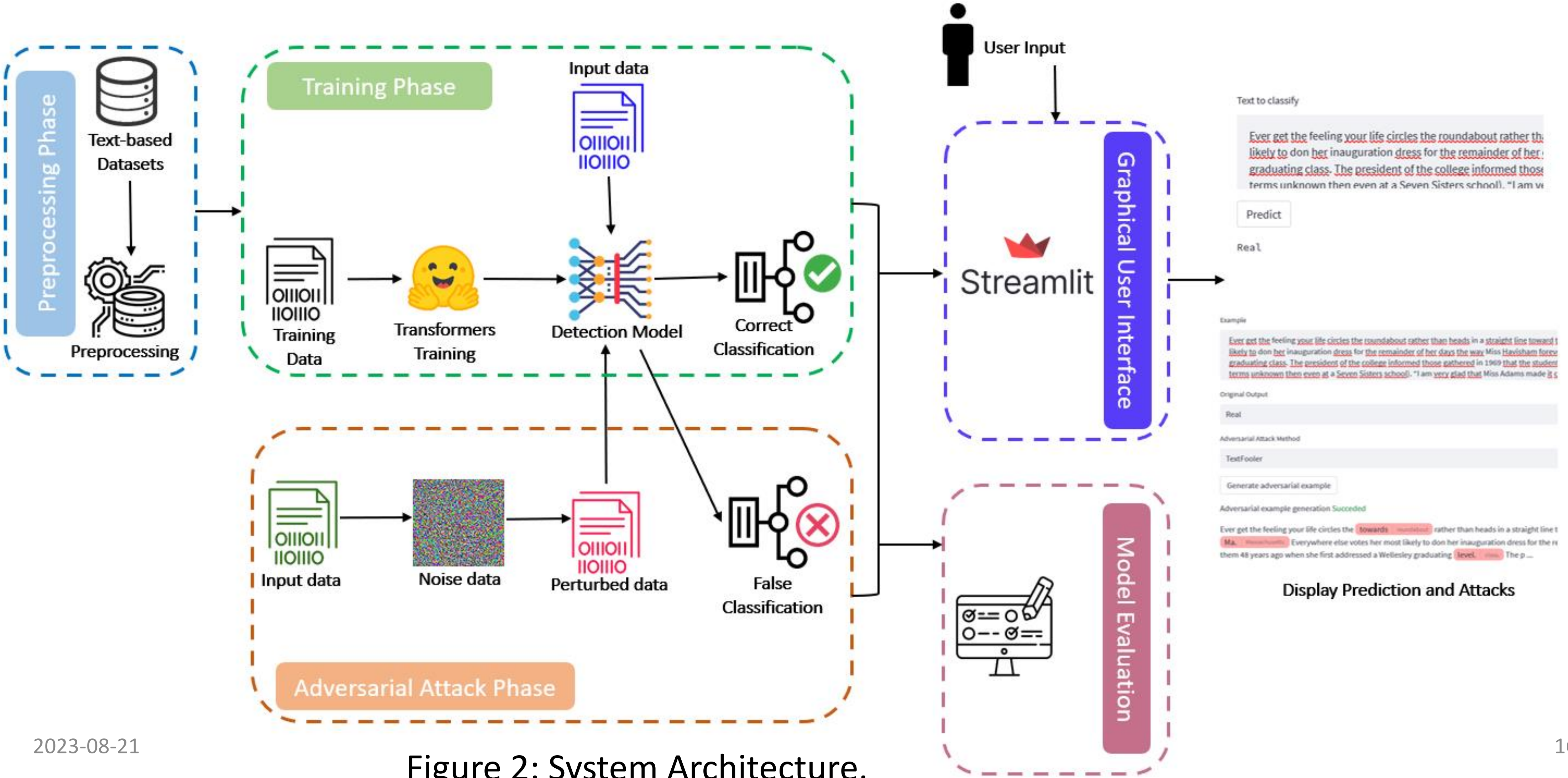


Figure 2: System Architecture.

Methodology (3/8): Models Architectures

Hugging Face

1 **BERT** (Bidirectional Encoder Representations from Transformers)

- Pre-trained by masking words in a sentence to understand context.
- Introduced attention mechanism for contextual word embeddings.

2 **DistilBERT** (Distill + BERT)

- A distilled version of BERT with reduced parameters.
- Maintains similar performance while being more resource-efficient.

3 **RoBERTa** (A Robustly Optimized BERT Pretraining Approach)

- An optimized version of BERT with modified pre-training techniques.
- Achieves state-of-the-art performance on various NLP tasks.

Methodology (4/8): Models Architectures

Hugging Face

4

XLNet

- Utilizes generalized autoregressive techniques to enable bidirectional context learning.
- Demonstrated superior performance over BERT in various tasks, including question answering,, sentiment analysis, … .

5

GPT-2

- Developed by OpenAI in February 2019.
- It has the capability to translate text, answer questions, summarize passages, and generate text.

6

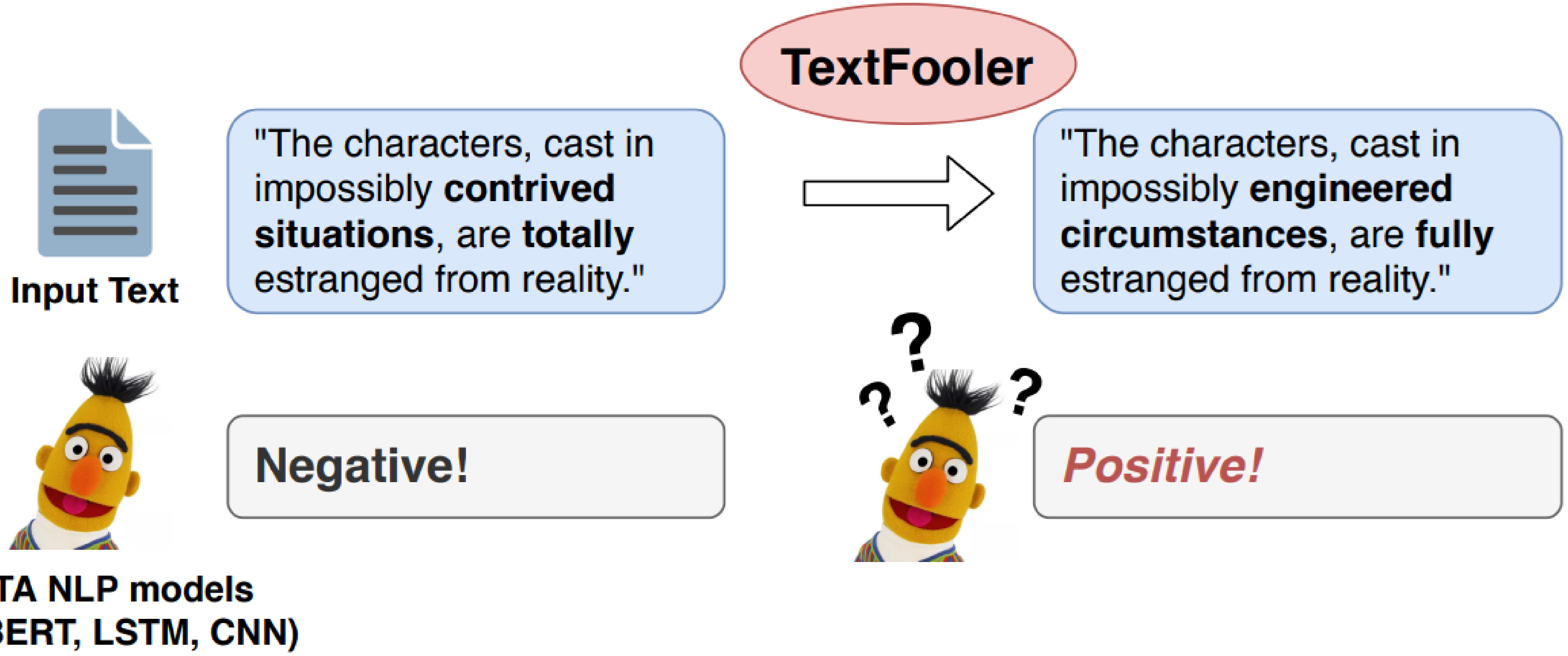
GPT-J

- Developed by EleutherAI.
- Comparable performance to OpenAI’s GPT-3 across various zero-shot downstream tasks and even surpasses it in code generation tasks

Methodology (5/8): Adversarial Attacks

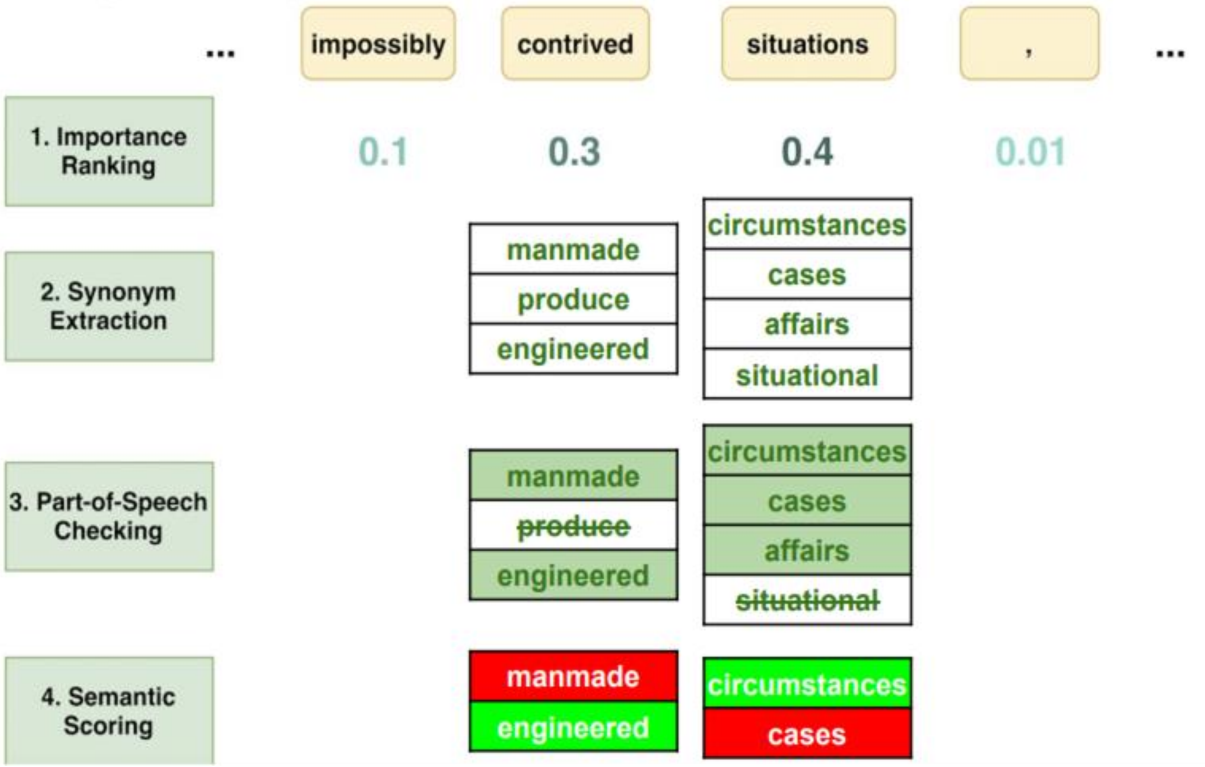
TextFooler

- Adversarial examples generation approach for text data.
- "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment" (Di Jin et al. ,2019)



Methodology (6/8): Adversarial Attacks

Input: The words are placed in incredibly forced situations, becomes completely divorced from reality.

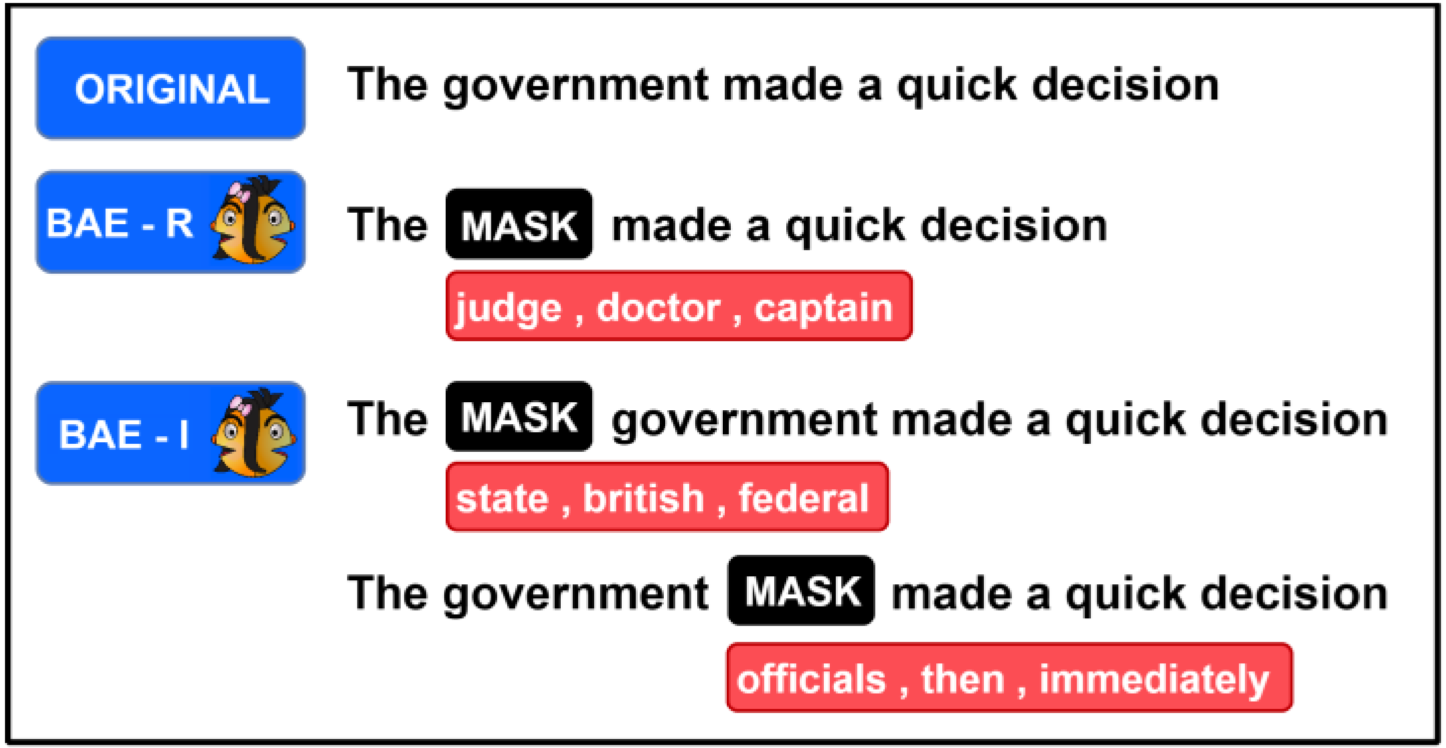


Output: The words are crafted within unrealistically engineered circumstances, is entirely detached from reality.

Methodology (7/8): Adversarial Attacks

Bert-Attack

- Adversarial examples generation approach using Bert.
- “BERT-ATTACK: Adversarial Attack Against BERT Using BERT”, (Linyang Li. , 2020)



Methodology (8/8): Adversarial Attacks

1/ Finding Vulnerable Words

- Select words in the sequence which have a high significance influence on the final output logit (o)
- Word Importance:

$$I_{w_i} = o_y(S) - o_y(S_{\backslash w_i})$$

2/ Word Replacement via BERT

Iteratively replace the words in list one by one to find perturbations that can mislead the target model.

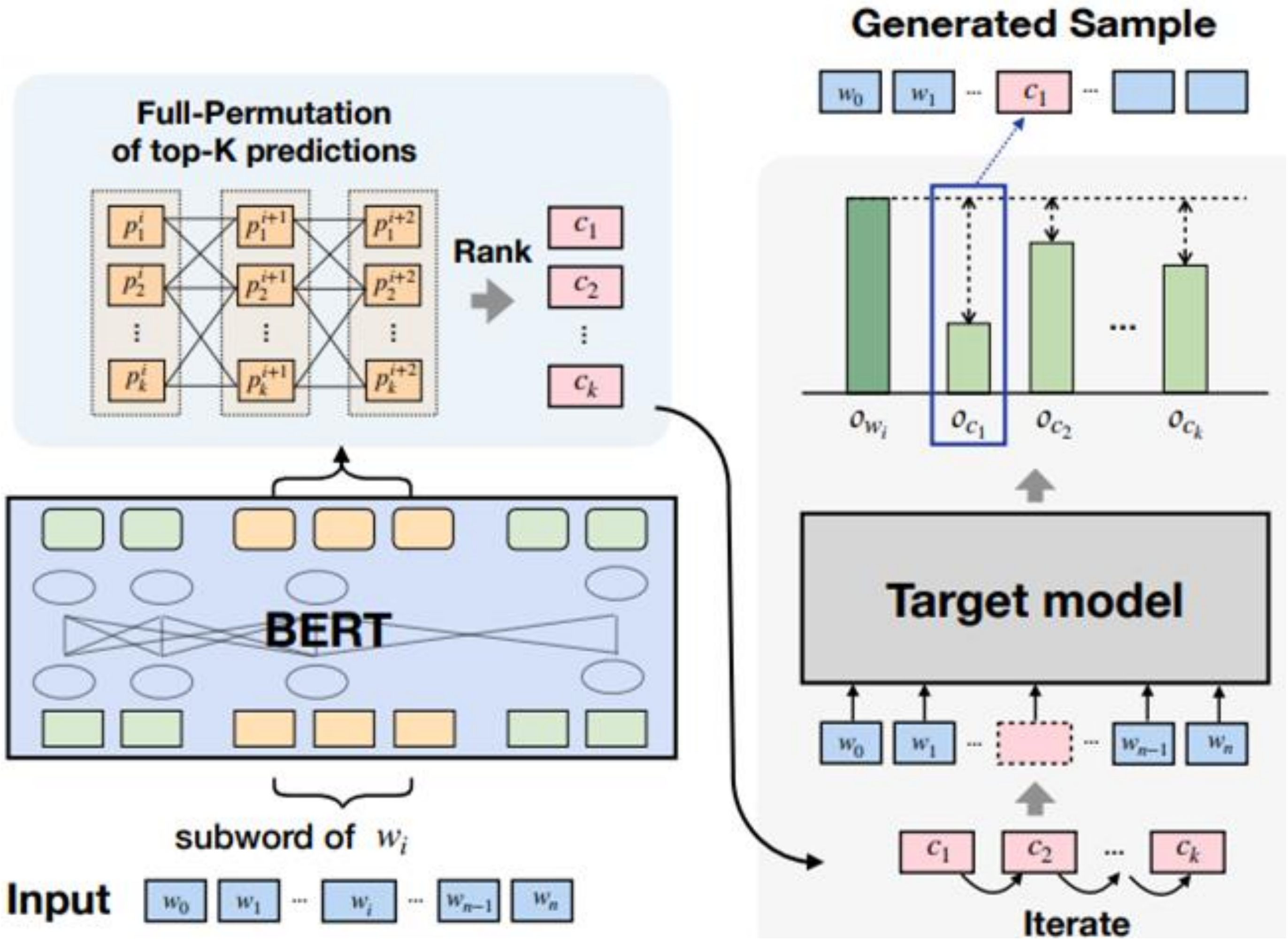


Table of Contents

Experiments (1/4): Datasets

Table : 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

Assess the performance of transformers across different languages, expanding beyond the conventional English-focused evaluations

Experiments (2/4): Experimental Setup

- Google Colab Pro
- Pytorch
- Hugging Face transformers

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

Experiments (3/4): Evaluation Metrics

Classification task evaluation:

- Accuracy
- Precision
- Recall
- F1-Score

Adversarial attacks evaluation:

- Number of successful attacks
- Accuracy under attack
- Average perturbed words (%)
- Average number of queries

Experiments (4/4): GUI Implementation

The Streamlit library is employed to create a **user-friendly** interface for visualizing and analyzing experimental results

Streamlit

Results (1/5): Assessing Performance

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

Results (1/5): Assessing Performance

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

Results (1/5): Assessing Performance

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

Results (1/5): Assessing Performance

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

Results (1/5): Assessing Performance

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

Results (2/5): Evaluating Robustness

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

Results (2/5): Evaluating Robustness

Table 4: Results of the Kaggle dataset attack utilizing DistilBERT.

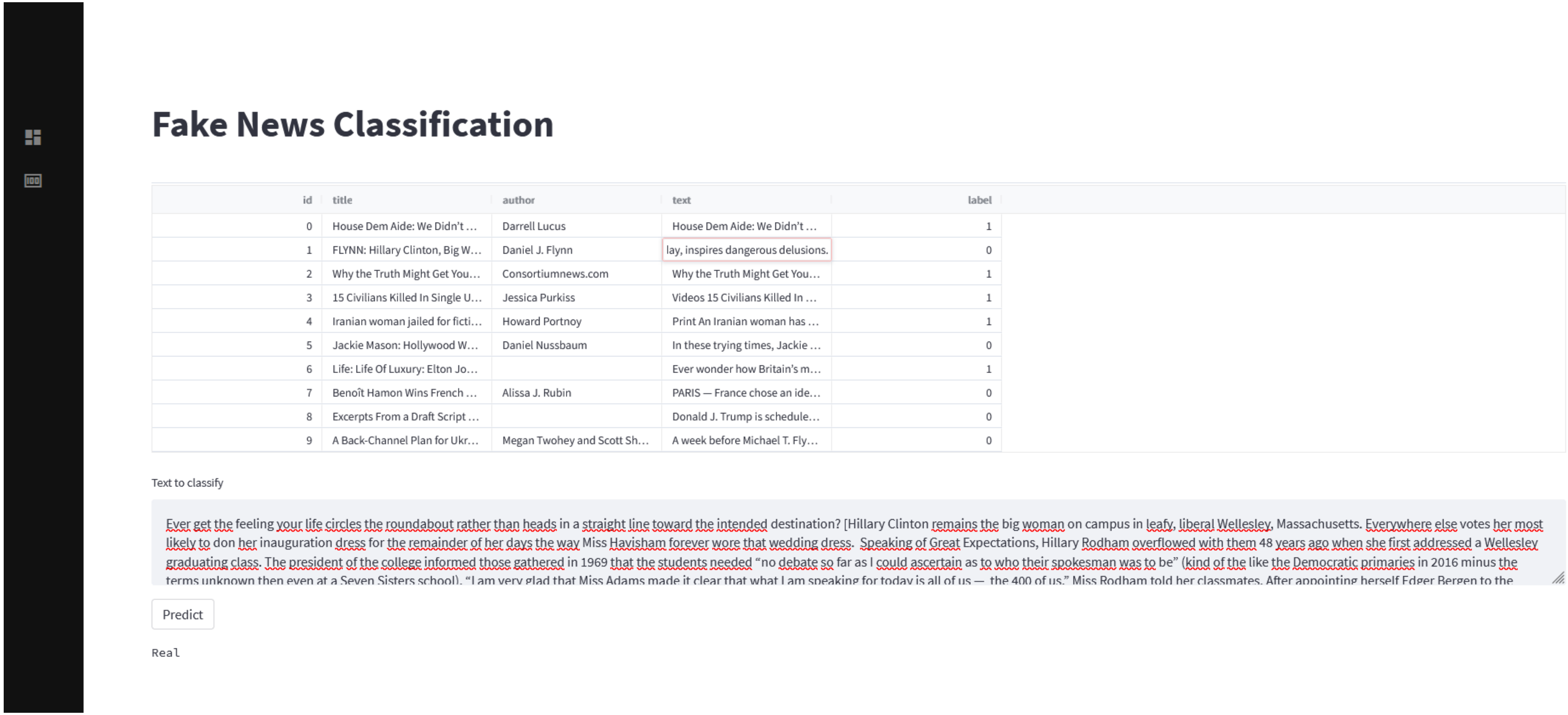
Metrics	TextFooler	BAE
Number of successful attacks	1730	2036
Number of failed attacks	896	590
Original accuracy	99%	99%
Accuracy under attack	31.73%	20.66%
Attack success rate	65.88%	77.53%
Average perturbated word	25.34%	27.53%
Average num. words per input	11.72%	11.72%
Avgerage num. queries	76.97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

Results (3/5): Adversarial Examples

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

Results (4/5): Interactive Interface



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

Predict

Real

Results (5/5): Interactive Interface

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 ...

Discussion (1/3): Dataset Variability

- Performance variation across datasets indicates dataset characteristics affect model performance significantly.

- The **language** factor plays a crucial role in a model's generalization and performance on a particular dataset.

- Transformer architecture advancements have been primarily tested and reported on high-resource languages like English (Kaggle dataset).

- Dataset **size** is another important factor to consider; the FNCS dataset with only 1248 items led to decreased model performance.

Discussion (2/3): Trade-off between Precision and Recall

- Variations in F1 scores on the CHECKED dataset suggest potential trade-offs between precision and recall.

- XLNet achieves high precision and recall.

- BERT and DistilBERT strike a slightly different balance, leading to slightly higher F1 scores.

Discussion (3/3): The Impact of Adversarial Attacks

- DistilBERT is among the best-performing models. However, the results expose its **vulnerability** to adversarial attacks.
- When subjected to TextFooler and BAE attacks, the model's accuracy significantly decreases.
- These findings highlight the importance of addressing DistilBERT's susceptibility to adversarial examples to ensure its reliability and robustness in real-world scenarios.

Table of Contents

Conclusion and Future Work

- Transformers' accuracy **vulnerable** to adversarial attacks.
- **Language** of training datasets impacts Transformers' performance.
- Comparative evaluation of models and attack techniques.
- Interface developed for visualizing experimental results.
- **Need** for reliable detection methods against fake news.

Future research:

- ➢ Investigating models' resilience.
- ➢ Identifying vulnerabilities.
- ➢ Implementation of adversarial training for protection.

References

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
- [2] 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
- [3] 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.
- [4] 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.
- [5] 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.
- [6] A. Aggarwal, A. Chauhan, D. Kumar, S. Verma, M. Mittal, Classification of fake news by finetuning deep bidirectional transformers based language model, EAI Endorsed Transactions on Scalable Information Systems 7 (2020) e10–e10
- [7] 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).
- [8] S. Gundapu, R. Mamidi, Transformer based automatic covid-19 fake news detection system, arXiv preprint arXiv:2101.00180 (2021).
- [9] 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).
- [10] 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).
- [11] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, Bertattack: Adversarial attack against bert using bert, 2020.

Thank you!