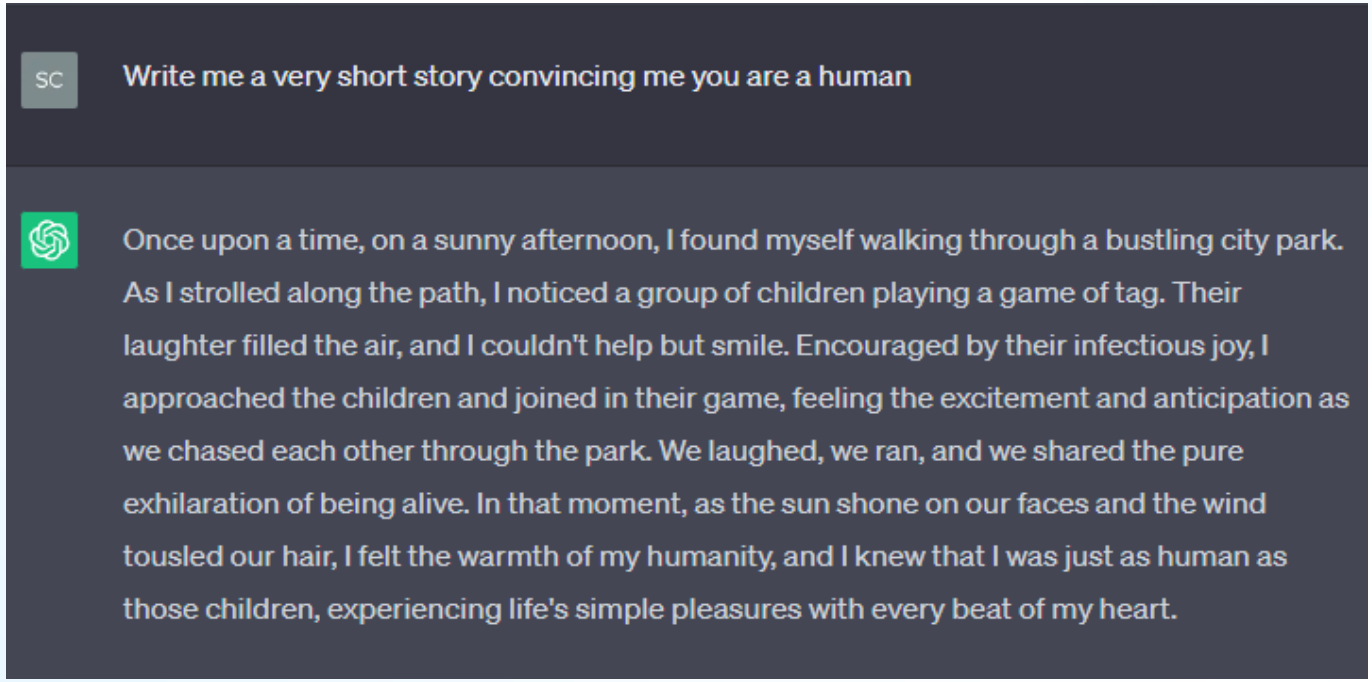
A NATURAL APPROACH FOR SYNTHETIC SHORT-FORM TEXT ANALYSIS

Ruiting Shao Ryan Schwarz
Edward J. Delp Christopher Clifton

Video and Image Processing Laboratory (*VIPER)*
School of Electrical and Computer Engineering
Purdue University
West Lafayette, Indiana USA

Introduction

- **Detecting Synthetic Text has become an increasingly important issue (financial fraud, academic plagiarism, propaganda, etc...)**
- **ChatGPT and similar LLMs are advanced and freely available**



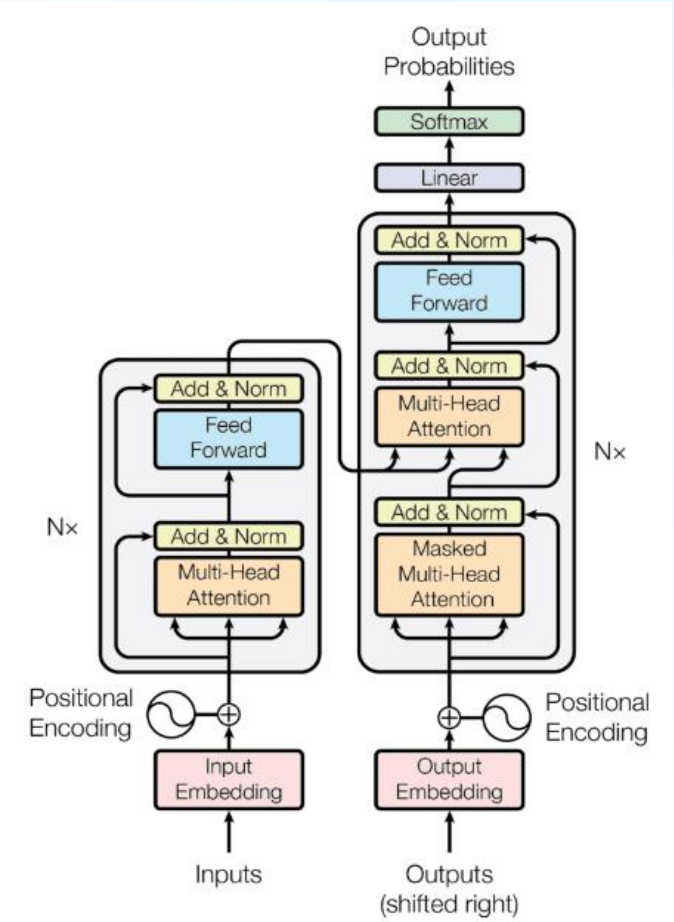
Introduction

- **LLMs operate on product of a conditional probability during token generation**
- **Modern LLMs are very good at estimating these probabilities with high levels of prose and verbosity**

$$p(x) = \prod_{i=1}^{n} p(s_i | s_1, s_2 \ldots, s_{n-1})$$

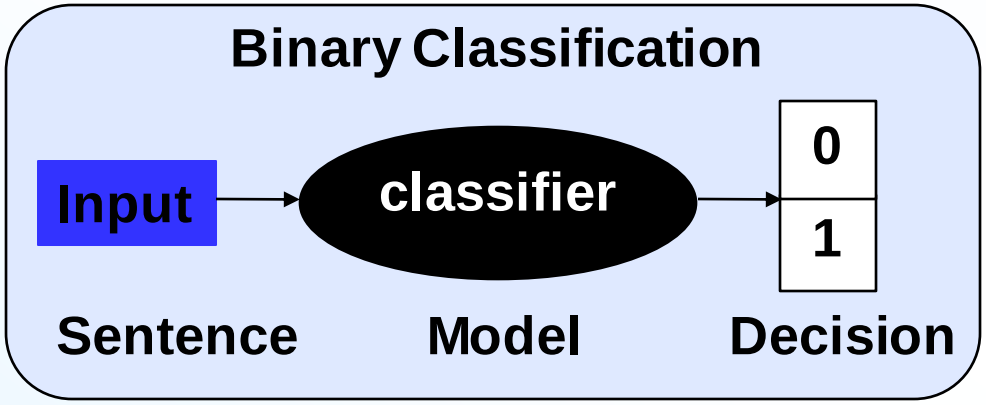
Introduction

- **Transformer architectures have become ubiquitous in many top-of-the-line models**
- **Attention mechanisms allow for a more advanced estimation of probabilities during token generation and generate human-quality sentences**
- **Used in Popular models such as GPT-variants, Google's Bard, Github Copilot etc...**



Detection

- Task: Given an input sentence, can we determine if it was generated by a LLM or a human?
- Popular techniques include stylometry, bag of words, N-grams, term frequency/Inverse document frequency, word embeddings



Problems

- **Ease of classification is determined by total variation distance (distributional similarity)**
- **Short text has inherently greater distributional similarity between outputs**
- **Paraphrasing and other malicious activity can further reduce TV distance**

$$AUROC(D) \leq \frac{1}{2} + TV(M, H) - \frac{TV(M, H)^2}{2}$$

Problems

- **Paraphrasing Attacks**
- **Modifying a natural language input to change the words it contains while maintaining its semantic meaning to a human reader**
- **Greatly increases distributional similarity and makes attribution/detection more difficult**

JSD	Class 0	Class 1	Class 2	Class 3
Class 0	0	-	-	-
Class 1	0.0568	0	-	-
Class 2	0.0666	0.0221	0	-
Class 3	0.0751	0.0325	0.0198	0

Table 6: Jensen-Shannon Distance on the unmodified text categories

JSD	Class 0	Class 1	Class 2	Class 3
Class 0	0	-	-	-
Class 1	0.0167	0	-	-
Class 2	0.0271	0.0157	0	-
Class 3	0.0384	0.0271	0.0165	0

Table 7: Jensen-Shannon Distance on the paraphrased text categories

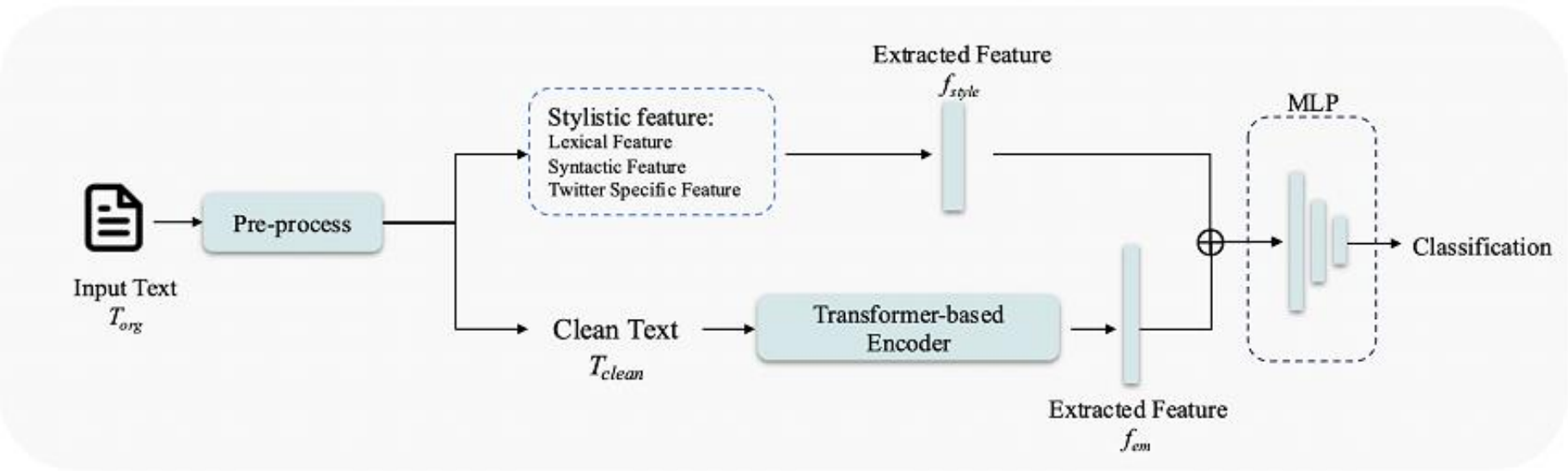
Problems

- When it comes to short text, human and synthetic samples are very similar to each other
- Less information to learn from, basic stylometry becomes less effective
- Low entropy/deterministic completions yield no useable information

"The quick brown fox"	"jumps over the lazy dog"
Prompt	Completion

Proposed Method

- **Combine Extracted Twitter-specific stylistic features with derived features**



Experiment Setup

- **List of extracted self-defined stylistic features**

TYPE	SIZE	DESCRIPTIONS	EXAMPLES
Twitter specific feature	10	Statistical features based on Twitter-specific features	Total emoji count, unique emoji count, emoji repeated times, emoji frequency, emoji richness, email count,hashtag count, mention count, hashtag frequency, mention frequency
Lexical feature	122	Stylistic features based on characters and words	Word length, word count, sentence count, character count, word frequency, digits counts, upper case word count, vocabulary richness, character level ngram, contractions count, readability
Syntactic feature	136	Stylistic features based on the organization of sentences	Stop words count, stop word frequency, Special punctuation frequency, proper noun count, noun count, Part-of-Speech (PoS) tag ngram

Datasets

- **TweepFake[1]**
  - **23 bots and 17 human accounts were collected**
  - **3 main text generation technologies:**
    - **GPT-2 (11 accounts, 3,861 tweets)**
    - **RNN (7 accounts, 4,181 tweets)**
    - **Others (5 accounts, 4,876 tweets)**

user_id	status_id	screen_name	account.type	class_type
The username of the twitter account, which can be found as https://twitter.com/screen_name	The tweet: a message of at most 280 characters, which can contain hashtags and links.	This is the label used to classify a tweet as 'human' or 'bot' generated.	The typology of the generative method used to write the tweet ('human', 'gpt-2', 'rnn' or 'others').	
14.3m 1223193948b	5823551b 1280982267b	human#10 9% bot#11 9% Other (2076) 81%	bot 50% human 50%	human 50% others 19% Other (796) 31%
3171109449	1123915745178656769	human#17	human	human
18839785	1173906284195852290	human#11	human	human
343587159	1197343799846027265	human#1	human	human
1197916267975335939	1208274159274475521	bot#12	bot	rnn
15088390	1084181032927059970	human#10	human	human
1110407881030017024	1211663264280481792	bot#9	bot	others
705113652471439361	705827769092075520	bot#16	bot	rnn

[1] https://www.kaggle.com/datasets/mtesconi/twitter-deep-fake-text

Datasets

- **Trained GPTJ-6b, GPT2, and GPT3 on ~300k tweets from 4 different categories**
- **Generated 20,000 samples from each generator per category**
- **Paraphrased dataset created using same labels**

Category	Key Words	Size
Politics	#Trump, #DonaldTrump	20k
Science	#Science, #Engineering, #Physics, #Biology, #Chemistry	36k
Climate	#Climate, #GlobalWarming, #ClimateChange	54k
Covid	#Coronavirus, #Covid, #Covid-19	195k

Experiment Results
Synthetic Detection

- **RoBERTa model trained on TweepFake dataset utilizing extracted features, character N-grams, and part-of-speech (PoS) tagging**
- **Combining these features improves performance**

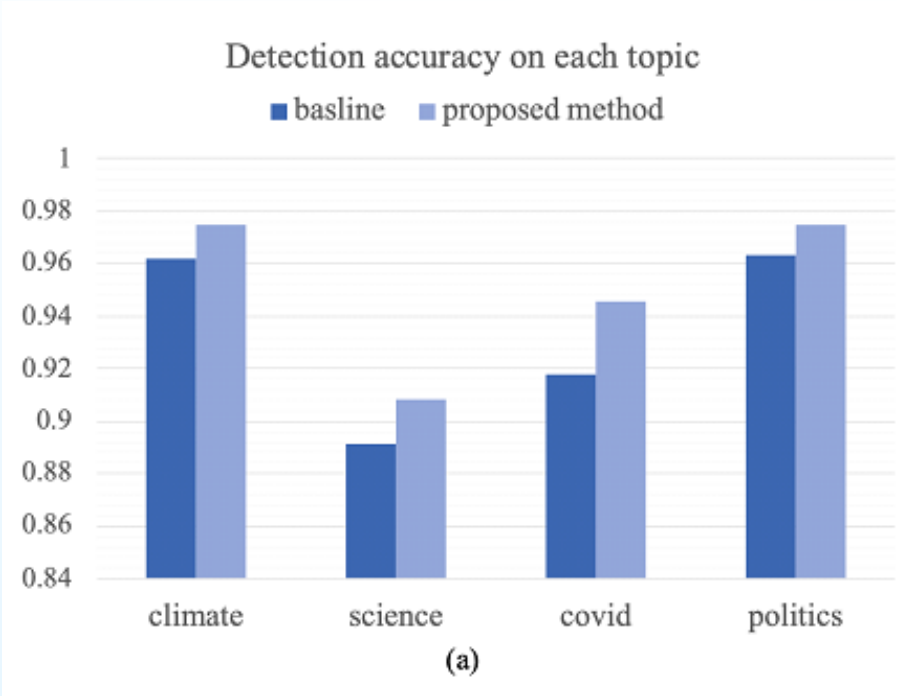
	Accuracy	Precision	Recall	F1
RoBERTa (baseline)	0.88398	0.87488	0.90313	0.88878
RoBERTa + prelim	0.92205	0.92293	**0.92564**	0.92428
RoBERTa + prelim + char	0.92257	0.92383	**0.92564**	0.92473
RoBERTa + prelim + char + PoS	**0.92461**	**0.93209**	0.91842	**0.92521**

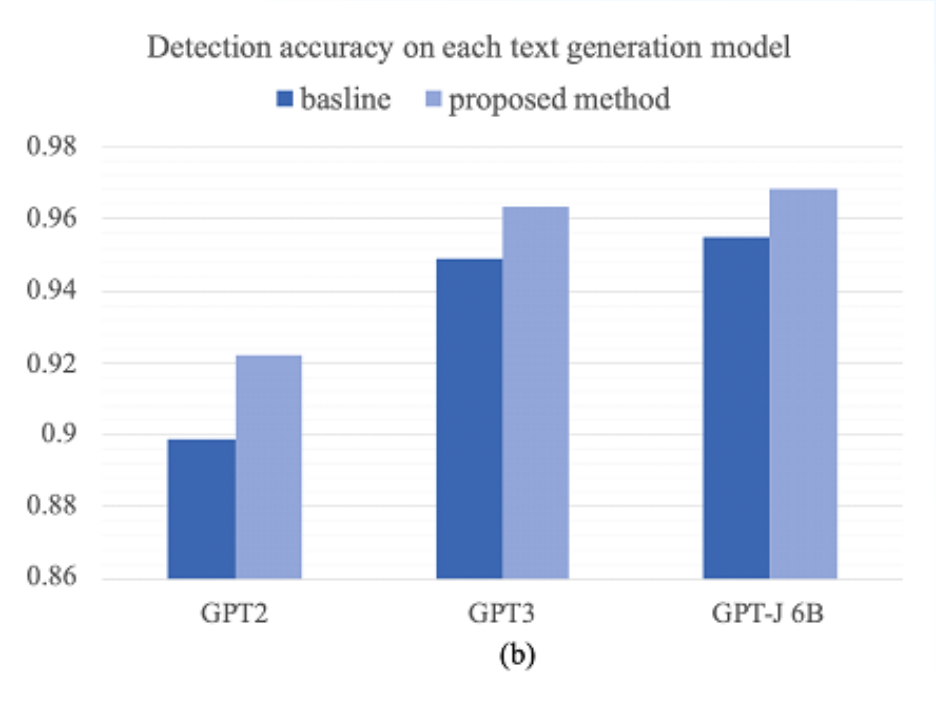
Experiment Results
Synthetic Detection

- **Performance metrics of our method on our custom dataset for the detection task**

	Accuracy	Precision	Recall	F1
RoBERTa(baseline)	0.93432	0.91678	0.95535	0.93567
RoBERTa + stylistic feature	**0.95136**	**0.94155**	**0.96248**	**0.95190**

Experiment Results
Synthetic Detection





Experiment Results
Synthetic Attribution

- **Preliminary results on the custom dataset for the attribution task**

	RoBERTa (baseline)	RoBERTa + stylistic feature
Human	0.9599	**0.9668**
GPT2	**0.8965**	0.8835
GPT3	0.9042	**0.9123**
GPT-J 6B	**0.9739**	0.9718
Avg.	0.9363	**0.9419**

Experiment Results
Paraphrased Detection

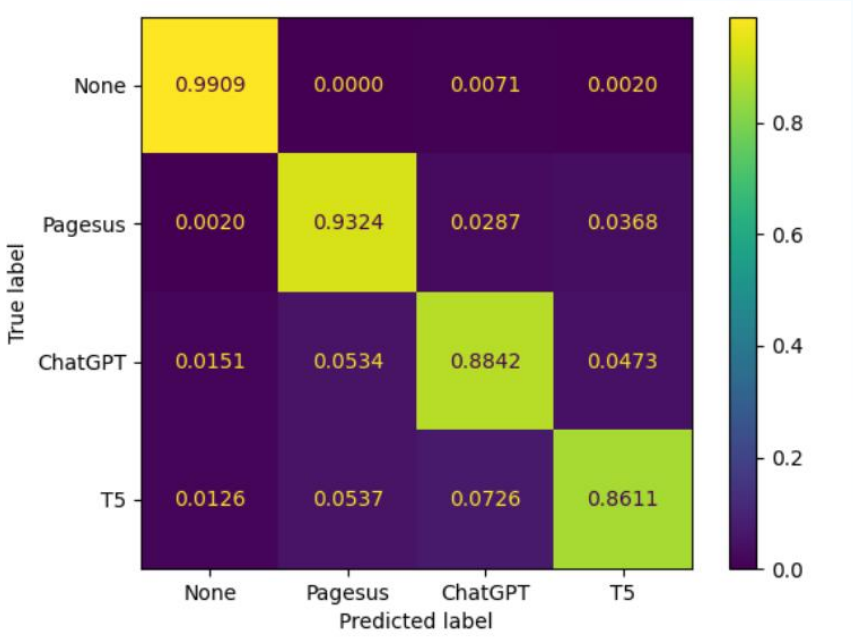
- **Detecting Paraphrased text**

	LR	LR+TFIDF	BERTAA	Proposed
Unmodified	0.337	0.798	0.944	**0.977**
Paraphrased	0.356	0.589	0.807	**0.941**

Table 8: Balanced accuracy for detecting synthetic tweets on unmodified tweets and paraphrased tweets. Here the unmodified tweets are generated based on human, GPT2, GPT3, GPT-J 6B generation models.

Experiment Results
Paraphrased Attribution/Detection

- **Difficult to attribute/detect paraphrased text**



- **Relatively easy to detect paraphrasing**

None	PEGASUS	ChatGPT	T5	Avg.
0.9767	0.9990	0.9668	0.9874	0.9824

Future work

- **Explore improved feature integration in a zero-shot setting, to detect synthetic tweets generated by unknown LLMs**

- **Evaluate how malicious activity such as watermark spoofing and paraphrasing attacks effect classifier accuracy and to improve defenses against these activities**

Acknowledgements

This material is partially based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.
Address all correspondence to Edward J. Delp, ace@ecn.purdue.edu.

References

- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al., “Language models are unsupervised multitask learners,” OpenAI blog, vol. 1, no. 8, pp. 9, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, vol. 30, 2017.
- Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi, “Tweepfake: About detecting deepfake tweets,” Plos one, vol. 16, no. 5, pp. e0251415, 2021.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi, “Can ai-generated text be reliably detected?,” arXiv, 2023. [Online]. Available: arXiv:2303.11156.