MMU NLP at CheckThat! 2024: Homoglyphs are Adversarial Attacks

Notebook for the CheckThat Lab at CLEF 2024

Charlie Roadhouse[1,*], Matthew Shardlow[1] and Ashley Williams[1]

[1]Manchester Metropolitan University, All Saints Campus, Metropolitan University, Manchester M15 6BH

Abstract

Work on adversarial attacks is becoming more necessary with the growing amount of misinformation online, as it can bolster future defence techniques. In this paper, we present a character-based approach for 2024 CheckThat! Lab Task 6, InCrediblAE. We pair an importance-based search technique with a homoglyph attack to replace characters with glyphs that are imperceptible to human readers but elicit a change in classification. The model is tested against three separate victim models across five distinct misinformation domains: COVID-19 misinformation, fact-checking, rumour detection, propaganda detection and style-based news bias. Results show that our model has comparable performance to the current state of the art, even exceeding current results in certain cases.

Keywords

Adversarial Attacks,
Homoglyphs,
Character-based Attacks,
Semantic Preservation

1. Introduction

In this paper, we focus on character-based homoglyph attacks, where visually similar characters are substituted to deceive the classification model while maintaining the appearance of the original text. This method aims to strike a balance between high attack success rates and maintaining semantic integrity. We compare our approach against state-of-the-art models, demonstrating its effectiveness and robustness across various domains. All our code is available on GitHub.[1]

Misinformation online is one of the challenges facing social media platforms today [1]. This has led to a large body of research on how to combat this [2]. Machine learning (ML) and other AI techniques are seen as the answer to solving the current issues of misinformation [3]. However, these solutions are not without flaws, with research showing that ML models can be susceptible to adversarial attacks [4]. By generating adversarial examples, it is possible to create new versions of the text that appear legitimate to a human reader but are given a different classification than the original.

By participating in the IncrediblIAE shared task at the CheckThat! lab of CLEF 2024, we hope to contribute to the growing body of research on adversarial attacks for classifying misinformation.

2. Background

There is a growing body of research focused on making classification models more robust against adversarial attacks [5, 6]. Some researchers view the generation of adversarial examples as a data augmentation task, aiming to train future models on adversarial data to improve their robustness [7]. Others focus on developing new defence methods based on evolving attack strategies [8].

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

charlie.roadhouse@mmu.ac.uk (C. Roadhouse); m.shardlow@mmu.ac.uk (M. Shardlow); ashley.williams@mmu.ac.uk (A. Williams)

[1]https://github.com/chroadhouse/MMU_NLP-HomoGlyph-Attack

Adversarial attack strategies can be broadly categorized into three types: character-based, word-based, and sentence-based attacks.

Character-based attacks modify individual characters within words using techniques such as insertion, deletion, replacement, and transposition. This method tends to preserve the original semantics well, as demonstrated by models like HotFlip [9] and TextFooler [10], though it can often be detected by spell checkers, making it less robust [11].

Word-based attacks affect entire words, employing techniques such as word removal, replacement, and insertion. This approach typically preserves semantics by using synonyms of the original word, but the resulting sentences can sometimes be easily identified by humans due to grammatical inconsistencies. Examples include BERTattack [12] and BAE [13].

Sentence-based attacks modify sentences either wholly or partially, using techniques like paraphrasing to generate reworded versions of given phrases [14, 15]. This level of attack can cause issues with the amount of semantic preservation retained [8].

A key aspect of adversarial attack methodologies is the search technique used to identify which parts of the text to attack. Two main techniques are prevalent in research: importance-based and gradient-based.

Importance-based search techniques generate perturbations only on words that are crucial for the classifier's decision. State-of-the-art models like BERTattack [12], BAE [13], and A2T [16] use the BERT model to mask words in the sequence, ranking the most important words to attack based on classification scores of these masked sequences.

Gradient-based search techniques, similar to the Fast Gradient Sign Method (FGSM) [17] used in adversarial image generation, are adapted to text by identifying perturbations that best affect the model's gradient [18]. However, this method requires a white-box environment with full access to the model.

3. Task Description

We participated in Task 6 InCredibleAE, as part of the CheckThat! 2024 Shared Task [19]. Task Six aimed to generate high-quality adversarial examples of text across five domains: COVID-19 misinformation (C19), Fact-Checking (FC), Rumour Detection (RD), Propaganda Detection (PR2), and Style Based News Bias (HN). The distinguishing feature of this task was its emphasis on producing attacks with minimal changes to the original text while effectively altering the classification.

To evaluate the quality of the generated attacks, the BODEGA framework [20] was employed. This framework provided an overall score based on the success of the classification change while also considering semantic preservation using BLEURT [21] and character-based similarities calculated using the Levenshtein distance metric [22].

Additionally, participants engaged in a manual annotation stage where adverse sentences were manually annotated based on their semantic similarity to the original sentences. This stage aimed to further assess the semantic preservation of the generated attacks and provide insights into the effectiveness of the methods employed.

4. Methodology

The task of generating adversarial examples can be split into two stages. Stage one involves a method that searches or evaluates text elements (such as words or characters) to be able to provide a target for attack. Stage two attacks those identified elements by either removing, adding, or replacing them. Both stages are important to delivering an effective classifier, and our approaches to each stage are outlined below.

4.1. Search Method

We used an importance-based approach to identify candidates. Our implementation follows a similar methodology to other popular state-of-the-art approaches [12, 13, 16]. Words are ranked based on their importance to the model's classification through the final logit output. There was no pre-processing performed on the text to preserve the original meaning. Masking the words in the sentence is done by iterating through each word and replacing it with the [MASK] token; the new masked sequence is then run through the classifier. We then sort this list of words in descending order based on the ranking. We take the top K words that have a high impact on the final logit output compared to the base classification output. Our implementation is outlined through pseudo-code in Algorithm 1.

Algorithm 1 Word Importance Search Algorithm

Aim: To identify vulnerable words
Input: Text Sequence X, Logit output for original label Y, Victim Classifier $F(\cdot)$, Threshold for important words K
Output: List of important words L_K

Initialise List L
for X_i in X **do**
    $X^* \leftarrow X$
    $M \leftarrow$ Replace X^* with [MASK]
    $X^* \leftarrow$ Replace X_i with M in X^*
    $W_{score} \leftarrow Y - F(X^*)$
    $L \leftarrow$ Add W_{score} to List
end for
$L_{ordered} \leftarrow$ Sort(L) in descending order
List$L_K \leftarrow$ Top K elements of $L_{ordered}$
return L_K

4.2. Attack Method

We investigated two main attack methods: **Homoglyph** and **Lexical replacement**. Both of these methods followed the same overall methodology. A vulnerable word in the sentence is replaced with a corresponding adversarial word, and this new sentence is submitted to the classifier. If this adversarial sentence elicited a change in classification then the goal is met, and the sentence is returned. If not, then each next attacked word would replace the current, and the sentence would run through the classifier again. For each new adversarial word, the distance is measured from the original classification logits to the new classification logits, with the word with the highest distance being stored. If no perturbations of the vulnerable word are successful in changing classification, then the word with the highest gap in classification is permanently added to the sentence, and the next vulnerable word goes through the same process. We found that without the addition of these additional words, the model did not perform as well in terms of successful classification.

4.2.1. Homoglyphs

Homoglyphs are symbols (glyphs) that can represent a letter, word, or character that look very similar to another letter/symbol but hold a different meaning. We can use these Homoglyphs to generate adversarial examples, as the attacked sentence reads the same, but the tokeniser treats them differently, having to split them apart into multiple unknown tokens. We propose two experiments for the way that Homoglyphs are attacked: **HG1** randomly attacking characters and **HG2** attacking the starting and ending characters of a word.

For HG1, we attacked the vulnerable word by randomly selecting characters from the word, ensuring that the combination of characters was only generated once, meaning that a word containing only

homoglyphs could be generated but would only appear once. A maximum of 50 homoglyphs would be created for each word, with less being created if that threshold is met through combinations of characters. The list of attacked words is then sorted ascending by the number of changes made. For generation of homoglyphs, a dictionary was created mapping each character to a range of homoglyphs, but there was not enough variation of homoglyphs to be effective, so we used a Python library[2] for generation of a multitude of homoglyphs. The full attack methodology is illustrated in more detail in Algorithm 2. HG2 followed the same procedures and had a focused attack strategy of attacking only the beginning and end characters of the word and would only attack the first and second characters of each side of the word unless it was 4 characters or less, which would then only attack the first and last.

Algorithm 2 HG1 Algorithm

Aim: Adversarial attack framework to fool neural text classifier
Input: Text Sequence X, label Y, Victim Classifier $F(\cdot)$
Output: Adversarial Text Sequence X^{adv}

$W \leftarrow Importance(X, Y)$ ▷ Search Method (See Algorithm 1)
for W_i in W **do**
  $H \leftarrow$ Random replacement of characters for Homoglyphs
  $H_{ordered} \leftarrow$ Sort(H) in ascending
  Initialise $score, highestScore$
  for H_i in $H_{ordered}$ **do**
    $X^* \leftarrow$ replacing W_i with H_i in X
    if $F(X^*) \neq Y$ **then**
      return $X^{adv} \leftarrow X^*$
    else
      $score \leftarrow Y - F(X^*)$
      if $score > highestScore$ **then**
        $highestScore \leftarrow score$
      end if
    end if
  end for
  if $highestScore > 0$ **then**
    $X \leftarrow$ replace W_i with $highestScoreH_i$
  end if
end for
return None

4.2.2. Lexical replacement

We used large language models (LLM) to generate substitutions for vulnerable words. We implemented a zero-shot prompting [23] approach, as prompts need to be able to work with multiple types of text. For model selection, we tested our prompts on both base Llama8B[3] and base Mistral 7B[4]. The Llama8B model was selected for being one of the most complex open-source LLMs available throughout the experimentation phase. Mistral was selected due to good performance on NLP tasks [24]. Table 1 provides an outline of the prompts which were experimented. Prompt 3 was used as a baseline but, in the end, performed best in terms of computational time to run and overall BODEGA score. One explanation for this could be that the reduced amount of instruction leads to less noise for the model.

[2] https://pypi.org/project/homoglyphs/
[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[4] https://huggingface.co/mistralai/Mistral-7B-v0.3

Table 1
Experimental Prompts used in LLM

Number	Prompt
1	Based on this sentence <sentence> identify the word <word> and replace it with semantically similar words, picking the best matches for the context of the sentence.
2	Replace a <word> in the given <sentence> with another word that conveys the same meaning. Understand the meaning of the sentence. Generate potential synonyms for the word to be replaced. Select the most contextually appropriate synonym.
3	Generate a list of similar words to <word> in the context of this sentence <sentence> without explanation.

Table 2
Success, Semantic and Character scores for experimental models

						Baseline
Domain	Score	HG1	Llama3	Mistral	HG2	TextFooler
RD	Success	0.37	0.37	0.3	0.32	**0.41**
	Character	0.99	**0.79**	0.8	0.97	0.9
	BLEURT	**0.56**	0.37	0.4	0.46	0.46
	BODEGA	0.15	0.12	0.2	0.14	**0.18**
C19	Success	**0.91**	0.88	0.9	0.78	0.74
	Character	0.93	0.88	**0.85**	0.97	0.92
	BLEURT	0.57	0.42	0.53	**0.6**	0.52
	BODEGA	**0.49**	0.42	0.47	0.46	0.35
HN	Success	0.9	0.93	**0.95**	0.79	0.94
	Character	0.99	**0.9**	0.97	0.98	0.95
	BLEURT	**0.61**	0.5	0.53	0.45	0.42
	BODEGA	**0.55**	0.53	0.43	0.45	0.39
PR2	Success	0.47	0.79	**0.92**	0.38	0.75
	Character	0.95	0.84	**0.8**	0.97	0.88
	BLEURT	0.76	0.64	0.61	**0.78**	0.64
	BODEGA	0.33	**0.44**	0.42	0.29	0.42
FC	Success	0.73	0.87	**0.94**	0.7	0.7
	Character	**0.77**	0.88	0.84	0.97	0.93
	BLEURT	**0.77**	0.53	0.56	0.7	0.69
	BODEGA	**0.55**	0.49	0.47	0.51	0.46

5. Results

Table 2 presents the overall BODEGA score of the experimental models across all domains on the BERT victim model on testing data. The model provides details on all three aspects that make up the BODEGA score: success percentage, BLEURT score for semantic preservation, and the Levenshtein distance for the character score. The TextFooler model is included as a baseline mode.

HG1's overall performance was the best-performing model out of all experimental models. It achieved the highest semantic similarity scores in four of the domains (bar PR2). Comparing HG1 and HG2, it shows that the refined HG2 attack technique seemed to have a negative impact on performance, with it only achieving higher than HG1 in semantic preservation scores for the PR2 and C19 domains. Both Homoglyph attack methods have a high Levenshtein distance, which is expected when attacking characters with potentially unknown characters through Homoglyphs.

The Mistral lexical replacement model outperforms the Llama3 model in terms of the success percentage of classification changes, beating the LLama3 model in all but the RD dataset. Both models' overall BODEGA scores are hindered by high Levenshtein distance and, in some cases, a lower semantic score. Both models are able to slightly outperform the HG2 model.

Table 3
BODEGA scores from different attackers on test data

		Model				
Victim Model	Domain	HG1	TextFooler [10]	Genetic [25]	PWWS [26]	BERTattack [12]
BERT	C19	**0.49**	0.32	0.36	0.3	0.42
	RD	0.15	0.16	0.2	0.15	**0.17**
	HN	**0.55**	0.38	0.37	0.36	**0.55**
	PR2	0.33	0.44	**0.5**	0.47	0.43
	FC	**0.55**	0.46	0.52	0.48	0.53
BiLSTM	C19	0.45	0.38	0.42	0.39	**0.46**
	RD	0.28	0.23	**0.31**	0.28	0.28
	HN	0.54	0.4	0.4	0.4	**0.57**
	PR2	0.40	0.51	**0.54**	0.53	0.53
	FC	0.49	0.55	**0.62**	0.57	0.6
RoBerta	C19	**0.46**	0.27	0.38	0.3	0.37
	RD	**0.17**	0.15	**0.17**	0.14	**0.17**
	HN	**0.45**	0.19	0.26	0.3	0.38
	PR2	0.26	0.23	**0.32**	0.28	0.23
	FC	0.52	0.44	0.51	0.47	**0.56**

5.1. Submission Results

To evaluate the final model, we compare our results to other adversarial generation models provided in the BODEGA framework. Table 3 illustrates five of these models, which achieved the highest BODEGA score across the five domains and three victim models on the test dataset.

The HG1's performance mirrors that of the other models, exhibiting similar fluctuations across various domains. Particularly, HN and RD yield lower BODEGA scores compared to domains with shorter sequence lengths, such as C19 and FC. This is evident in the average word length of HN and RD being 605 and 253, respectively, in contrast to C19's 27 and FC's 38 for the testing dataset. It suggests a common challenge among all models when processing longer content forms. One theory behind this is that longer content requires more alterations to affect a successful classification change, consequently diminishing both semantic and character scores as more words/characters are replaced.

HG1 performs well on BERT-based victim models, having the highest BODEGA score in C19, HN, and FC for the BERT victim and C19, RD, and HN for Roberta. However, the model performs very poorly in the BiLSTM model. The model performs poorly in the PR2 data; one theory for this is due to the shorter word length of these sequences (average word count of 21). It could also show that the Homoglyph attack does not work well on propaganda text. Overall, HG1 is able to outperform TextFooler, another character-based attack methodology, only being outperformed in PR2 and FC for the BiLSTM model. Though this could show that Homoglyphs are more effective in character-based attacks, it is more likely to demonstrate that an importance-based search method is more effective for adversarial attacks. Comparing HG1 to BERTattack (the other model using Importance-based search), both models perform similarly but in different domains.

Table 5 (see Appendix A) presents one example of each domain where the attack was successful on the same piece of text across the three victim models from the testing dataset. This helps us understand what the model is doing while also being able to identify other key features, such as discrepancies across the different domains and victim models. Initial themes show that the models require different strengths of attack in terms of the number of words that are attacked. Some words appear to be important for the entire domain, showing up in all victim classes. For example, "breaking" is attacked in all three classes for the RD domain, with "news" being attacked in just BiLSTM and Roberta. This shows that "breaking" is an important word for rumour detection classification, which correlates with what rumour detection is trying to do, as classifiers would classify based on language that draws people in.

Parts of this reflect the overall BODEGA score presented in Table 3. For instance, HG1's FC perfor-

Table 4
Average number of Minutes on each example

	Model				
Domain	HG1	TextFooler [10]	Genetic [25]	PWWS [26]	BERTattack [12]
C19	0.39	0.38	0.46	**0.31**	0.35
RD	**0.14**	0.5	0.63	0.58	0.15
HN	0.38	**0.26**	0.29	0.34	0.35
PR2	0.06	0.07	0.09	**0.01**	**0.01**
FC	0.08	**0.02**	**0.02**	**0.02**	0.03

mance on the BiLSTM model was lower than both the BERT-based models and also lower than the other state-of-the-art models. When inspecting the number of changes required to change classification, BiLSTM (14) requires 12 more than BERT (2) and RoBERTa (2). However, this pattern is not followed through the whole as HG1's best-performing victim model for the HN domain is the BERT model, yet the text shows that it requires the most changes to bring about a change in classification.

Table 4 illustrates HG1's performance in terms of runtime compared to the other state of the art models, running the testing data on the RoBERTa victim model. Overall it shows that HG1 follows the same patterns in performance with the models taking longer to attack some domains compared to others. It shows us that importance based search techniques which are employed in models such as HG1 and BERTattack perform better on RD data compared to the others.

Finally, in the context of the other competing teams that submitted to InCrediblAE; our Homoglyph model performed low in the overall BODEGA score; placing 5th out of the 6 teams. However in the manual evaluation stage; our model scored the second highest for semantic preservation [19]. Overall showing that the models that were better at semantic preservation to humans were less effective in terms of BODEGA score, but the sentences read closer to the original sentence than the high BODEGA scoring models.

6. Conclusion and Future Work

With this contribution, our main attention was on the use of Homoglyphs as a viable technique for successful adversarial attacks, with a specific focus on minimising the amount of semantic preservation lost. We also show that these homoglyph replacement techniques cannot only compete with but exceed state-of-the-art adversarial attacks, showing that a randomised attack of characters can be more effective than a specific character attack in terms of success rates. Future work will consist of integrating more character-based methodologies to identify a more efficient way of replacing characters. Even though our random attack pattern was successful in changing classification, a more refined approach could help lower the character distance score while increasing semantic preservation.

References

[1] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, The spreading of misinformation online, Proceedings of the national academy of Sciences 113 (2016) 554–559.

[2] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, Y. Liu, Combating fake news: A survey on identification and mitigation techniques, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–42.

[3] M. R. Kondamudi, S. R. Sahoo, L. Chouhan, N. Yadav, A comprehensive survey of fake news in social networks: Attributes, features, and detection approaches, Journal of King Saud University - Computer and Information Sciences 35 (2023) 101571. URL: https://www.sciencedirect.com/science/article/pii/S1319157823001258. doi:https://doi.org/10.1016/j.jksuci.2023.101571.

[4] N. Chattopadhyay, A. Goswami, A. Chattopadhyay, Adversarial attacks and dimensionality in text classifiers, 2024. arXiv:2404.02660.

[5] D. Pruthi, B. Dhingra, Z. C. Lipton, Combating adversarial misspellings with robust word recognition, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5582–5591.

[6] N. Papernot, P. McDaniel, X. Wu, S. Jha, A. Swami, Distillation as a defense to adversarial perturbations against deep neural networks, in: 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 582–597. doi:10.1109/SP.2016.41.

[7] X. Liu, S. Dai, G. Fiumara, P. De Meo, An adversarial training method for text classification, Journal of King Saud University - Computer and Information Sciences 35 (2023) 101697. URL: https://www.sciencedirect.com/science/article/pii/S1319157823002513. doi:https://doi.org/10.1016/j.jksuci.2023.101697.

[8] L. Huber, M. A. Kühn, E. Mosca, G. Groh, Detecting word-level adversarial text attacks via SHapley additive exPlanations, in: S. Gella, H. He, B. P. Majumder, B. Can, E. Giunchiglia, S. Cahyawijaya, S. Min, M. Mozes, X. L. Li, I. Augenstein, A. Rogers, K. Cho, E. Grefenstette, L. Rimell, C. Dyer (Eds.), Proceedings of the 7th Workshop on Representation Learning for NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 156–166. URL: https://aclanthology.org/2022.repl4nlp-1.16. doi:10.18653/v1/2022.repl4nlp-1.16.

[9] J. Ebrahimi, A. Rao, D. Lowd, D. Dou, Hotflip: White-box adversarial examples for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 31–36.

[10] D. Jin, Z. Jin, J. T. Zhou, P. Szolovits, Is bert really robust? a strong baseline for natural language attack on text classification and entailment, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8018–8025.

[11] E. Jones, R. Jia, A. Raghunathan, P. Liang, Robust encodings: A framework for combating adversarial typos, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 2752–2765. URL: https://aclanthology.org/2020.acl-main.245. doi:10.18653/v1/2020.acl-main.245.

[12] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, Bert-attack: Adversarial attack against bert using bert, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6193–6202.

[13] S. Garg, G. Ramakrishnan, BAE: BERT-based adversarial examples for text classification, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6174–6181. URL: https://aclanthology.org/2020.emnlp-main.498. doi:10.18653/v1/2020.emnlp-main.498.

[14] M. Iyyer, J. Wieting, K. Gimpel, L. Zettlemoyer, Adversarial example generation with syntactically controlled paraphrase networks, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1875–1885. URL: https://aclanthology.org/N18-1170. doi:10.18653/v1/N18-1170.

[15] Y. Zhang, J. Baldridge, L. He, Paws: Paraphrase adversaries from word scrambling, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 1298–1308.

[16] J. Y. Yoo, Y. Qi, Towards improving adversarial training of nlp models, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 945–956.

[17] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, stat 1050 (2015) 20.

[18] C. Guo, A. Sablayrolles, H. Jégou, D. Kiela, Gradient-based adversarial attacks against text

transformers, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5747–5757.
[19] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, X. Song, R. Suwaileh, The clef-2024 checkthat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2024, pp. 449–458.
[20] P. Przybyła, A. Shvets, H. Saggion, Verifying the robustness of automatic credibility assessment, 2023. arXiv:2303.08032.
[21] T. Sellam, D. Das, A. Parikh, Bleurt: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 7881–7892.
[22] V. I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, volume 10, Soviet Union, 1966, pp. 707–710.
[23] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, arXiv e-prints (2022) arXiv–2205.
[24] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Singh Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv e-prints (2023) arXiv–2310.
[25] K. W. Church, V. Kordoni, Emerging trends: Sota-chasing, Natural Language Engineering 28 (2022) 249–269. doi:10.1017/S1351324922000043.
[26] S. Ren, Y. Deng, K. He, W. Che, Generating natural language adversarial examples through probability weighted word saliency, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1085–1097. URL: https://aclanthology.org/P19-1103. doi:10.18653/v1/P19-1103.

A. Additional Tables

Table 5
Homoglyph Attacked Sequence for each Domain and Victim Model

Model	Datatype	Homoglyph Attacked sequence
BERT	RD	~~Breaking~~ Breaking news: at lease five people ~~taken~~ taken ~~hostage~~ hostage at a ~~grocery~~ grocery ~~store~~ store in ~~eastern~~ eastern #paris.@shropshirestar ~~#Kosher~~ #kosher ~~grocery~~ grocery store
BiLSTM	RD	~~Breaking~~ breaking ~~news~~ news" at lease ~~five~~ five people taken ~~hostage~~ hostage at a grocery store in eastern #paris.@shropshirestar #kosher grocery store
RoBERTa	RD	~~Breaking~~ break,ng ~~news~~ news: at lease five people ~~taken~~ taken hostage at a grocery store in eastern #paris.@shropshirestar #kosher grocery store
BERT	FC	hannah and her sisters. hannah and her sisters is a 1986 american comedy-drama film which tells the intertwined stories of an extended family over two years that begins and ends with a family thanksgiving dinner. ~ hannah and her sisters is an american ~~1986~~ 1986 film.
BiLSTM	FC	~~Hannah~~ hannah and her sisters.~~Hannah~~ hannah and her ~~sisters~~ sisters is a ~~1986~~ 1986 american ~~comedy-drama~~ comedy-drama film which tells the ~~intertwined~~ intertwined stories of an extended ~~family~~ family over ~~two~~ two years that ~~begins~~ begins and ~~ends~~ ends with a ~~family~~ family thanksgiving dinner. ~ ~~hannah~~ hannah and her ~~sisters~~ sisters is an ~~American~~ american 1986 film
RoBERTa	FC	hannah and her sisters. hannah and her sisters is a ~~1986~~1986 american comedy-drama ~~film~~ film which tells the intertwined stories of an extended family over two years that begins and ends with a family thanksgiving dinner. ~ hannah and her sisters is an american 1986 film.
BERT	PR2	what ~~scandalizes~~ scandalizes believers, and even some fellow bishops, is not only your having appointed such men to be shepherds of the church, but that you also seem silent in the face of their teaching and pastoral ~~practice~~ practice.
BiLSTM	PR2	what ~~scandalizes~~ scandalizes believers, and even some fellow bishops, is not only your having appointed such men to be shepherds of the church, but that you also seem silent in the face of their teaching and pastoral ~~practice~~ practice
RoBERTa	PR2	what ~~scandalizes~~ scandalizes believers, and even some fellow bishops, is not only your having appointed such men to be shepherds of the church, but that you also seem silent in the face of their teaching and pastoral practice….
BERT	C19	~~Take~~ take these ~~steps~~ steps to help prevent the #coronavirus from infecting your #cannabis #stocks https://t.co/vucw2ct96r ~ cannabis may stop coronavirus from infecting people, study finds.
BiLSTM	C19	take these steps to help prevent the ~~#coronavirus~~ #coronavirus from ~~infecting~~ infecting your #cannabis ~~#stocks~~ #stocks ~~https://t.co/vucw2ct96r~~ https://t.co/vucw2ct96r ~ “cannabis may stop coronavirus from infecting people, study finds.”
RoBERTa	C19	take these steps to ~~help~~ help prevent the #coronavirus from ~~infecting~~ infecting your #cannabis #stocks https://t.co/vucw2ct96r ~ “cannabis may stop coronavirus from infecting people, study finds.”
BERT	HN	~~Planned~~ planned parenthood to ~~target~~ target 4 vulnerable gop senatorsplanned parenthood ~~will~~ will ~~run~~ run ~~ads~~ ads targeting four incumbent ~~republican~~ republican ~~senators~~ senators in tight races. yes, the organization that cuts through the faces of babies with still beating hearts to ~~extract~~ extract the ~~brain~~ brain has the gall to attack ~~vulnerable~~ vulnerable republican ~~senators~~ senators while asking people to stand up for planned parenthood healthcare. from the hill: the ads will run in the home states of sens. ~~Kelly~~ kelly ayotte (r-n.h.), rob ~~Portman~~ portman (r-ohio), ron johnson (r-wis.) and pat ~~toomey~~ toomey (r-pa.), all of whom ~~face~~ face tough ~~reelection~~ reelection races next year. those ~~four~~ four races are among those that ~~democrats~~ democrats will target as well in an effort to regain control of the ~~senate~~ senate. a comment on from a reader on the ~~article~~ article from the hill ~~asks~~ asks a ~~great~~ great question: does planned ~~parenthood~~ parenthood need our tax dollars if they can afford to throw money into senate races?
BiLSTM	HN	planned parenthood to target 4 vulnerable gop ~~Senators\nPlanned~~ senatorsplanned parenthood ~~will~~ will run ads targeting four incumbent republican senators in tight races. yes, the organization that cuts through the faces of babies with still beating hearts to extract the brain has the gall to attack vulnerable republican senators while asking people to stand up for planned parenthood healthcare. from the hill: the ads will run in the home states of sens. kelly ayotte (r-n.h.), rob portman (r-ohio), ron johnson (r-wis.) and pat toomey (r-pa.), all of whom face tough reelection races next year. those four races are among those that democrats will target as well in an effort to regain control of the senate. a comment on from a reader on the article from the hill asks a great question: does planned parenthood need our tax dollars if they can afford to throw money into senate races?
RoBERTa	HN	planned parenthood to ~~Target~~ target 4 vulnerable gop ~~Senators\nPlanned~~ senatorsplanned parenthood will run ads ~~targeting~~ targeting four incumbent republican senators in tight races. ~~\nYes~~ yes, the organization that cuts through the faces of babies with still beating hearts to extract the brain has the gall to attack vulnerable republican senators while asking people to stand up for “planned parenthood ~~healthcare~~ healthcare.” from the ~~hill~~hill: the ads will run in the home states of sens. kelly ayotte (r-n.h.), rob portman (r-ohio), ron johnson (r-wis.) and pat toomey (r-pa.), all of whom face tough reelection races next year. those four races are among those that democrats will target as well in an effort to regain control of the senate. a ~~comment~~ comment on from a reader on the article from the hill asks a great question: does planned parenthood need our tax dollars if they can afford to throw money into senate races?