Palöri at CheckThat! 2024 Shared Task 6: GloTa - Combining GloVe Embeddings with RoBERTa for Adversarial Attack*

Notebook for the CheckThat! Lab at CLEF 2024

Haokun He[1,†], Yafeng Song[1,*,†] and Dylan Massey[1,†]

[1] University of Zurich

Abstract

This paper describes the submission of attack methods and results for shared task 6 at CheckThat! Lab at CLEF 2024. We present two novel attack methods to test the robustness of credibility assessment (CA) classifiers across five tasks: fact-checking, COVID-19 misinformation detection, propaganda detection, style-based news bias assessment, and rumor detection. The methods were evaluated using the BODEGA score, which accounts for the success of the attack while preserving the original text's meaning. Our GloTa method, combining GloVe embeddings with RoBERTa-based substitutions, demonstrated superior effectiveness in most tasks compared to baselines. Notably, GloTa achieved the highest BODEGA scores in propaganda detection and fact-checking, indicating significant vulnerability in these areas. However, the method showed comparable performance to baselines in style-based news bias and rumor detection, reflecting the inherent robustness of classifiers in these tasks. Against a more robust pre-trained RoBERTa classifier, GloTa still outperformed RoBERTa-ATTACK, although with generally lower success rates. These findings highlight the need for continuous improvement in adversarial attack techniques to enhance the robustness of CA systems against evolving threats.

Keywords

robustness, adversarial attack, BODEGA score, GloVe embeddings, credibility assessment, RoBERTa

1. Introduction

Credibility assessment (CA) can be understood as a family of tasks that have their goal in determining whether a given textual document adheres to constraints, such as factuality, or not [1]. Advances in NLP techniques and increased availability of high-quality domain-specific data have made classifiers for CA viable for real-world deployment in contexts such as *automated moderation* of comments in online platforms.

However, recent studies [2, 3] indicate that text classifiers can be easily deceived through simple manipulations. For example, a user might circumvent a misinformation classifier by selectively replacing alphabetic characters with numbers. In the statement *drinking water kills*, this would lead to a perturbation such as *drinking w**4**ter k**1**lls*, which a classifier might not robustly handle, leading to misclassification. Such alterations, both simple and sophisticated, highlight that classifiers still lack the **robustness** needed to withstand attacks from users with potentially malicious intent.

The robustness of classifiers can be systematically assessed by automatically perturbing initially correctly classified input examples until the classifier's decision is altered. From an attacker's perspective, the goal is to develop an algorithm that generates adversarial examples for each text sequence, resulting in an opposite label compared to the original text sequence. If a decision can be changed (i.e., the classifier gets *confused*), the attack is considered successful. To ensure that perturbations remain human-readable and convey the original content, semantic and character-based distance metrics can be employed in systematic robustness assessments.

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

*Corresponding author.

† These authors contributed equally.

haokun.he@uzh.ch (H. He); yafeng.song@uzh.ch (Y. Song); dylan.massey@uzh.ch (D. Massey)

In the present working notes, we detail **two attack methods**[1] used to assess the robustness of five tasks given by this shared task: fact checking (FC), COVID-19 misinformation detection (C19), propaganda detection (PR), style-based news bias assessment (HN), and rumor detection (RD). These tasks are interpreted as binary classification tasks aimed at determining whether a given piece of text is credible or not.

We evaluate our attack methods on various classifier models, referred to as *victim models*, which differ in architecture but are applied to the same tasks for comparison. The evaluation metric employed is the BODEGA score, as proposed by Przybyła et al. [1], which measures the success rate of confusion under the constraint of meaning preservation.

2. Background

Credibility assessment (CA) is the high-level task concerned with determining whether some given natural language expression is credible with regards to some aspect, e.g. veracity, or not. Assessing the robustness of classifiers performing CA is vital, for otherwise users with malicious intent might easily bypass such classifiers in contexts such as automated content moderation, such as the screening of contributions to online forums. We restrict our focus to five CA tasks. The first task, FC is concerned with classifying whether a given natural language statement is true or false relative to some body of knowledge [1]. The fact checking classifiers we attack are based on data from Thorne et al. [4]. The other four tasks are similar to FC, but differ by text type and subsequently by the datasets they were trained and evaluated on. These tasks have the goal of *assessing* whether some given text is misinformation [5], a rumor [6], propaganda [7] or fake news [8] respectively. A summary of the statistical information for these tasks is presented in the following Table 1.

Table 1
Detailed information about the task dataset used to attack the victim models.

Task Name	**HN**	**PR**	**FC**	**RD**	**C19**
Domain	News Bias	Propaganda Detection	Fact Checking	Rumor Detection	COVID-19 Misinformation
Number of Texts[2]	400	416	405	415	541
Average Words Per Text	323	21	47	147	43

Previous work on classifier attack strategies can be broadly classified by information available to an attacker (black-box, grey-box, white-box) and perturbation granularity (sentence-level, word-level, character-level). In a white-box setting an attacker has full visibility of the models internals, including model weights [9]. In a black-box scenario – as understood in the context of this task – an attacker only can obtain information of the (binary) classification decision / confidence scores.

We solve for the *grey-box* version of the task as outlined in Przybyła et al. [1]. An attacker hence, (1) can obtain the confidence scores from the model, (2) is provided information about the high-level architecture (not model parameters though), and (3) has access to training and development datasets as also the evaluation method. Further, an attacker is free to query the model as many times as needed in order to confuse it. Przybyła et al. [1] introduce an attack effectiveness metric, called BODEGA score, that is composed of the confusion success rate, while also punishing semantic and lexical (character-level) distance. They detail BODEGA scores for a number of methods on the five aforementioned tasks (FC, C19, PR, HN, RD) subsuming three different model types – BERT [10], BiLSTM and RoBERTa [11]. The datasets used to train the victim models are openly available[3].

[1] The detailed code for our methods is available at: https://github.com/yafengsong/InCrediblAE-2024-GloTa

[2] The numbers of texts here are from the data used for the BERT Classifier. The number of texts may vary slightly for the other two models.

[3] cf.: https://gitlab.com/checkthat_lab/clef2024-checkthat-lab/-/tree/main/task6/incrediblAE_public_release

The measure of evaluation, the BODEGA score, is computed as the product of one binary and two real-valued numbers $S_{\mathrm{BODEGA}} = \mathrm{succ} * \mathrm{sem}_{\mathrm{dist}} * \mathrm{chr}_{\mathrm{dist}}$. Where $\mathrm{succ} \in 0, 1$. The success variable (succ) takes 1 when confusion is achieved and 0 when not. The other two values, semantic distance ($\mathrm{sem}_{\mathrm{dist}}$) and character-edit distance ($\mathrm{chr}_{\mathrm{dist}}$) are $\in [0, 1]$. 1 indicates that similarity is preserved relative to the original in both cases, whereas a value closer to 0 signifies a higher divergence. The BODEGA scores over the individual adversarial generations are then mean-aggregated across test data points and tasks, to generate a final score. Thus, a high BODEGA score against a classifier implies *low robustness*, but a *high fecundity* of the attack method. The BODEGA score can only consider attacks successful that *targeted* towards a potential attacker's goals, i.e. we are only interested in changes from $1 \rightarrow 0$, or consider both confusion directions as a success, which can be understood as *untargeted*. We consider only the *untargeted* scenario.

As basis for our experiments, we follow two promising methods, of which the latter also serves as one of our baselines. Firstly, Li et al. [9], who replace words using nearest neighbor search and second Li et al. [12], who use BERT to detect potential replacements for each input instance. Li et al. [12]'s method, BERT-ATTACK, consists of probing the victim model for words that have high potential to change the classification confidence and then in a subsequent step looking for suitable replacement words for the most vulnerable words that still preserve the meaning. Their method outperforms previous methods and is shown to work relatively well independent of the specific classifier architecture or task (model-agnostic). The second baseline method is from Alzantot et al. [13], who use a genetic algorithm over multiple generations to generate adversarial samples that are maximally fit to confuse the classifier. As a framework for evaluation, we rely on the OpenAttack toolkit developed by Zeng et al. [14]. While our work focuses on the word-level, some approaches have addressed perturbations on a more coarse- respectively fine-grained level [15], such as character-switching [16] or paraphrasing [17].

Our contributions in the *CLEF CheckThat! 2024* edition [18, 19] of the Shared Task on *Robustness of Credibility Assessment with Adversarial Examples (IncrediblAE)* [20] can be summarized as follows:

- We introduce two novel methods to efficiently generate adversarial text samples for robustness assessment of CA classifiers.
- We outperform previous baselines in the majority of CA tasks, with our GloTa approach appearing as the most promising.

3. Methods

Motivated by previous work, we initially attempted to address this task using either rule-based algorithms or neural network-based methods. However, our experiments indicated that rule-based algorithms, such as randomly arranging characters or replacing words using a preset synonym list, did not achieve satisfactory performance across the five test tasks. Consequently, we focused primarily on developing a new model-based method to solve this shared task.

3.1. Contextual Embedding with RoBERTa Attacker

Inspired by BERTAttacker[4] [12], which provides a framework for automatically generating adversarial samples, we first adopted a similar approach. BERTAttacker calculates an importance score for each word and generates a candidate word list for substitution. However, we opted to use the RoBERTa model [11] instead of BERT, given RoBERTa's focus on masked word prediction with dynamic masking and its training on a larger dataset, which should enhance its semantic understanding. We use the RoBERTa-base model to generate importance scores for each word by calculating the difference in output probability distributions between the original input sequence and the masked input sequence.

Once the importance scores for each word are obtained, we rank the words in the sequence based on these scores. This ranking identifies the most vulnerable words, with the highest-ranking word

[4] An implementation can be found at: https://github.com/thunlp/OpenAttack/blob/master/OpenAttack/attackers/bert_attack/__init__.py

being the most susceptible to an attack that could alter the classifier's output. Then, we iteratively substitute these words to execute the attack on the victim model. For each substitution, we identify the word's position and extract its contextual embedding from the second-to-last layer of the RoBERTa model. We select k (in our case k=36) other words from the masked sequence's predictions at that position and extract their contextual embeddings after substituting the original word in the RoBERTa model. We experimented with different values of k and determined that 36 offers the optimal balance between attack success rate and semantic preservation. By comparing the original word's contextual embedding with those of the k selected words, we retain only those with a similarity score above a preset threshold(=0.3) as candidates for substitution.

After obtaining a list of candidate words for each position in the original sequence, we substitute each original word with candidates and check if the substitution fools the victim model. If successful, we stop and return the modified sequence. If none of the candidates succeed, we retain the word that most reduces the confidence in the original label and repeat the process for the next position. This continues until either the attack is successful or all words in the input sequence have been processed. The whole process of this method is shown in Algorithm 1.

Algorithm 1 Adversarial Attack using RoBERTa

```
Require: Original sequence $\mathbf{X}$, victim model $M$, RoBERTa model $R$, number of candidates $k$, similarity threshold $\tau$
Ensure: Modified sequence $\mathbf{X}'$
$\mathbf{S} \leftarrow$ CalculateImportanceScores$(\mathbf{X}, R)$
$\mathbf{W} \leftarrow$ RankWordsByImportance$(\mathbf{S})$
for word $w_i$ in $\mathbf{W}$ do
    $e_i \leftarrow$ ExtractEmbedding$(w_i, R)$
    $\mathbf{C} \leftarrow$ GenerateCandidateWords$(w_i, k, R)$
    $\mathbf{E} \leftarrow$ ExtractEmbeddings$(\mathbf{C}, R)$
    $\mathbf{C}' \leftarrow \{c \in \mathbf{C} \mid \text{Similarity}(e_i, e_c) > \tau\}$
    for candidate $c$ in $\mathbf{C}'$ do
        $\mathbf{X}' \leftarrow$ SubstituteWord$(\mathbf{X}, w_i, c)$
        if $M(\mathbf{X}') \neq M(\mathbf{X})$ then
            return $\mathbf{X}'$
        end if
    end for
    $w_i' \leftarrow$ WordThatReducesConfidenceMost$(\mathbf{C}', M)$
    $\mathbf{X} \leftarrow$ SubstituteWord$(\mathbf{X}, w_i, w_i')$
end for
return $\mathbf{X}$
```

3.2. GloTa: Combining GloVe Embeddings with RoBERTa

GloTa, which stands for **Glo**Ve and RoBer**Ta**, represents a method that combines GloVe [21] embeddings and RoBERTa to enhance adversarial attack techniques. Applying the aforementioned method yielded a high success rate, but the semantic score was often low due to extensive substitution by RoBERTa-generated candidates. These substitutions do not necessarily preserve the original meaning and may even introduce opposite meanings. For example, RoBERTa-generated candidates for the word *love* in the sentence *I love you* might include *miss, forgive,* or *hate.* To address this issue, we use GloVe embeddings to generate candidate lists for substituting vulnerable words in the input sequence.

The candidate lists are generated using a process akin to the initial step in Genetic algorithm [13]. We build a large synonym dataset by computing the N nearest neighbors of each selected word based on distance in the GloVe embedding space (*Common Crawl, 840B tokens, 2.2M vocab*), using the *aclImdb* dataset [22] of movie reviews from IMDB as dictionary to construct the synonym dictionary, thereby mitigating the semantic loss associated with candidates generated from masked language models. We still use the *aclImdb* dataset because it was employed in the original Genetic algorithm paper, allowing us to maintain consistency and compare our results with the original findings.

Algorithm 2 Adversarial Attack using RoBERTa and GloVe

```
Require: Original sequence $\mathbf{X}$, victim model $M$, RoBERTa model $R$, GloVe embeddings $G$, synonym
    dictionary $D$, number of candidates $k$, similarity threshold $\tau$
Ensure: Modified sequence $\mathbf{X}'$
    $\mathbf{S} \leftarrow$ CalculateImportanceScores($\mathbf{X}, R$)
    $\mathbf{W} \leftarrow$ RankWordsByImportance($\mathbf{S}$)
    for word $w_i$ in $\mathbf{W}$ do
        if $w_i \in D$ then
            $\mathbf{C} \leftarrow D[w_i]$
        else
            $e_i \leftarrow$ ExtractEmbedding($w_i, R$)
            $\mathbf{C} \leftarrow$ GenerateCandidateWords($w_i, k, R$)
            $\mathbf{E} \leftarrow$ ExtractEmbeddings($\mathbf{C}, R$)
            $\mathbf{C}' \leftarrow \{c \in \mathbf{C} \mid \text{Similarity}(e_i, e_c) > \tau\}$
            $\mathbf{C} \leftarrow$ ReRankCandidatesByBLEURT($\mathbf{C}', \mathbf{X}$)
        end if
        for candidate $c$ in $\mathbf{C}$ do
            $\mathbf{X}' \leftarrow$ SubstituteWord($\mathbf{X}, w_i, c$)
            if $M(\mathbf{X}') \neq M(\mathbf{X})$ then
                return $\mathbf{X}'$
            end if
        end for
        $w_i' \leftarrow$ WordThatReducesConfidenceMost($\mathbf{C}, M$)
        $\mathbf{X} \leftarrow$ SubstituteWord($\mathbf{X}, w_i, w_i'$)
    end for
    return $\mathbf{X}$
```

Once the synonym dictionary is constructed, we extract the most vulnerable word by the method in Algorithm 1, generate a list of the closest words in the GloVe embeddings as candidates, and then substitute the original word with candidates until the decision is flipped. If none of the words in the list achieves this, we proceed to the next vulnerable word in the sequence and repeat the process. However, because the input sequences span five different domains and may even include URLs and emojis, many words are absent from the constructed synonym dictionary. For these out-of-vocabulary words, we revert to using RoBERTa to generate candidate lists and then re-rank the substitution words by BLEURT [23] to prioritize semantically close words for substitution. The entire process is illustrated in Algorithm 2.

Additionally, we set two types of hyperparameters that can be tuned for different input datasets: (1) *max_candidates* (2) *max_substitutes* and *max_sub_rate*. The first hyperparameter is the number of substitution words. A longer list may increase the success rate but reduce semantic integrity. The second type includes two hyperparameters: the number of substitutions made to the original sequence and the substitution rate compared to the original input sentences. Ideally, we can substitute each word, but excessive substitutions may significantly alter the semantic meaning. Therefore, we establish these thresholds to balance the trade-off between the semantic score and success score. The attack on the sequence will terminate when either threshold is reached. In our experiment, we tested different parameter values and set the optimal parameters as follows: *max_candidates* to 30, *max_substitutes* to 80, and *max_sub_rate* to 0.5.

4. Results

We conducted the RoBERTa-ATTACK and our GloTa method, along with BERT-ATTACK and Genetic algorithms as baselines, on five tasks to attack three victim models: the BERT Classifier, the Bi-LSTM Classifier, and the RoBERTa Classifier. The last model, the RoBERTa Classifier, was introduced as a "surprise" model for this shared task. The results are summarized in Table 2. GloTa achieved the highest BODEGA scores in the Propaganda Detection (PR) and Fact Checking (FC) tasks. Analyzing the sub-scores of these tasks, the gains were primarily from the success scores compared to the baseline methods, indicating that our method is more effective at fooling the classifier in these tasks.

Table 2
Performance comparison of different attack methods on BERT and Bi-LSTM classifiers. Evaluation measures include BODEGA score (BO), success score (suc), semantic score (sem), and character score (char). The best score in each task and scenario is in boldface.

Task	Method	BERT Classifier				Bi-LSTM Classifier				RoBERTa Classifier			
		BO	suc	sem	cha	BO	suc	sem	cha	BO	suc	sem	cha
HN	BERT-ATTACK	**0.60**	0.96	0.64	0.97	**0.64**	0.98	0.66	0.99	-	-	-	-
	Genetic	0.40	0.86	0.47	0.98	0.44	0.94	0.48	0.98	-	-	-	-
	RoBERTa-ATTACK	0.58	0.95	0.62	0.98	0.63	0.99	0.64	0.99	0.34	0.57	0.60	0.98
	GloTa	0.59	0.96	0.62	0.98	0.63	0.99	0.64	0.99	**0.44**	0.77	0.59	0.97
PR	BERT-ATTACK	0.43	0.70	0.68	0.90	0.53	0.80	0.72	0.91	-	-	-	-
	Genetic	0.50	0.84	0.65	0.89	0.54	0.88	0.67	0.89	-	-	-	-
	RoBERTa-ATTACK	0.53	0.95	0.62	0.88	0.58	0.97	0.65	0.89	0.25	0.54	0.54	0.82
	GloTa	**0.56**	0.97	0.64	0.88	**0.60**	0.98	0.68	0.90	**0.45**	0.95	0.56	0.81
FC	BERT-ATTACK	0.53	0.77	0.73	0.95	0.60	0.86	0.73	0.95	-	-	-	-
	Genetic	0.52	0.79	0.70	0.95	0.61	0.90	0.71	0.95	-	-	-	-
	RoBERTa-ATTACK	**0.62**	1.00	0.64	0.96	0.67	0.99	0.70	0.97	**0.67**	0.99	0.69	0.97
	GloTa	**0.62**	0.98	0.66	0.96	**0.69**	1.00	0.71	0.97	**0.67**	0.99	0.70	0.96
RD	BERT-ATTACK	0.18	0.44	0.43	0.96	0.29	0.79	0.41	0.89	-	-	-	-
	Genetic	**0.20**	0.46	0.45	0.96	**0.32**	0.71	0.47	0.96	-	-	-	-
	RoBERTa-ATTACK	**0.20**	0.52	0.42	0.89	0.31	0.76	0.44	0.92	0.20	0.47	0.44	0.93
	GloTa	0.19	0.45	0.44	0.94	0.31	0.71	0.46	0.95	**0.21**	0.52	0.44	0.92
C19	RoBERTa-ATTACK	**0.53**	0.99	0.57	0.92	**0.53**	1.00	0.57	0.92	**0.47**	0.99	0.51	0.89
	GloTa	0.52	0.96	0.58	0.93	**0.53**	1.00	0.57	0.92	0.45	0.96	0.51	0.89

For the Style-based News Bias Assessment (HN) and Rumor Detection (RD) tasks, our GloTa method achieved similar BODEGA scores as the baseline. In the HN task, the baseline BERT-ATTACK already achieved a high success score, limiting the potential for additional gains. In the meantime, GloTa's semantic score did not outperform the baseline, resulting in a comparable BODEGA score. In the RD task, the victim model's robustness led to low success scores across all methods, limiting GloTa's effectiveness in this context.

The COVID-19 Misinformation Detection (C19) task is a new dataset released for this shared task, lacking prior baselines for comparison. Both GloTa and RoBERTa-ATTACK achieved a relatively high success score, indicating this task's susceptibility. However, the semantic score was not high due to the presence of non-word tokens such as URLs, hashtags, and emojis from Twitter data, making it challenging to find semantically similar alternative words.

The newly introduced classifier in this shared task, the RoBERTa classifier, is purported to be more robust. We employed both RoBERTa-ATTACK and GloTa methods to attack this model, although we did not have baseline results for comparison. GloTa significantly outperformed RoBERTa-ATTACK in the HN and PR tasks, while achieving comparable results in the FC and RD tasks. These outcomes from the RoBERTa classifier align with the comparative performance between GloTa and RoBERTa-ATTACK observed in the BERT and Bi-LSTM classifier results. However, the performance gap between them has widened, suggesting that GloTa is more effective at attacking more robust classifiers. Nonetheless, RoBERTa-ATTACK marginally outperformed GloTa in the C19 task, primarily due to its higher success score. Furthermore, when compared to all the BODEGA scores in BERT and Bi-LSTM classifiers, only the FC task exhibited a higher BODEGA score in the RoBERTa classifier, while the other tasks showed lower scores due to decreased success score or semantic score. This indicates that the RoBERTa classifier is indeed more robust.

5. Discussion and Future Work

Given that the BODEGA score consists of three components—success score, semantic score, and character score—the evaluation of adversarial attacks can be analyzed within these divisions. As shown in Table 2, the character scores consistently remain high across different tasks. Therefore, our discussion is focused on improving the success score and semantic score.

5.1. Success Score

Our GloTa method achieves near-perfect success scores across all three classifiers, with the exception of the RD task in these three classifiers and the HN task in the RoBERTa classifier. We attribute this high success rate primarily to our "greedy" attack method, which identifies vulnerable words and replaces them sequentially until the classification is altered. While it is possible to replace all words in the input sentences, this approach would significantly slow the process and lower the semantic score, given that GloTa creates word-level replacements without considering the context. Therefore, we introduced hyperparameters to control the number of substitutions in the original sequences, balancing the trade-off between success and semantic scores. Additionally, as shown in Table 1, we observed that HN and RD tasks involve longer texts compared to other tasks. This can lead to a more robust trained victim model, which requires either replacing more words to alter the result or failing to attack when the thresholds set by hyperparameters are reached. Future research could investigate allowing varying numbers of substitutions in low-success tasks, such as RD, to determine if increasing the success rate can enhance the overall BODEGA score.

5.2. Semantic Score

Our initial goal of combining BERT-ATTACK and Genetic was due to their superior performance in the experiment conducted by Przybyła et al. [1]. We found that using a masked language model to replace words could not maintain a relatively high semantic score, and character-level replacement resulted in a low success rate. Therefore, we employed GloVe embeddings at word level to improve the semantic score in the BODEGA calculation while maintaining a high success rate. However, results indicate that the semantic score of our GloTa method showed only minor improvement compared to using words from masked language models. For some tasks, such as FC, the semantic score decreased despite an increase in the success score, due to more words being replaced in the original sequences. This may be because the synonym dictionary was constructed from the aclImdb dataset, which differs from the domains of the test tasks (HN, PR, FC, RD, and C19). As a result, many out-of-vocabulary words were found in the original sequences, which required using RoBERTa to generate alternative words. This likely explains the similarity in semantic scores between RoBERTa-ATTACK and GloTa.

In addition, our system achieved an average BODEGA score of 0.4776 across all tasks and classifiers with an average semantic score of 0.5867, ranking 4th out of 6 teams. However, in human evaluations, our system managed to preserve the meaning in only 14% of the attack samples, also ranking 4th out of 6 teams. The details of the ranking and human evaluation methodology are explained in the overview paper of this shared task by Barrón-Cedeño et al. [18]. The human evaluation result is notably lower compared to the semantic score achieved by our system. This discrepancy is likely due to several factors. First, in the automatic evaluation, the BODEGA score utilizes BLEURT to calculate the semantic score, where substituting a single word with its close synonym often results in a high score, whereas human evaluation considers the entire context. Second, as previously mentioned, the synonym dictionary did not adequately cover the domain of the test tasks, leading to candidate words that were not semantically similar enough to the original words. Third, word-level replacements do not consider the full context, leading to decreased sentence fluency and greater deviation from the original meanings as more replacements are performed. However, we found that using GloVe embeddings performed better in human evaluation compared to other teams that used only masked language models for word substitution. This is likely because, as noted in Section 3.2, the words generated by the masked language model do not necessarily preserve the original meaning and may even introduce opposite meanings.

To further improve semantic scores, constructing the synonym dictionary using text data from the five test tasks could reduce the occurrence of out-of-vocabulary words, thereby enhancing the semantic score. Additionally, methods such as DeepWordBug [24], which performs character-level modifications, could be explored to enhance the semantic score in both automatic and human evaluations after identifying vulnerable words by the masked language model. Furthermore, our current methodology overlooks the impact of emojis, which were ignored despite their importance in conveying emotional information. Future research should incorporate emoji embeddings to enhance semantic understanding and model performance.

6. Conclusion

In this shared task, we explored the robustness of classifiers for credibility assessment (CA) tasks by developing and evaluating two attack methods: RoBERTa-ATTACK and GloTa. The GloTa method, which combines GloVe embeddings and RoBERTa to enhance adversarial attack capabilities, demonstrated superior effectiveness, achieving the highest BODEGA scores in propaganda detection (PR) and fact-checking (FC) tasks. This indicates a significant vulnerability in these areas. However, its performance was on par with baselines in style-based news bias assessment (HN) and rumor detection (RD), reflecting the inherent robustness of classifiers in these tasks.

When tested against a more robust RoBERTa classifier, GloTa outperformed RoBERTa-ATTACK, although with generally lower success rates, underscoring the enhanced robustness of the RoBERTa classifier. These findings highlight the trade-off between achieving high attack success rates and maintaining semantic integrity. Our study enhances the understanding of CA classifier robustness and demonstrates that using a masked language model to identify vulnerable words and replace them with similar word embeddings in the original texts can be an effective method for adversarial attacks.

Acknowledgments

Special thanks to all the organizers of the CheckThat! Lab at CLEF 2024. We would also like to extend our gratitude to Andrianos Michail, Simon Clematide and the Department of Computational Linguistics at the University of Zurich for valuable suggestions and continuous support for this shared task.

References

[1] P. Przybyła, A. V. Shvets, H. Saggion, Verifying the robustness of automatic credibility assessment, 2023. URL: https://api.semanticscholar.org/CorpusID:257505431.

[2] B. Liang, H. Li, M. Su, P. Bian, X. Li, W. Shi, Deep text classification can be fooled, arXiv preprint arXiv:1704.08006 (2017). URL: https://arxiv.org/abs/1704.08006.

[3] Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu, F. Li, A survey on adversarial attack in the age of artificial intelligence, Wireless Communications and Mobile Computing 2021 (2021) 1–22. doi:10.1155/2021/4907754.

[4] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal (Eds.), Proceedings of the First Workshop on Fact Extraction and VERification

(FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1–9. URL: https://aclanthology.org/W18-5501. doi:10.18653/v1/W18-5501.

[5] Y. Jiang, X. Song, C. Scarton, I. Singh, A. Aker, K. Bontcheva, Categorising fine-to-coarse grained misinformation: An empirical study of the COVID-19 infodemic, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 556–567. URL: https://aclanthology.org/2023.ranlp-1.61.

[6] S. Han, J. Gao, F. Ciravegna, Neural language model based training data augmentation for weakly supervised early rumor detection, in: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ACM, Vancouver British Columbia Canada, 2019, pp. 105–112. URL: https://dl.acm.org/doi/10.1145/3341161.3342892. doi:10.1145/3341161.3342892.

[7] G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, P. Nakov, SemEval-2020 task 11: Detection of propaganda techniques in news articles, in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 1377–1414. URL: https://aclanthology.org/2020.semeval-1.186. doi:10.18653/v1/2020.semeval-1.186.

[8] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 231–240. URL: https://aclanthology.org/P18-1022. doi:10.18653/v1/P18-1022.

[9] J. Li, S. Ji, T. Du, B. Li, T. Wang, Textbugger: Generating adversarial text against real-world applications, ArXiv abs/1812.05271 (2018). URL: https://api.semanticscholar.org/CorpusID:54815878.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[11] 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 abs/1907.11692 (2019). URL: https://api.semanticscholar.org/CorpusID:198953378.

[12] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, BERT-ATTACK: Adversarial attack against BERT using BERT, ArXiv abs/2004.09984 (2020). URL: https://api.semanticscholar.org/CorpusID:216036179.

[13] M. F. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. B. Srivastava, K.-W. Chang, Generating natural language adversarial examples, ArXiv abs/1804.07998 (2018). URL: https://api.semanticscholar.org/CorpusID:5076191.

[14] G. Zeng, F. Qi, Q. Zhou, T. Zhang, Z. Ma, B. Hou, Y. Zang, Z. Liu, M. Sun, OpenAttack: An open-source textual adversarial attack toolkit, in: H. Ji, J. C. Park, R. Xia (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2021, pp. 363–371. URL: https://aclanthology.org/2021.acl-demo.43. doi:10.18653/v1/2021.acl-demo.43.

[15] X. Han, Y. Zhang, W. Wang, B. Wang, Text Adversarial Attacks and Defenses: Issues, Taxonomy, and Perspectives, Security and Communication Networks 2022 (2022) 6458488. URL: https://doi.org/10.1155/2022/6458488. doi:10.1155/2022/6458488, publisher: Hindawi.

[16] J. Gao, J. Lanchantin, M. L. Soffa, Y. Qi, Black-box generation of adversarial text sequences to evade deep learning classifiers, in: 2018 IEEE Security and Privacy Workshops (SPW), 2018, pp. 50–56. doi:10.1109/SPW.2018.00016.

[17] 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.

[18] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, C. Li, J. Piskorski, F. Ruggeri, X. Song, R. Suwaileh, Overview of the CLEF-2024 CheckThat! Lab: Check-Worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[19] G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France, 2024.

[20] P. Przybyła, B. Wu, A. Shvets, Y. Mu, K. C. Sheang, X. Song, H. Saggion, Overview of the CLEF-2024 CheckThat! lab task 6 on robustness of credibility assessment with adversarial examples (InCrediblAE), in: [19], 2024.

[21] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 142–150. URL: http://www.aclweb.org/anthology/P11-1015.

[23] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, 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. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704. doi:10.18653/v1/2020.acl-main.704.

[24] J. Gao, J. Lanchantin, M. L. Soffa, Y. Qi, Black-box generation of adversarial text sequences to evade deep learning classifiers, in: 2018 IEEE Security and Privacy Workshops (SPW), IEEE, 2018, pp. 50–56. URL: http://arxiv.org/abs/1801.04354.