Overview of the CLEF-2024 CheckThat! Lab Task 6 on Robustness of Credibility Assessment with Adversarial Examples (InCrediblAE)

Piotr Przybyła[1,2,*], Ben Wu[3], Alexander Shvets[1], Yida Mu[3], Kim Cheng Sheang[1], Xingyi Song[3] and Horacio Saggion[3]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[3]University of Sheffield, Sheffield, UK

Abstract

Task 6 at CheckThat! Lab, organised at CLEF-2024, is devoted to assessing the robustness of misinformation detection solutions implemented as text classification models. The participants of the task were provided with prediction models and data examples for several problems of credibility estimation and their goal was to come up with adversarial examples (AEs): small modifications to the provided text fragments, such that the original meaning is preserved, but the victim classifier changes its decision. The evaluation involved five domains (detection of: biased news, propaganda techniques, false claims, rumours and COVID-19 misinformation) and three classifiers (BiLSTM, BERT and adversarially fine-tuned RoBERTa). Six teams participated in the task, representing a variety of approaches and substantially outperforming previous AE generation solutions. We also performed manual evaluation, which highlighted some modification techniques that are particularly likely to pass unnoticed by human readers. Overall, the task results emphasise the need to assess the robustness of text classification solutions before implementing them in content filtering on large platforms, such as social media.

Keywords

adversarial examples, robustness, misinformation detection, credibility, text classification, natural language processing

1. Introduction

The challenges of misinformation have been taken up with great energy and vigour by the NLP and IR communities. The main reasons for such enthusiastic adoption of the new tasks are wide availability of textual data to train on and tantalisingly simple dichotomy of *fake* vs. *real*, clearly fitting the familiar task of binary classification. Among a great deal of work in the domain [1, 2], this framework has also enabled numerous shared tasks, including detecting hyperpartisan news [3], propaganda [4], bots [5], false claims [6] and more. These research results quickly found applications in content moderation for large media platforms, which increasingly rely on ML tools to support, but also to replace the human effort [7].

However, a shared task framework is far from the real-world application scenario, where the test data are not fixed, but are generated continuously by users. This means that if a malicious actor sees their non-credible content rejected by the system, they are likely to try to modify it to pass the filters, rather than simply abandon their goals. Unfortunately, the deep learning architectures, which many of the best-performing solutions use, are known for their susceptibility to *adversarial examples*, i.e. data instances modified with the intent of fooling a classifier [8]. While discovering adversarial examples for text is more challenging than in other domains, it is definitely possible [9]. Thus, investigating

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
piotr.przybyla@upf.edu (P. Przybyła); bpwu1@sheffield.ac.uk (B. Wu); y.mu@sheffield.ac.uk (Y. Mu); kimcheng.sheang@upf.edu (K. C. Sheang); x.song@sheffield.ac.uk (X. Song); horacio.saggion@upf.edu (H. Saggion)
0000-0001-9043-6817 (P. Przybyła); 0009-0002-0918-526X (B. Wu); 0000-0002-8255-9435 (Y. Mu); 0000-0002-4662-0358 (K. C. Sheang); 0000-0002-4188-6974 (X. Song); 0000-0003-0016-7807 (H. Saggion)

the robustness of text credibility assessment solutions is indispensable for making them applicable in real-world adversarial scenarios.

Here we report on the shared task on **In**vestigating Robustness of **Credibil**ity Assessment with **A**dversarial **E**xamples (**InCrediblAE**), which was organised as Task 6 of the CheckThat! 2024 evaluation lab [10, 11] at the CLEF 2024 conference. In InCrediblAE, participants get access to the following resources for each *domain*:

1. three *victim classifiers*, assessing the credibility of the input text and returning a score,
2. an *attack dataset*, including around 400 instances unseen by the classifier in training.

In the task we take into account five domains, corresponding to important challenges in the credibility assessment – see section 2 for details.

The goal of the participants is to make modification to the instances in the attack dataset, turning them into adversarial examples. Each adversarial example (AE) is evaluated on *meaning preservation*, i.e. how similar it is to the original; and *classifier confusion*, i.e. whether the output of the victim classifier is different than for the original.

The evaluation consists of two stages. The **automatic evaluation** follows the framework established in the field of adversarial learning, with the above factors assessed through automatic measures, i.e. BODEGA score [12], leading to a ranking list – see section 3.1. In the **manual evaluation** we use human judgement to assess the semantic similarity between attack sentences and their original counterparts. This process aims to highlight potential errors arising from automatic evaluation metrics, as well as to create a high-quality text similarity dataset for the future development and evaluation of metrics – see section 3.2.

The task has attracted six teams submitting various approaches (section 4.1), some of which have clearly outperformed previous solutions applied to the same problems (section 4.2), generally confirming the high vulnerability of popular text classifiers to adversarial attacks. What is more, the results of the manual evaluation (section 4.3) highlight cases where AEs might appear far from the original text if judged by automatic measures, but are in fact quite convincing to human annotators. The results of the manual annotation are made openly available for future research[1]. We also share the code[2] and data[3] for the automatic evaluation.

2. Task data

The shared task uses the foundation of the BODEGA framework [12], which has been created to enable systematic robustness testing in the area of misinformation detection. Within this framework, several *domains* are available, each organised around a credibility assessment problem, defined as a binary text classification task. Within a domain, an expert-annotated corpus of documents is used to train victim classifiers (train subset) and test the attack performance (attack subset).

The participants are provided with Python code, hosted on Google Colab, allowing to interact with the victim models in an attack scenario through *OpenAttack* interface [13]. Their goal is to prepare a procedure that modifies the text fragments in the attack dataset to achieve a different decision of a classifier with a minimal meaning alteration. The participant submission includes the AEs, as well as the number of victim model queries needed to find them.

2.1. Datasets

Four of the domains in InCrediblAE have been prepared in BODEGA based on previously published corpora. The final one (C19) is new and was not available before.

Style-based news bias assessment (HN) is a task of verifying credibility of a news article based on its overall writing style. It relies on previous work indicating that stylistic analysis of fake news

[1] https://github.com/GateNLP/CLEF2024_InCrediblAE_Manual_Evaluation_Dataset
[2] https://github.com/piotrmp/BODEGA
[3] https://gitlab.com/checkthat_lab/clef2024-checkthat-lab/-/tree/main/task6

Table 1
The sizes of subsets in each domain and the percentage of cases labelled as positive.

Task	Training	Attack	Development	Positive
HN	60,235	400	3,600	50.00%
PR	12,675	416	3,320	29.42%
FC	172,763	405	19,010	51.27%
RD	8,694	415	2,070	32.68%
C19	1,130	595	0	42.55%

outlets can be used to distinguish them from credible sources [14, 15]. The corpus contains news bias annotations assigned at the level of the source (whole website) by journalists from *BuzzFeed* and *MediaBiasFactCheck.com*[4]. For the purpose of BODEGA, 10% of the training instances were used and the non-credibility label was assigned to articles from sources marked as hyperpartisan, both left- and right-wing.

Propaganda detection (PR) is focused on recognising specific manipulation techniques appealing to emotions [16], for example name-calling, flag-waving or straw-man fallacy. This approach has the advantage of being fine-grained by highlighting manipulative fragments in text, akin to the NER (Named Entity Recognition) tasks. We rely on the token-level annotations of 14 techniques, marked by professional annotators for the SemEval 2020 Task 11 (*Detection of Propaganda Techniques in News Articles*) [4], for which the training set is public[5]. In order to cast the task as binary text classification for BODEGA, the corpus was split into sentences and those including some tokens marked as propaganda were labelled as non-credible.

Fact checking (FC) is an approach to misinformation detection based on extracting claims made in a piece of text and verifying them with respect to a trusted knowledge base [17]. In order to represent the problem as binary classification, we focus on the final stage of the workflow, when a claim is compared to relevant evidence from the knowledge base, which either confirms its validity or refutes it. In BODEGA, the data from the FEVER shared task [18] is used, consisting of claims that were paired with relevant passages from Wikipedia articles. The instances where a claim is supported by the evidence were labelled as credible, and those when it is refuted as non-credible.

Rumour detection (RD) is aimed at detecting information spreading widely over social media despite it not coming from a credible source. Rumours can be detected using many indicators [19], but here we focus on the textual content of a social media post, as well as the reactions of other users. In BODEGA, this is achieved thanks to the *augmented dataset of rumours and non-rumours for rumour detection* [20], created from Twitter threads relevant to six real-world events, labelled by experts according to the source reliability of the initial post. One of the events (Charlie Hebdo shooting) was set aside as the attack dataset.

COVID-19 misinformation detection (C19) focuses on binary classification[6] of misinformation related to COVID-19 [21, 22]. Given a known false claim about the disease, the task is to determine whether a user's tweet supports that false claim. If so, the tweet is classified as COVID-19 misinformation (positive class). Alternative responses, such as contradicting, questioning, commenting, or irrelevance regarding the false claim are reserved for the negative class. Appendix B provides examples from the dataset.

Table 1 summarises the information on datasets, including the sizes of subsets: training (for training victim classifiers), development (reserved for future use) and attack (to be modified into AEs), as well as the percentage of positive (non-credible) instances.

[4]https://zenodo.org/record/1489920
[5]https://zenodo.org/record/3952415
[6]We merged IRRELEVANT and DEBUNK class as non-misinfo to covert original dataset into a binary classification

2.2. Victim classifiers

Training datasets were used to prepare victim models, representing popular approaches to text classification. Two of the models (BiLSTM and BERT) were trained as in BODEGA framework, but the *surprise classifier* was trained specifically for the shared task and it was revealed to the participants in the test phase, one week before the submissions.

BiLSTM classifier consists of an embedding layer (token representations of size 32), two LSTM [23] layers (forwards and backwards, hidden representation of size 128) and a dense linear layer converting the text fragment representation (of size 256) into two-class probability, normalised using softmax.

BERT classifier is a `bert-base-uncased` model [24] from the *HuggingFace Transformers* library [25], fine-tuned for sequence classification using Adam optimiser with linear weight decay [26] for 5 epochs.

Surprise classifier is a RoBERTa model [27], i.e. `roberta-base` from *HuggingFace Transformers*, adversarially-trained to be more robust to adversarial attacks. We use data augmentation to improve robustness: First, the model is fine-tuned for one epoch on the train dataset, then adversarial examples are generated from the entire train dataset using BERT-ATTACK [28], and then the model is fine-tuned for one epoch on a combination of the train dataset and the successful adversarial examples. We train with constant learning rate 2×10^{-5} and the Adam optimiser. We use batch of size 32 for all tasks except PR, which uses a batch of size 64. Due to computational constraints, for HN we only generate adversarial examples from a subset (6000 samples) of the training data.

3. Evaluation

The evaluation procedure consists of two stages. Firstly, the BODEGA framework is used to automatically assess the attack effectiveness of each participant in 15 scenarios (5 domains $\times$ 3 victims). The average BODEGA score is used to create the leaderboard, expressing the overall performance. Secondly, the task most challenging for automatic evaluation (fact-checking) is used to perform manual annotation of meaning preservation in selected instances.

3.1. Automatic evaluation

In automatic evaluation, when an example x_i is modified into AE x_i^*, the quality of the transformation is assessed through BODEGA score defined as follows [12]:

$$\text{BODEGA\_score}(x_i, x_i^*) = \text{Con\_score}(x_i, x_i^*) \times \text{Sem\_score}(x_i, x_i^*) \times \text{Char\_score}(x_i, x_i^*),$$

where:

- Con_score, i.e. confusion score, takes value of 1 when the attacked classifier predicts a different class for x_i^* than it did for x_i, and 0 otherwise.
- Sem_score, i.e. semantic similarity score, is a measure of meaning preservation between x_i and x_i^*, computed using the *BLEURT* [29] evaluation measure (`BLEURT-20` variant), clipped to the (0-1) range.
- Char_score, i.e. character similarity score, is a measure of similarity of x_i and x_i^* as character sequences, computed through Levenshtein distance [30], scaled to (0-1) similarity.

We can see that an AE will be ranked highly if it changes the output of the classifier ($\text{Con\_score}(x_i, x_i^*) = 1$), but at the same time preserves both the meaning ($\text{Sem\_score}(x_i, x_i^*) \approx 1$) and the appearance ($\text{Char\_score}(x_i, x_i^*) \approx 1$) of the original text.

To measure the overall attack success in a particular scenario, the BODEGA score averaged over all instances in the attack set is employed. However, the constituent scores, also averaged over the dataset, can be used to understand the results. We also report the number of queries that is needed (on average) for a single AE to be found.

Table 2
Category definitions for the manual evaluation.

Category	Definitions
Preserve the Semantic Meaning	This label is used when the semantic content of the attack sample closely aligns with that of the original sample. Participants should use this label if the meaning, context, and intent of the compared texts remain essentially unchanged, indicating that the attack sample has effectively maintained the core message of the original.
Change the Semantic Meaning	Participants should apply this label when there is a noticeable alteration in the semantic content between the original and the attack sample. This label indicates that while the attack sample may be related or similar in some aspects to the original, it diverges enough in meaning or intent to be considered distinct or modified. For example, name entities (including Year, Name, Location, etc.) are changed in the modified text.
No sense	The content of the attack sample does not make any sense.

Table 3
Confidence scores and definitions used by Mu et al. [31].

Confidence	Definitions
1	Extremely unconfident about the annotation (I'm really unsure about the annotation. It may belong to another category as well, you may wish to discard this instance from the training.)
2	Not confident about the annotation (I'm not sure about the annotation, it seems it also belongs to other categories, but you can still include this instance as a "silver standard instance" in training.)
3	Pretty confident about the annotation (I'm pretty sure about the annotation, but might be in high chance other annotators may label it in a different category.)
4	Fairly confident about the annotation (I'm confident about the annotation, but might be in small chance other annotators may label it in a different category.)
5	Extremely confident about the annotation (I'm certain about the annotation without a doubt.)

3.2. Manual evaluation

The goal of the manual evaluation is to highlight the cases where the automatic evaluation measures, especially regarding semantic similarity, might not be an accurate representation of human reception of the adversarially modified content. To that end, we have selected the samples from the fact-checking domain and the surprise victim, where even small changes in text can alter the meaning and, consequently, the credibility label.

Task Description We aim to gather assessments regarding the semantic similarities between attack samples and the original samples. Participants in the shared task are requested to dedicate approximately 60 minutes to this manual evaluation (i.e., 100 samples per participant), which are conducted using an open-source, collaborative annotation platform, i.e., GATE Teamware 2 [32]. Judges rate the sample pairs based on the following scale: (a). Preserve the Semantic Meaning, (b). Change the Semantic Meaning, and (c). No sense. Table 2 demonstrates categories and descriptions. Similar to the work of Mu et al. [31], annotators are required to indicate the confidence level (see Table 3) of their assigned class in the 'Confidence Row'.

Data Sampling We randomly select 100 paired samples (i.e., original and modified texts) from each submission, resulting in a total of 600 paired samples. Note that only the successful attacks are considered.

Annotator Training We train the annotators by providing a training document detailing the annotation pipeline, which includes (i) a step-by-step tutorial for using the GATE Teamware platform, (ii) a user information sheet to inform about any potential issues and risks that may occur during data annotation, and (iii) a user consent sheet as required by the ethical approval from the University of Sheffield, where the annotation was performed.

A total of 12 annotators (i.e., 6 participants and 6 organisers) were recruited to manually annotate the paired samples. These 12 annotators were further divided into 6 separate groups (i.e., two annotators per group). In each group, 100 tweets were assigned to each annotator. Finally, this process yielded 100 double-annotated paired samples from each group, resulting in 600 double-annotated samples in total.

Annotation Methodology and Quality Assurance All samples are double-annotated by the shared task organizers and participants. Briefly, each paired sample is annotated by one participant and one shared task organiser. A third annotator from the shared-task organisers is used to resolve any conflicts. Given that there are three categories in total, the annotation with the highest confidence score will be considered in the case of three differing annotations.

4. Results

Here we outline the results of the InCrediblAE shared task in three steps: first, we describe the solutions submitted by the participants (section 4.1), then we present the results of the automatic (section 4.2) and manual evaluation (section 4.3).

4.1. Participating solutions

The SINAI team [33] proposes a method for adversarial attacks based on the substitution of characters by homoglyphs (e.g. characters which resemble the target such as $l \approx 1$). The method uses exhaustive search with two variants: with memory and without memory. They ground their approach in the fact that homoglyphs could deceive the human eye while at the same time provoke a Large Language Model classifier to reverse its prediction due to the presence of an unexpected token. According to the official leader board, the approach is sub-optimal, obtaining the last rank in the task according to the official evaluation metrics. However, human evaluation of content preservation is ranked high.

The MMU_NLP participation [34] features a system to attack classifiers based on lexical substitution and character replacement. The proposed method searches for candidate words to attack followed by a word replacement mechanism. The word search mechanism masks words to check their vulnerability with those words having a high impact on the classifier performance retained for the attack. The replacement step uses homoglyphs for character replacement or lexical substitution. Character replacement is tested in two different conditions: random character attack or begin/end of word attack. The lexical replacement attack uses a large language model to retrieve a word similar to the target word. Overall result fall short compared to other participants, however the proposed methods improves over the baseline in several settings. The character attack method seems more effective than the word replacement approach.

The Palöri team [35] proposes an approach to identify vulnerable words by computing (relying on a masked language model) the difference between the probability distribution of the original sentence and the sentence with a word masked. These differences are used as scores to rank words by their "vulnerability" according to the model. The ranked words (most to least vulnerable) are then replaced using a word from a list of substitutes proposed by the language model. The sentence with the replaced word is used to attack the victim classifier. In case of success, the new sentence is returned, otherwise the method loops using the sentence with the "best" possible substitution (i.e. one which reduced the victim's confidence the most). The method produces successful attacks which however do not preserve the original sentence's meaning. To address this problem a "synonym" dictionary is created, using GloVe embeddings and the aclImbd dataset, to draw substitutes from. The new method only contributes minor improvements over the masked language model.

The solution of the OpenFact team [36] consists in a coupling of various word-substitution approaches in an ensemble in such a way that if the first approach does not succeed in changing the classifier's decision, then the second one is called. In particular, a modification of BERT-ATTACK [28] was proposed (it features a change in parameter values, an alternative selection of a replacement position by an exhaustive search of candidates that provide the largest difference in probabilities for a predicted class, and an iterative replacement from 1 to 7 words, including punctuation and digits at the latter, until the success of an attack) and backed up by a genetic algorithm [37] realised in the OpenAttack framework [13]. Another ensemble was compiled of a proposed greedy search by word swap with synonyms in the word embedding space (prebuilt "counter-fitted" GloVe embeddings [38]) and another model available in the OpenAttack, TextFooler [39], which unlike other similar approaches replaces

words in agreement with the syntax of the attacked text. Apart from ensemble models, approaches from the TextAttack framework [40] were used. They demonstrated superior performance over the baseline methods in automatic scores. In particular, CLARE [41] – a model that implements a special mask-then-infill procedure that incorporates replace/insert/merge operations allowing for outcomes varied in length – was applied for PR, FC and C19 tasks and consistently yielded better results in all automatic scores. Overall, this solution gained the best automatic scores in most of the domains for most of the victims, which made it the first in the leaderboard created by averaging the scores across all scenarios. However, it was ranked very low within the human evaluation, as in the majority of the cases the meaning of the text was changed.

The TurQUaz [42] team leverages a genetic algorithm to look for a combination of character modifications. The modifications that are introduced using a mutation operator include homoglyph replacement, three options of word splitting (random, favouring existing words in the subword outcome, and special heuristic-wise), insertion or removal of individual random letters, and shuffling the order of the letters within the word. The search is carried out until the first flip is found. Only the use of homoglyphs and word splits proved to be efficient. Apart from the genetic algorithm, the team experimented with attacks made by utilizing large language models (LLMs) such as Llama 3[7] and Mistral[8]. Three approaches have been tried: (i) prompting a model for text paraphrasing, (ii) leveraging a model for identifying words to be changed, and (iii) generating adversarial examples with one LLM and verifying whether an attack is going to be successful with another LLM. None of the approaches outperformed the genetic algorithm, however, the team believes in the potential of LLMs and suggests fine-tuning the models specifically for this task to improve their performance in the future. The primary solution has been ranked third based on the BODEGA score.

The TextTrojaners team [43] introduces a BeamAttack method that makes attacks at a word level using RoBERTa [27] and a beam search as a backbone to produce contextually appropriate word substitutions. Beam search algorithm adapted by enabling operations of replacing, skipping, or removing words allows for generating and evaluating multiple alternative word replacement combinations in a single run. For the identification of the most vulnerable words to be replaced, the team experiments with two ranking approaches that use the explainable AI framework LIME [44] and logit-based importance scores, as proposed in [28]. In a series of ablation studies, they show that the choice of a ranking method varies across victims and depends on the dataset. The solution achieved the second-best result on the BODEGA evaluation metric but gained rather low manual evaluation scores.

4.2. Automatic evaluation

Table 4 includes the results of the evaluation against the BiLSTM victim. Generally, we can see that different approaches dominate in different scenarios. However, in every domain, the best BODEGA score (in boldface) is achieved by a solution submitted to InCrediblAE, rather than a reference solution from previous work (BERT-ATTACK or DeepWordBug). HN appears to be the easiest domain, with the leading solution (TextTrojaners) achieving BODEGA score of 0.91 through 100% confusion with 91% semantic similarity and 99% character similarity. This result, closely followed by OpenFact, is especially impressive when compared to the scores of BERT-ATTACK (BODEGA score of 0.64). The TextTrojaners approach also leads in PR, but we need to note its high amount of queries needed – in this case, 593 compared to TurQUaz achieving almost the same result (0.68 instead of 0.70) with six times less queries. The C19 appears to be the domain most challenging for attacks, although even here we note a vast improvement over the reference method (OpenFact: 0.72 vs BERT-ATTACK: 0.50).

In the attacks against the BERT victim, evaluated in table 5, the OpenFact method dominates, ceding only in FC to homoglyph-based SINAI. We can also note that the best BODEGA score is either equivalent (for FC, HN and C19) or significantly lower (for PR and RD) than for BiLSTM, indicating higher difficulty of attacking a Transformer-based classifier. This also results in a higher number of queries necessary to

[7] https://huggingface.co/docs/transformers/model_doc/llama3
[8] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Table 4
Results of the automatic evaluation of attacks against the **BiLSTM** victims, for each of the five domains, expressed through BODEGA score and its constituents, as well as the number of queries (except for OpenFact, whose submission did not include this value).

Domain	Method	BODEGA	Confusion	Semantic	Character	Queries
PR	MMU_NLP	0.40	0.57	0.74	0.95	390.42
	OpenFact	0.65	0.94	0.77	0.89	-
	Palöri	0.60	0.98	0.68	0.90	52.31
	SINAI	0.34	0.37	0.93	0.98	318.27
	TextTrojaners	**0.70**	0.97	0.80	0.90	593.38
	TurQUaz	0.68	0.95	0.76	0.94	96.22
	BERT-ATTACK	0.53	0.80	0.72	0.91	61.41
	DeepWordBug	0.29	0.38	0.79	0.96	27.45
FC	MMU_NLP	0.49	0.66	0.77	0.96	175.32
	OpenFact	0.80	0.98	0.84	0.97	-
	Palöri	0.69	1.00	0.71	0.97	76.47
	SINAI	**0.82**	0.98	0.86	0.96	507.16
	TextTrojaners	0.76	1.00	0.81	0.94	1,549.12
	TurQUaz	0.79	1.00	0.82	0.97	39.44
	BERT-ATTACK	0.60	0.86	0.73	0.95	132.80
	DeepWordBug	0.48	0.58	0.85	0.98	54.36
RD	MMU_NLP	0.12	0.28	0.44	0.97	2,198.42
	OpenFact	**0.84**	0.95	0.91	0.98	-
	Palöri	0.30	0.71	0.44	0.95	532.04
	SINAI	0.14	0.17	0.85	1.00	140.96
	TextTrojaners	0.83	1.00	0.87	0.96	3,831.72
	TurQUaz	0.39	0.67	0.61	0.95	276.93
	BERT-ATTACK	0.29	0.79	0.41	0.89	985.52
	DeepWordBug	0.16	0.24	0.68	0.99	232.75
HN	MMU_NLP	0.54	0.91	0.59	0.99	898.83
	OpenFact	0.89	0.97	0.93	0.99	-
	Palöri	0.64	0.99	0.65	0.99	373.22
	SINAI	0.41	0.48	0.87	1.00	172.92
	TextTrojaners	**0.91**	1.00	0.91	0.99	936.96
	TurQUaz	0.62	1.00	0.65	0.96	68.30
	BERT-ATTACK	0.64	0.98	0.66	0.99	487.85
	DeepWordBug	0.41	0.53	0.77	1.00	396.18
C19	MMU_NLP	0.45	0.78	0.60	0.95	149.15
	OpenFact	**0.72**	0.91	0.83	0.96	-
	Palöri	0.53	0.99	0.57	0.93	110.90
	SINAI	0.26	0.30	0.88	1.00	33.03
	TextTrojaners	0.72	0.99	0.77	0.92	837.35
	TurQUaz	0.58	0.93	0.65	0.96	103.10
	BERT-ATTACK	0.50	0.84	0.62	0.95	127.17
	DeepWordBug	0.33	0.48	0.70	0.99	61.15

find an AE, e.g. while the leading method for HN domain needed 937 queries to obtain the score of 0.91 with BiLSTM, attacking BERT with the same approach requires 4328 queries on average.

The results for the surprise victim (table 6) show that the adversarially-trained classifier indeed is more challenging to attack, leading to lower scores in PR and RD domains. However, the general view remains similar, with OpenFact again dominating with the exception of FC. Their approach clearly works very well with Transformer-based victims, just with an exception of this single task. The increasing level of difficulty is again reflected in a rising numbers of queries, reaching a record value of over 15000 for TextTrojaners attacking in the RD domain.

Finally, table 7 shows the final leaderboard obtained by averaging the BODEGA scores across

Table 5
Results of the automatic evaluation of attacks against the **BERT** victims, for each of the five domains, expressed through BODEGA score and its constituents, as well as the number of queries (except for OpenFact, whose submission did not include this value).

Domain	Method	BODEGA	Confusion	Semantic	Character	Queries
PR	MMU_NLP	0.33	0.47	0.75	0.95	438.26
	OpenFact	**0.68**	0.97	0.77	0.89	-
	Palöri	0.56	0.97	0.64	0.88	65.37
	SINAI	0.38	0.43	0.92	0.98	288.73
	TextTrojaners	0.62	0.99	0.71	0.86	4,097.37
	TurQUaz	0.46	0.68	0.72	0.94	254.75
	BERT-ATTACK	0.43	0.70	0.68	0.90	80.16
	DeepWordBug	0.28	0.36	0.79	0.96	27.43
FC	MMU_NLP	0.55	0.73	0.78	0.96	710.29
	OpenFact	0.80	1.00	0.83	0.97	-
	Palöri	0.62	0.98	0.66	0.96	102.52
	SINAI	**0.82**	0.97	0.86	0.98	250.74
	TextTrojaners	0.79	1.00	0.83	0.96	1,390.83
	TurQUaz	0.74	1.00	0.78	0.95	70.70
	BERT-ATTACK	0.53	0.77	0.73	0.95	146.73
	DeepWordBug	0.44	0.53	0.84	0.98	54.32
RD	MMU_NLP	0.15	0.37	0.42	0.93	986.24
	OpenFact	**0.65**	0.78	0.86	0.95	-
	Palöri	0.18	0.45	0.42	0.94	1,077.24
	SINAI	0.12	0.14	0.87	1.00	143.06
	TextTrojaners	0.59	0.80	0.79	0.91	10,618.93
	TurQUaz	0.22	0.38	0.61	0.95	417.75
	BERT-ATTACK	0.18	0.44	0.43	0.96	774.31
	DeepWordBug	0.16	0.23	0.70	0.99	232.74
HN	MMU_NLP	0.47	0.86	0.55	0.97	806.60
	OpenFact	**0.91**	1.00	0.92	0.99	-
	Palöri	0.60	0.96	0.64	0.98	502.03
	SINAI	0.24	0.27	0.87	1.00	245.10
	TextTrojaners	0.85	1.00	0.87	0.97	4,327.67
	TurQUaz	0.46	0.84	0.59	0.93	223.54
	BERT-ATTACK	0.60	0.96	0.64	0.97	648.41
	DeepWordBug	0.22	0.29	0.78	1.00	395.94
C19	MMU_NLP	0.45	0.82	0.58	0.95	142.18
	OpenFact	**0.72**	0.91	0.82	0.96	-
	Palöri	0.52	0.96	0.57	0.93	201.01
	SINAI	0.41	0.47	0.89	1.00	32.16
	TextTrojaners	0.71	0.98	0.78	0.92	2,628.90
	TurQUaz	0.57	0.96	0.62	0.95	102.92
	BERT-ATTACK	0.42	0.74	0.60	0.95	161.70
	DeepWordBug	0.27	0.39	0.71	0.99	61.06

victims and domains. We can see that all solutions submitted to the task have beaten the DeepWordBug reference and most have also outperformed BERT-ATTACK, which is a strong reference point. OpenFact and TextTrojaners are clear leaders, with the former slightly better-performing, especially against Transformers victim. However, we need to emphasise that the averaged ranking does not show the whole picture and various methods work best for various scenarios. For example, SINAI is the best approach for the BERT-FC combination.

Table 6
Results of the automatic evaluation of attacks against the **Surprise** victims, for each of the five domains, expressed through BODEGA score and its constituents, as well as the number of queries (except for OpenFact, whose submission did not include this value).

Domain	Method	BODEGA	Confusion	Semantic	Character	Queries
PR	MMU_NLP	0.28	0.40	0.76	0.94	525.66
	OpenFact	**0.62**	0.93	0.75	0.87	-
	Palöri	0.25	0.54	0.55	0.83	482.21
	SINAI	0.26	0.31	0.89	0.97	374.20
	TextTrojaners	0.45	0.97	0.55	0.79	10,286.82
	TurQUaz	0.20	0.26	0.78	0.95	471.40
	BERT-ATTACK	0.20	0.32	0.69	0.91	117.64
	DeepWordBug	0.13	0.17	0.81	0.96	26.87
FC	MMU_NLP	0.51	0.68	0.78	0.96	201.15
	OpenFact	0.80	1.00	0.82	0.97	-
	Palöri	0.66	1.00	0.68	0.97	117.77
	SINAI	0.44	0.50	0.89	0.99	43.14
	TextTrojaners	**0.82**	1.00	0.84	0.97	498.93
	TurQUaz	0.71	1.00	0.75	0.94	90.55
	BERT-ATTACK	0.56	0.79	0.73	0.96	164.07
	DeepWordBug	0.37	0.46	0.83	0.98	53.39
RD	MMU_NLP	0.16	0.35	0.46	0.97	2,894.29
	OpenFact	**0.55**	0.71	0.82	0.93	-
	Palöri	0.19	0.47	0.43	0.93	1,513.25
	SINAI	0.09	0.10	0.85	1.00	149.12
	TextTrojaners	0.54	0.87	0.69	0.84	15,458.12
	TurQUaz	0.17	0.28	0.63	0.96	466.13
	BERT-ATTACK	0.17	0.41	0.42	0.95	951.87
	DeepWordBug	0.12	0.18	0.69	0.99	229.56
HN	MMU_NLP	0.47	0.77	0.62	0.97	713.89
	OpenFact	**0.83**	0.99	0.86	0.97	-
	Palöri	0.34	0.57	0.62	0.98	1,453.38
	SINAI	0.36	0.41	0.88	1.00	202.67
	TextTrojaners	0.67	1.00	0.72	0.92	4,596.62
	TurQUaz	0.28	0.47	0.61	0.94	376.77
	BERT-ATTACK	0.38	0.67	0.60	0.95	1,781.97
	DeepWordBug	0.16	0.21	0.76	1.00	384.34
C19	MMU_NLP	0.42	0.76	0.58	0.94	155.45
	OpenFact	**0.72**	0.99	0.78	0.93	-
	Palöri	0.46	0.99	0.51	0.89	299.37
	SINAI	0.17	0.18	0.92	1.00	37.97
	TextTrojaners	0.65	1.00	0.71	0.91	6,491.39
	TurQUaz	0.41	0.75	0.59	0.92	253.78
	BERT-ATTACK	0.37	0.68	0.58	0.93	198.26
	DeepWordBug	0.20	0.28	0.72	0.98	60.94

4.3. Manual Evaluation

We randomly selected 100 successful adversarial samples from each team's submission to the scenario including fact checking and surprise victim. During the evaluation, both the original and adversarial samples were presented to the annotators, with differences highlighted. The annotators were asked to categorise each sample pair into one of the three categories described in Section 3.2. They also provided a confidence score for each annotation (5: very confident, 1: not confident). Each sample pair was judged by at least two annotators. The annotation agreement of the initial annotators was 0.52 (Cohen's Kappa). A third annotator was invited if there was a conflict between the two initial annotators.

We determined the ranking of participants by the number of samples that fell into the 'Preserve the

Table 7
Final leaderboard, created by averaging BODEGA scores across all scenarios (five domains and three victims).

#	Method	BODEGA avg.
1.	OpenFact	0.7458
2.	TextTrojaners	0.7074
3.	TurQUaz	0.4859
4.	Palöri	0.4776
5.	MMU_NLP	0.3848
6.	SINAI	0.3507
-	BERT-ATTACK	0.4261
-	DeepWordBug	0.2682

Table 8
Manual evaluation results.

Team	% of Preserve the meaning
SINAI	99%
MMU_NLP	96%
TurQUaz	62%
Plagori	14%
OpenFact	11%
TextTrojaners	7%

Semantic Meaning' category, adhering to the principle that a higher count indicates better performance. This ranking method was used because the task demands that the adversarial samples maintain their original meaning. Table 8 presents the final manual evaluation results for all participants.

In general, we observe we observe a discrepancy between the manual and automatic evaluations (see Table 7). This may be because the manual evaluation task is a fact-checking task. Even a slight replacement of named entities (such as changing the year from 1990 to 1991) could result in a change of meaning.

The leading team (SINAI) in manual evaluation uses an adversarial attack method mainly based on the substitution of characters with homoglyphs and achieves a 99% score. This suggests that the use of homoglyphs can successfully deceive the annotator's eye. Similarly, team MMU employs similar attack approaches, such as lexical substitution and character replacement, which also achieve a high manual evaluation score (96%).

Team TurQUaz proposes a method of inserting white space to split English words, which achieved third place on the leaderboard (62%). However, this method may sometimes change the meaning of the original fact-checking document, resulting in a lower manual evaluation score compared to the methods proposed by teams SINAI and MMU. Besides, by inserting white space in the original text, the modified text may become non-interpretable by humans, resulting in a high proportion of third category submissions by Team TurQUaz, i.e., "The sentence does not make any sense."

We observe that OpenFact (11%) and TextTrojaners (7%) obtain lower manual evaluation scores. By manually investigating text modified by Team OpenFact, we notice that some key information, such as time and location, has been changed. Note that such named entities play a vital role in the context of fact-checking downstream tasks. Therefore, changes to these named entities result in a higher number of samples labelled as "Change the Semantic Meaning." As for the solution of TextTrojaners, the choice of alternative words solely depends on the context rather than the word to be replaced. This can significantly deviate the meaning. In addition, the operation of word removal without further word agreement adjustment may lead to nonsensical sentences.

5. Discussion

The first conclusion from the results obtained is clear: the state of the art in AE generation for misinformation detection, established by previous solutions, has advanced considerably. Various solutions were submitted to the shared task, but they are based on the established lines of research in the area: word replacements (preserving meaning similarity) or character replacements (preserving visual similarity).

The word-level solutions (esp. TextTrojaners and OpenFact) were performing the best in most scenarios, but not all of them. Fact checking is a clear outlier due to its nature: every word matters, making it hard to perform any change without drastically affecting the meaning. This opens the avenue for character-level modifications and, indeed, such solution (SINAI) provided the best results in manual evaluation.

We also need to acknowledge the limitations of the evaluation setup. It aims to predict the likelihood of an AE fooling the victim classifier and transmitting the intended message, as encoded through BODEGA score. However, human readers are the intended recipients of misinformation, and they are also able to refuse to engage with a message that seems suspicious, artificial or distorted, e.g. due to use of letter with non-standard shapes. Thus, the success of AEs will also depend on the visual appearance of the manipulated content, which is not directly evaluated in the current setup. Quantifying this effect would be challenging in manual evaluation and even more so in an automatic setup.

In any case, the results that we do have leave one thing clear: popular architectures for text classification are very vulnerable to attack with AEs. While the adversarially-trained model posed slightly harder challenge, ultimately AEs were found for nearly all cases in these scenarios as well.

How can we protect the real-world deployments of text classifiers against such attack? The first barrier can be established by limiting the access to the victim model. We can see from the results that the advance over previous state of the art was accompanied by a raise in the number of queries sent, well into hundreds and thousands for each generated AE sample. Nevertheless, all machine-learning-based solutions for content filtering should be only deployed after a thorough analysis of their adversarial robustness.

6. Conclusion

In the InCrediblAE shared task, six teams participated with various solutions, both operating with the word-level and character-level changes. The participants' approaches (and two reference solutions) were evaluated using five misinformation-detection scenarios and three victim models.

In total, 53,544 text modifications were considered and automatically assessed in terms of classifier confusion, meaning preservation and character similarity. The submitted solutions easily outperformed previous work in all of the tested scenarios. The manual evaluation highlighted the special role of the fact-checking tasks and the efficacy of character replacement in performing modification imperceptible to humans.

We hope that the combined effort of the participants and organisers of the InCrediblAE shared task will succeed in both highlighting the importance of robustness testing and showcasing the best solutions. To facilitate this outcome, the code and resources necessary for performing the automatic evaluation remain openly available.[9]

Acknowledgments

The work of P. Przybyła is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. This work has been also partially funded by the European Commission under contract numbers HE-101070278 and ISF-101080090. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting

[9] https://github.com/piotrmp/BODEGA

authority can be held responsible for them. We also acknowledge support from Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021) and from Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MCIN/AEI /10.13039/501100011033.

References

[1] J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, B. Nyhan, Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature, Technical Report, Hewlett Foundation, 2018. URL: https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/.

[2] S. van der Linden, Misinformation: susceptibility, spread, and interventions to immunize the public, Nature Medicine 2022 28:3 28 (2022) 460–467. URL: https://www.nature.com/articles/s41591-022-01713-6. doi:10.1038/s41591-022-01713-6.

[3] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, M. Potthast, SemEval-2019 Task 4: Hyperpartisan News Detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 829–839. URL: https://aclanthology.org/S19-2145. doi:10.18653/v1/S19-2145.

[4] 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: Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020), 2020, pp. 1377–1414. URL: http://propaganda.qcri.org/annotations/definitions.htmlhttp://arxiv.org/abs/2009.02696. arXiv:2009.02696.

[5] F. Rangel, P. Rosso, Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CEUR-WS.org, 2019.

[6] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The FEVER2.0 Shared Task, in: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 2018.

[7] M. Singhal, C. Ling, P. Paudel, P. Thota, N. Kumarswamy, G. Stringhini, S. Nilizadeh, SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice, in: The 8th IEEE European Symposium on Security and Privacy (EuroS&P 2023), IEEE, 2022. URL: https://arxiv.org/abs/2206.14855v2. doi:10.48550/arxiv.2206.14855. arXiv:2206.14855.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv: 1312.6199 (2013). URL: https://arxiv.org/abs/1312.6199v4. doi:10.48550/arxiv.1312.6199. arXiv:1312.6199.

[9] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial Attacks on Deep-learning Models in Natural Language Processing, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020). URL: https://dl.acm.org/doi/10.1145/3374217. doi:10.1145/3374217.

[10] 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.

[11] 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.

[12] P. Przybyła, A. Shvets, H. Saggion, Verifying the Robustness of Automatic Credibility Assessment, arXiv preprint arXiv:2303.08032 (2023). URL: https://arxiv.org/abs/2303.08032v1. arXiv:2303.08032.

[13] 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: ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the System Demonstrations, Association for Computational Linguistics (ACL), 2021, pp. 363–371. URL: https://aclanthology.org/2021.acl-demo.43. doi:10.18653/V1/2021.ACL-DEMO.43. arXiv:2009.09191.

[14] B. D. Horne, S. Adali, This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News, in: Proceedings of the 2nd International Workshop on News and Public Opinion at ICWSM, Association for the Advancement of Artificial Intelligence, 2017. URL: http://arxiv.org/abs/1703.09398. arXiv:1703.09398.

[15] P. Przybyła, Capturing the Style of Fake News, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), volume 34, AAAI Press, New York, USA, 2020, pp. 490–497. URL: https://aaai.org/ojs/index.php/AAAI/article/view/5386. doi:10.1609/aaai.v34i01.5386.

[16] T. J. Smith, Propaganda: A Pluralistic Perspective, Praeger, 1989.

[17] L. Graves, Understanding the Promise and Limits of Automated Fact-Checking, Technical Report, Reuters Institute, University of Oxford, 2018. URL: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves{_}factsheet{_}180226FINAL.pdf. arXiv:arXiv:1011.1669v3.

[18] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The Fact Extraction and VERification (FEVER) Shared Task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018. arXiv:1811.10971v1.

[19] M. Al-Sarem, W. Boulila, M. Al-Harby, J. Qadir, A. Alsaeedi, Deep learning-based rumor detection on microblogging platforms: A systematic review, IEEE Access 7 (2019) 152788–152812. doi:10.1109/ACCESS.2019.2947855.

[20] 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, ASONAM 2019, Association for Computing Machinery, Inc, 2019, pp. 105–112. URL: https://dl.acm.org/doi/10.1145/3341161.3342892. doi:10.1145/3341161.3342892. arXiv:1907.07033.

[21] 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: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023, pp. 556–567.

[22] Y. Mu, Y. Jiang, F. Heppell, I. Singh, C. Scarton, K. Bontcheva, X. Song, A large-scale comparative study of accurate covid-19 information versus misinformation, arXiv preprint arXiv:2304.04811 (2023).

[23] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[24] 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, Association for Computational Linguistics, 2018, pp. 4171–4186. URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[26] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7. arXiv:1711.05101v3.

[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov,

P. G. Allen, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). URL: https://arxiv.org/abs/1907.11692v1. doi:10.48550/arxiv.1907.11692. arXiv:1907.11692.

[28] 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), Association for Computational Linguistics, 2020, pp. 6193–6202. URL: https://aclanthology.org/2020.emnlp-main.500. doi:10.18653/v1/2020.emnlp-main.500.

[29] 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, 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.

[30] V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady 10 (1966) 707–710.

[31] Y. Mu, M. Jin, C. Grimshaw, C. Scarton, K. Bontcheva, X. Song, Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 17, 2023, pp. 1052–1062.

[32] D. Wilby, T. Karmakharm, I. Roberts, X. Song, K. Bontcheva, GATE Teamware 2: An open-source tool for collaborative document classification annotation, in: D. Croce, L. Soldaini (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 145–151. URL: https://aclanthology.org/2023.eacl-demo.17. doi:10.18653/v1/2023.eacl-demo.17.

[33] J. Valle Aguilera, A. J. Gutiérrez Megías, S. M. Jiménez Zafra, L. A. Ureña López, E. Martínez Cámara, SINAI at CheckThat! 2024: Stealthy character-level adversarial attacks using homoglyphs and search, iterative, in: [45], 2024.

[34] C. Roadhouse, M. Shardlow, A. Williams, MMU NLP at CheckThat! 2024: Homoglyphs are adversarial attacks, in: [45], 2024.

[35] H. He, Y. Song, D. Massey, Palöri at CheckThat! 2024 shared task 6: Glota - combining glove embeddings with roberta for adversarial attack, in: [45], 2024.

[36] W. Lewoniewski, P. Stolarski, M. Stróżyna, E. Lewańska, A. Wojewoda, E. Księżniak, M. Sawiński, OpenFact at CheckThat! 2024: Combining multiple attack methods for effective adversarial text generation, in: [45], 2024.

[37] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, K.-W. Chang, Generating Natural Language Adversarial Examples, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2890–2896. URL: https://aclanthology.org/D18-1316. doi:10.18653/v1/D18-1316.

[38] N. Mrkšić, D. Ó. Séaghdha, B. Thomson, M. Gasic, L. M. R. Barahona, P.-H. Su, D. Vandyke, T.-H. Wen, S. Young, Counter-fitting word vectors to linguistic constraints, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 142–148.

[39] 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: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020,, AAAI Press, 2020, pp. 8018–8025. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6311.

[40] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, Y. Qi, TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 119–126. URL: https://aclanthology.org/2020.emnlp-demos.16. doi:10.18653/v1/2020.emnlp-demos.16.

[41] D. Li, Y. Zhang, H. Peng, L. Chen, C. Brockett, M.-T. Sun, W. B. Dolan, Contextualized perturbation for textual adversarial attack, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5053–5069.

[42] B. Demirok, M. Kutlu, S. Mergen, B. Oz, TurQUaz at CheckThat! 2024: Creating adversarial examples using genetic algorithm, in: [45], 2024.
[43] D. Guzman Piedrahita, A. Fazla, L. Krauter, TextTrojaners at CheckThat! 2024: Robustness of credibility assessment with adversarial examples through beamattack, in: [45], 2024.
[44] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[45] 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.

A. Annotation Guidelines

InCrediblAE Shared Task Annotation All participants who submit entries will have their submissions manually evaluated for semantic similarity. To aid in this human evaluation process, participants in the shared tasks will be required to manually assess a selection of the submitted examples.

Manual Scoring We will gather assessments regarding the semantic similarities between attack samples and the original samples. Participants in the shared task are requested to dedicate approximately 60 mins to this manual evaluation (i.e., 100 samples per participant), which will be conducted using this online annotation tool. Judges will rate the sample pairs based on the following scale: (a). Preserve the Semantic Meaning, (b). Change the Semantic Meaning, (c). No sense.

Labels and Definitions (a). Preserve the Semantic Meaning: This label is used when the semantic content of the attack sample closely aligns with that of the original sample. Participants should use this label if the meaning, context, and intent of the compared texts remain essentially unchanged, indicating that the attack sample has effectively maintained the core message of the original. For example:

> **Text Original:** middle-earth. middle-earth is the fictional setting of much of british writer j. r. r. tolkien 's legendarium . middle-earth is a setting.
>
> **Text Modified:** middle-earth. middle-earth is the fictional background of much of british writer j. r. r. tolkien 's legendarium . middle-earth is a setting.

(b). Change the Semantic Meaning: Participants should apply this label when there is a noticeable alteration in the semantic content between the original and the attack sample. This label indicates that while the attack sample may be related or similar in some aspects to the original, it diverges enough in meaning or intent to be considered distinct or modified. For example, name entities (including Year, Name, Location, etc.) are changed in the modified text. For example:

> **Text Original:** carol danvers. the character has also been known as binary , warbird , and captain marvel at various points in her history , and has been featured in other marvel licensed products including video games , animated television series , and merchandise such as trading cards . carol danvers is a real person.
>
> **Text Modified:** carol danvers. the character has also been known as binary , warbird , and captain marvel at various points in her history , and has been featured in other marvel licensed products including video games , animated television series , and merchandise such as trading cards . carol danvers is a fictional person.

(c). No sense: The sentence does not make any sense.

Confidence Scores Please indicate how confident you are in your assigned class in the Confidence Row. The confidence scores range from 1 to 5.

A.1. Information Sheet for InCrediblAE Participants

You are invited to participate in this research by contributing to the evaluation of the semantic similarity of adversarial examples. It is important for you to understand the goals of the task and what your participation will involve. Please take the time to read the following information. Please, ask us if there is anything that is not clear or if you would like more information. Thank you very much for your time.

What is the task's purpose? The aim of this research is to evaluate the robustness of the text classifier in adversarial attacks (a detailed description of the task can be found here: https://checkthat.gitlab.io/clef2024/task6/). In this manual evaluation, we will gather assessments regarding the semantic similarities between attack samples and the original samples. Participants in the shared task are requested to dedicate approximately 8 hours to this manual evaluation, which will be conducted using an online tool. Judges will rate the sample pairs based on the following scale: 3 (Preserve the Semantic Meaning), 2 (Change the Semantic Meaning), and 1 (No sense).

Your participation, what it involves and why we are grateful It is up to you to decide whether or not you want to participate in this annotation task. If you do decide to support us in the project, you will be given this information sheet to keep (and be asked to sign a separate consent form). You can still withdraw at any time without any consequences and without giving any reason. The entire data collected from withdrawn participants will be destroyed immediately, and no personal information will be kept. If participants submitted a solution to the shared task that also means their submission to the shared tasks will not be manually scored. If you wish to withdraw, please contact Dr Xingyi Song (details in section 10). You will be asked to annotate adversarial examples into 4 categories using the GATE Teamware Platform (https://annotate.gate.ac.uk/).

What are the possible advantages, disadvantages and risks of being involved? Participating in this evaluation will support our research on the development of a more accurate assessment of the robustness of text classifiers and effectiveness of the adversarial attack methods. No major disadvantages or risks are foreseen, however, it is worth mentioning that the content being annotated may case distress. You will be in charge of selecting the content to be annotated, therefore, you are free to only select content that you are comfortable with. If, at any part of the experiment you feel uncomfortable with the content you are accessing, please talk to one of the responsible researchers.

Will my involvement be kept confidential? All the information that we collect from you and about you during the course of the research will be kept strictly confidential and will only be accessible to members of the research team. You will not be able to be identified in any reports or publications unless you have given your explicit consent for this on your participant consent form.

What is the legal basis for processing my personal data? According to data protection legislation, we are required to inform you that the legal basis we are applying in order to process your personal data is that 'processing is necessary for the performance of a task carried out in the public interest' (Article 6(1)(e)). Further information can be found in the University of Sheffield's Privacy Notice, which is available online under https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

What will happen to the data collected in this study? The data gathered during this scoring task will be used to assess the effectiveness of adversarial examples generation. Since this data will also benefit other researchers, we plan to release a version of the annotated dataset. You will not be identified and your scoring will be aggregated with multiple other annotators. In the case that you provided us with your e-mail address at the beginning of the scoring task, we will destroy it after the scoring task is finished. The University of Sheffield will act as the Data Controller for this study.

Who is organising and funding the research? This study is organised jointly by Universitat Pompeu Fabra and the University of Sheffield at the CheckThat! Lab at CLEF 2024. Universitat Pompeu Fabra is funded by the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. The University of Sheffield is funded by the the UK's innovation agency (Innovate UK) grant 10039055 (approved under the Horizon Europe Programme as vera.ai EU grant agreement 101070093).

Who has ethically reviewed this study? This project has been ethically approved via the University of Sheffield's Ethics Review Procedure, as administered by the Computer Science Department.

What if something goes wrong and I wish to complain about the research? If you have any complaints, either from the researcher or something occurring during or following your participation in the project (e.g. a reportable serious adverse event), please contact Dr. Xingyi Song (contact details in section 10). Should you feel your complaint has not been handled to your satisfaction, you can also contact the Head of Department at the University of Sheffield, Professor Heidi Christensen (heidi.christensen@sheffield.ac.uk) who will then escalate the complaint through the appropriate channels. If the complaint relates to how your personal data has been handled, information about how to raise a complaint can be found in the University's Privacy Notice: https://www.sheffield.ac.uk/govern/data-protection/privacy/general.

Contact for further information Details of who you should contact if you wish to obtain further information are as follows: Dr. Xingyi Song, Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK. E-mail: x.song@sheffield.ac.uk Telephone: +44 114 222 18577.

B. COVID-19 Misinformation Dataset Examples

Table 9
Positive and negative examples from the COVID-19 misinformation dataset (C19).

Label	False Claim	User Text
1 (Misinformation related to false claim)	N95 masks block few, if any COVID-19 particles due to their size	COVID-19 the average diameter of the virus particles is around 120 nm (.12μm). Any mask including N95 masks can't filter this particle size, to filter it completely would prevent the subject being able to breath. Therefore, the mask are just the new normal, a fashion statement.
0 (Other)	N95 masks block few, if any COVID-19 particles due to their size	A #COVID19 particle is about 1 to 4 microns, an N95 will block 95% of tiny air particles, down to 0.3 microns, and surgical masks aren't effective at blocking particles smaller than 100 microns. Hence surgical masks cannot stop the spread of #COVID19.