Natural Language Processing (2024), 1–29
doi:10.1017/nlp.2024.54

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Verifying the robustness of automatic credibility assessment

Piotr Przybyła[1,2], Alexander Shvets[1] and Horacio Saggion[1]

[1]Universitat Pompeu Fabra, Barcelona, Spain and [2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
Corresponding author: Piotr Przybyła; Email: piotr.przybyla@upf.edu

(Received 28 May 2024; revised 10 October 2024; accepted 10 October 2024)

Abstract

Text classification methods have been widely investigated as a way to detect content of low credibility: fake news, social media bots, propaganda, etc. Quite accurate models (likely based on deep neural networks) help in moderating public electronic platforms and often cause content creators to face rejection of their submissions or removal of already published texts. Having the incentive to evade further detection, content creators try to come up with a slightly modified version of the text (known as an attack with an adversarial example) that exploit the weaknesses of classifiers and result in a different output. Here we systematically test the robustness of common text classifiers against available attacking techniques and discover that, indeed, meaning-preserving changes in input text can mislead the models. The approaches we test focus on finding vulnerable spans in text and replacing individual characters or words, taking into account the similarity between the original and replacement content. We also introduce BODEGA: a benchmark for testing both victim models and attack methods on four misinformation detection tasks in an evaluation framework designed to simulate real use cases of content moderation. The attacked tasks include (1) fact checking and detection of (2) hyperpartisan news, (3) propaganda, and (4) rumours. Our experimental results show that modern large language models are often more vulnerable to attacks than previous, smaller solutions, e.g. attacks on GEMMA being up to 27% more successful than those on BERT. Finally, we manually analyse a subset adversarial examples and check what kinds of modifications are used in successful attacks.

Keywords: Adversarial examples; credibility assessment; robustness; misinformation; benchmark

1. Introduction

Misinformation is one of the most commonly recognised problems in modern digital societies (Lewandowsky, Ecker, and Cook 2017; Akers *et al.* 2018; Tucker *et al.* 2018). Under this term, we understand the publication and spreading of information that is not *credible*, including fake news, manipulative propaganda, social media bots activity, rumours, hyperpartisan and biased journalism. While these problems differ in many aspects, what they have in common is non-credible (fake or malicious) content masquerading as credible: fake news as reliable news, bots as genuine users, falsehoods as facts, etc. (Tucker *et al.* 2018; van der Linden 2022).

Given that both credible and non-credible content is abundant on the Internet, the assessment of credibility has fast been recognised as a task for machine learning (ML) or wider artificial intelligence (AI) solutions (Ciampaglia *et al.* 2018). It is common practice among major platforms with user-generated content to use such models for moderation, either as preliminary filtering

before human judgement (Singhal *et al.* 2022) or as an automated detection system, for example in *Google*[a] and *Twitter* (Paul and Dang 2022).

Are the state-of-the-art techniques of ML and, in particular, Natural Language Processing (NLP), up for a task of great importance to society? The standard analysis of model implementation with traditional accuracy metrics does not suffice here as it neglects how possible it is to systematically come up with variants of malicious text, known as *adversarial examples* (AEs), that fulfil the original goal but evade detection (Carter, Tsikerdekis, and Zeadally, 2021). A realistic analysis in such a use case has to take into account an *adversary*, that is the author of the non-credible content, who has both motivation and opportunity to experiment with the filtering system to find out its vulnerabilities.

For example, consider a scenario in which a foreign actor aims to incite panic by spreading false information about a hazardous fallout, under alarming headings such as *Radioactive dust approaching after fire in a Ukrainian power plant!*.[b] If analogous scenarios were explored in the past, the content-filtering systems in social media platforms will likely block such a message. But the adversary might come up with an adversarial example *Radioactive dust* ***coming*** *after fire in a Ukrainian power plant!*. If the classifier is not robust and returns a different decision for this variant, the attacker succeeds.

Looking for such weaknesses via designing AE, to assess the *robustness* of an investigated model, is a well-established problem in ML. However, its application to misinformation-oriented NLP tasks is relatively rare, despite the suitability of the adversarial scenario in this domain. Moreover, similarly to the situation in other domains, the adversarial attack performance depends on a variety of factors, such as the data used for training and testing, the attack goal, disturbance constraints, attacked models, and evaluation measures. The common approach to measuring the attack success, that is by computing accuracy reduction, requires the definition of the maximum allowed change, with no clear way to define it across various tasks. It also ignores the number of queries to the victim model, which can decide the practical applicability of an attack.

In order to fill the need for reproducible and comprehensive evaluation in this field, we have created BODEGA (Benchmark fOr aDversarial Example Generation in credibility Assessment), intended as a common framework for comparing AE generation solutions to inform the creation of "better-defended" content credibility classifiers. We have used it to assess the robustness of the popular text classifiers, including state-of-the-art large language models, by simulating attacks using various AE generation solutions.

Thus, our contributions include the following:

1. The BODEGA evaluation framework, consisting of elements simulating the misinformation detection scenario:
   (a) A collection of four NLP tasks from the domain of misinformation, cast as binary text classification problems (Section 4),
   (b) A training and test dataset for each of the above tasks,
   (c) Two attack scenarios, specifying what information is available to an adversary and what is their goal (Section 5),
   (d) An evaluation procedure, involving a success measure designed specifically for this scenario (Section 6).
2. A systematic evaluation of the robustness of common text classification solutions of various sizes, answering several questions (Section 9):
   - Q1: Which attack method delivers the best performance?

[a] https://support.google.com/youtube/thread/192701791/updates-on-comment-spam-abuse?hl=en
[b] Similar messages were shared in a 2020 wide-scale misinformation campaign in Poland (Mierzyńska 2020).

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

- Q2: Are the modern large language models less vulnerable to attacks than their predecessors?
- Q3: How many queries are needed to find adversarial examples?
- Q4: Does targeting (selecting only some examples for AE generation) make a difference in attack difficulty?

3. A manual analysis of the most promising cases, revealing the kinds of modifications used by the AE solutions to confuse the victim models (Section 9.5).

BODEGA, based on the *OpenAttack* framework and existing misinformation datasets, is openly available for download.[c] It can be used to evaluate the effectiveness of emerging attack strategies, as well as to test the robustness of a classifier being prepared for deployment. Both of these applications can serve to improve the reliability of text classification, in content filtering and elsewhere.

2. Related work

2.1 Adversarial examples in NLP

Searching for adversarial examples can be seen within wider efforts to investigate the *robustness* of ML models, that is their ability to maintain good performance when confronted with data instances unlike those seen in training: anomalous, rare, adversarial or edge cases. This effort is especially important for deep learning models, which are not inherently interpretable, making it harder to predict their behaviour at the design stage. The seminal work on the subject by Szegedy *et al.* (2013) demonstrated the low robustness of neural networks used to recognise images. The adversarial examples were prepared by adding specially prepared noise to the original image, which forced the change of the classifier's decision even though the changes were barely perceptible visually and the original label remained valid.

Given the prevalence of neural networks in language processing, a lot of work has been done on investigating AEs in the context of NLP tasks (Zhang *et al.* 2020b), but the transition from the domain of images to text is far from trivial. Firstly, it can be a challenge to make changes small enough to the text, such that the original label remains applicable—there is no equivalent of *imperceptible noise* in text. The problem has been approached on several levels: of characters, making alterations that will likely remain unnoticed by a reader (Gao *et al.* 2018; Eger *et al.* 2019); of words, replaced while preserving the meaning by relying on thesauri (Ren *et al.* 2019) or language models (Jin *et al.* 2020; Li *et al.* 2020) and, finally, of sentences, by employing paraphrasing techniques (Iyyer *et al.* 2018; Ribeiro, Singh, and Guestrin 2018). Secondly, the discrete nature of text means that methods based on exploring a feature space (e.g. guided by a gradient) might suggest points that do not correspond to real text. Most of the approaches solve this by only considering modifications on the text level, but there are other solutions, for example finding the optimal location in the embedding space followed by choosing its nearest neighbour that is a real word (Gong *et al.* 2018), or generating text samples from a distribution described by continuous parameters (Guo *et al.* 2021). Note that these solutions are evaluated on different datasets, making it hard to compare their performance. We are aware of only one previous attempt to establish a reusable benchmark (Yoo *et al.* 2022), which relies on datasets for the classification of topics and sentiment.

Apart from AE generation, a public-facing text classifier may be subject to many other types of attacks, including manipulations to output desired value when a trigger word is used (Bagdasaryan and Shmatikov 2022) or perform an arbitrary task chosen by the attacker (Neekhara *et al.* 2019). Finally, verifying the trustworthiness of a model aimed for deployment should also take into

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

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

account undesirable behaviours exhibited without adversarial actions, for example its response to modification of protected attributes, such as gender, in the input (Srivastava *et al.* 2023).

2.2 Robustness of credibility assessment

The understanding that some deployment scenarios of NLP models justify expecting adversary actions predates the popularisation of deep neural networks, with the first considerations based on spam detection (Dalvi *et al.* 2004). The work that followed was varied in the explored tasks, attack scenarios and approaches.

The first attempts to experimentally verify the robustness of misinformation detection were based on simple manual changes (Zhou *et al.* 2019). The approach of targeting a specific weakness and manually designing rules to exploit it has been particularly popular in attacking fact-checking solutions (Thorne *et al.* 2019; Hidey *et al.* 2020).

In the domain of social media analysis, Le *et al.* (2020) have examined the possibility of changing the output of a text credibility classifier by concatenating it with adversarial text, for example added as a comment below the main text. The main solution was working in the white-box scenario, with the black-box variant made possible by training a surrogate classifier on the original training data.[d] It has also been shown that social media bot detection using AdaBoost is vulnerable to adversarial examples (Kantartopoulos *et al.* 2020). Adversarial scenarios have also been considered with user-generated content classification for other tasks, for example hate speech or satire (Alsmadi *et al.* 2022).

Fake news corpora have been used to verify the effectiveness of AE generation techniques, for example in the study introducing TextFooler (Jin *et al.* 2020). Interestingly, the study has shown that the classifier for fake news was significantly more resistant to attacks compared to those for other tasks, that is topic detection or sentiment analysis. This task also encouraged exploration of vulnerability to manually crafted modifications of input text (Jaime, Flores, and Hao 2022). In general, the fake news classification task has been a common subject of robustness assessment, involving both neural networks (Ali *et al.* 2021; Koenders *et al.* 2021) and non-neural classifiers (Brown *et al.* 2020; Smith *et al.* 2021).

To sum up, while there have been several experiments examining the vulnerability of misinformation detection to adversarial attacks, virtually each of them has used a different dataset, a different classifier and a different attack technique, making it hard to draw conclusions and make comparisons. Our study is the first to analyse credibility assessment tasks and systematically evaluate their vulnerability to various attacks.

2.3 Resources for adversarial examples

The efforts of finding AEs are relatively new for NLP, and there exist multiple approaches to evaluation procedures and datasets. The variety of studies for the misinformation tasks is reflective of the whole domain—see the list of datasets used for evaluation provided by Zhang *et al.* (2020b). Hopefully, as the field matures, some standard practice measures will emerge, facilitating the comparison of approaches. We see BODEGA as a step in this direction.

Two types of existing efforts to bring the community together are worth mentioning. Firstly, some related shared tasks have been organised. The *Build It Break It, The Language Edition* task (Ettinger *et al.* 2017) covered sentiment analysis and question answering, addressed by both 'builders' (building solutions) and 'breakers' (finding adversarial examples). The low number of breaker teams—four for sentiment analysis and one for question answering—makes it difficult to draw conclusions, but the majority of deployed techniques involved manually inserted changes targeting suspected weaknesses of the classifiers. The FEVER 2.0 shared task (Thorne *et al.* 2018b),

[d]We explain white- and black-box scenarios in Section 5.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

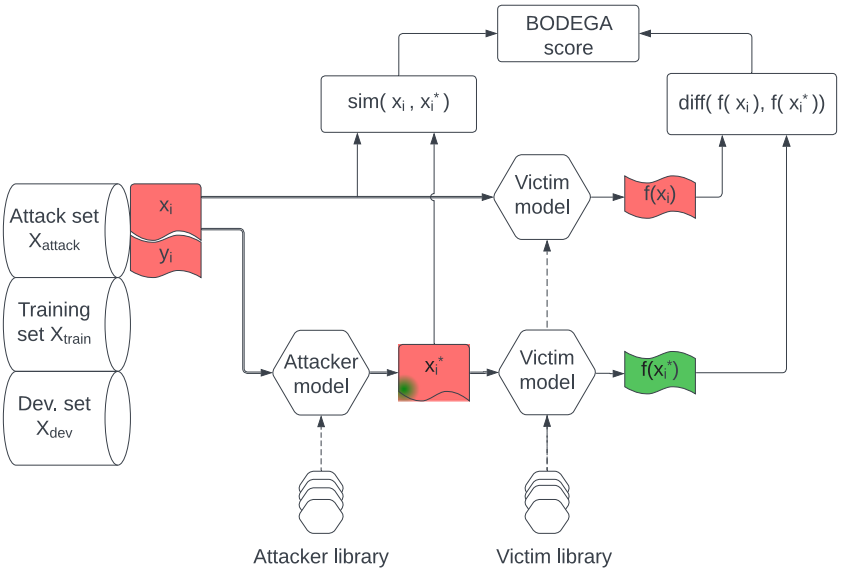

Figure 1. An overview of the evaluation of an adversarial attack using BODEGA. For each task, three datasets are available: development (X_{dev}), training (X_{train}), and attack (X_{attack}). During an evaluation of an attack involving an Attacker and Victim models from the library of available models, the Attacker takes the text of the ith instance from the attack dataset (x_i), e.g. a news piece, and modifies it into an adversarial example (x_i^*). The Victim model is used to assess the credibility of both the original ($f(x_i)$) and modified text ($f(x_i^*)$). The BODEGA score assesses the quality of an AE, checking the similarity between the original and modified sample (sim(x_i, x_i^*)), as well as the change in the victim's output (diff($f(x_i), f(x_i^*)$)).

focusing on fact checking, had a 'Build-It' and 'Break-It' phases with a similar setup, except the adversarial examples were generated and annotated from scratch, with no correspondence to existing true examples, as in *Build It Break It* or BODEGA. The three valid submissions concentrated around manual introduction of issues known as challenging for automated fact checking, including multi-hop or temporal reasoning, ambiguous entities, arithmetic calculations and vague statements.

Secondly, two software packages were released to aid evaluation: *TextAttack* (Morris *et al.* 2020) and *OpenAttack* (Zeng *et al.* 2021). They both provide a software skeleton for setting up the attack and implementations of several AE generation methods. A user can add the implementation of their own victims and attackers and perform the evaluation. BODEGA code has been developed based on OpenAttack by providing access to misinformation-specific datasets, classifiers and evaluation measures.

3. Adversarial example generation

Adversarial example generation is a task aimed at testing the robustness of ML models, known as *victims* in this context. The goal is to find small modifications to the input data that will change the model output even though the original meaning is preserved and the correct response remains the same. If such changed instances, known as adversarial examples, could be systematically found, it means the victim classifier is vulnerable to the attack and not robust.

In the context of classification, this setup (illustrated in Fig. 1) could be formalised through the following:

- A training set X_{train} and an attack set X_{attack}, each containing instances (x_i, y_i), coupling the i-th instance features x_i with its true class y_i,

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

- A victim model f, predicting a class label $\hat{y}_i$ based on instance features: $\hat{y}_i = f(x_i)$,
- A modification function (attack model) m, turning x_i into an adversarial example $x_i^* = m(x_i)$.

Throughout this study, we use $y_i = 1$ (positive class) to denote non-credible information and 0 for credible content.

The goal of the attacker is to come up with the m function. This process typically involves generating numerous variations of x_i and querying the model's response to them until the best candidate is selected. An evaluation procedure assesses the success of the attack on the set X_{attack} by comparing x_i to x_i^* (which should be maximally similar) and $f(x_i)$ to $f(x_i^*)$ (which should be maximally different).

Consider the following real example observed in our evaluation:

1. Within the propaganda recognition task, one of the instances in X_{attack} contains a text fragment x_i ='*Despite the hysteria of the left, it is impossible to see the Trump administration as anything but firm in its dealing with Russia.*', labelled as $y_i = 1$ (propaganda technique used).
2. The victim classifier (BiLSTM) correctly assigns the label $f(x_i) = 1$ with 94.76% certainty.
3. An attacker (BERT-ATTACK) tests 26 different reformulations of the text, until it comes up with the modified version: $x_i^* = m(x_i)$ ='***Given*** *the hysteria of the left, it is impossible to see the Trump administration as anything but firm in its dealing with Russia.*'
4. The victim classifier changes its decision after the modification, assigning $f(x_i^*) = 0$ (no propaganda) with 54.65% certainty.
5. This example is considered a good-quality AE, since it achieves a change in the classifier's decision ($f(x_i) \neq f(x_i^*)$) with a small change in text meaning.

4. BODEGA tasks

In BODEGA, we include four misinformation detection tasks:

- Hyperpartisan news (HN),
- Propaganda recognition (PR),
- Fact checking (FC),
- Rumour detection (RD).

For each of these problems, we rely on an already established dataset with credibility labels provided by expert annotators. The tasks are all presented as text classification.

Whenever data split is released with a corpus, the training subset is included as X_{train}—otherwise we perform a random split. In order to enable the evaluation of AE generation solutions that carry a high computational cost, we define the X_{attack} subset which is restricted to around 400 instances taken from the test set. The rest of the cases in the original test set are left out for future use as a development subset. Table 1 summarises the data obtained.

Table 2 includes some examples of the credible and non-credible content in each task. We can see how the non-credible examples often focus on particularly politically charged topics, trying to provoke an emotional reaction in readers. This is a well-known aspect of misinformation (Bakir and McStay 2017; Allcott and Gentzkow 2017). In the following subsections, we outline the motivation, origin and data processing within each of the tasks.

4.1 HN: hyperpartisan news

Solutions for news credibility assessment, sometimes equated with *fake news* detection, usually rely on one of three factors: (1) writing style (Horne and Adali 2017; Przybyła, 2020), (2) veracity

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 1. Four datasets used in BODEGA, with the task ID (see descriptions in text), number of instances in training, attack and development subsets, and an overall percentage of positive (non-credible) class

Task	Training	Attack	Dev.	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%

of included claims (Vlachos and Riedel 2014; Graves 2018) or (3) context of social and traditional media (Shu, Wang, and Liu 2019; Liu and Wu 2020).

In this task, we focus on the writing style. This means a whole news article is provided to a classifier, which has no ability to check facts against external sources, but has been trained on enough articles to recognise stylistic cues. The training data include numerous articles coming from sources with known credibility, allowing one to learn writing styles typical for credible and non-credible outlets.

In BODEGA, we employ a corpus of news articles (Potthast *et al.* 2018) used for the task of *Hyperpartisan News Detection* at SemEval-2019 (Kiesel *et al.* 2019). The credibility was assigned based on the overall bias of the source, assessed by journalists from *BuzzFeed* and *MediaBiasFactCheck.com*.[e] We use 1/10th of the training set (60,235 articles) and assign label 1 (non-credible) to articles from sources annotated as hyperpartisan, both right- and left-wing.

See the first row of Table 2 for examples: credible from *Albuquerque journal*[f] and non-credible from *Crooks and Liars*.[g]

4.2 PR: propaganda recognition

The task of propaganda recognition involves detecting text passages, whose author tries to influence the reader by means other than objective presentation of the facts, for example by appealing to emotions or exploiting common fallacies (Smith 1989). The usage of propaganda techniques does not necessarily imply falsehood, but in the context of journalism it is associated with manipulative, dishonest and hyperpartisan writing. In BODEGA, we use the corpus accompanying SemEval 2020 Task 11 (*Detection of Propaganda Techniques in News Articles*), with 14 propaganda techniques annotated in 371 newspaper articles by professional annotators (da San Martino *et al.* 2020).

Propaganda recognition is a fine-grained task, with SemEval data annotated on the token level, akin to a Named Entity Recognition task. In order to cast it as a text classification problem as others here, we split the text on sentence level and assign target label equal 1 to sentences overlapping with any propaganda instances and 0 to the rest. Because only the training subset is made publicly available,[h] we randomly extract 20 per cent of documents for attack and development subsets.

See the second row of Table 2 for examples—the credible fragment with no propaganda technique and the non-credible, annotated as including flag-waving.

[e] https://zenodo.org/record/1489920
[f] https://abqjournal.com/328734/syria-blamed-for-missed-deadline-on-weapons.html
[g] http://crooksandliars.com/2014/12/foxs-cavuto-and-stein-try-conflate
[h] https://zenodo.org/record/3952415

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 2. Examples of credible and non-credible content in each of the tasks: hyperpartisan news (HN), propaganda recognition (PR), fact checking (FC) and rumour detection (RD). See main text for references to data sources and labelling criteria

Task	Credible example	Non-credible example
HN	Syria blamed for missed deadline on chemical arsenal U.S. officials conceded that a Tuesday deadline for ridding Syria of hundreds of tons of liquid poisons would not be met, citing stalled progress in transporting the chemicals across war-ravaged countryside to ships that will carry them out of the region. But the officials insisted that the overall effort to destroy President Bashar Assad's chemical arsenal was on track. "We continue to make progress, which has been the important part," State Department spokeswoman Marie Harf told reporters. "It was always an ambitious timeline, but we are still operating on the June 30th timeline for the complete destruction." (. . .)	Fox's Cavuto And Stein Try To Conflate 'Grubergate' With Vietnam And The Pentagon Papers Over at Faux "news" this Tuesday, rather than focus on the newly released Senate torture report, it's been all Jonathan Gruber and "Grubergate" all the time and wall to wall coverage of another one of Darrell Issa's Obamacare witch hunts, otherwise known as a House Oversight Committee hearing. As soon as I heard the hearing was scheduled I knew that it meant things were going to get ugly over at Fox, but not even in my wildest imagination could I have come up with this big giant turd that Neil Cavuto and his buddy Ben Stein managed to toss against the wall to attack Obamacare and Gruber. (. . .)
PR	Leading Democratic senators like Robert Menendez, Ben Cardin and Chuck Schumer, who opposed Obama's Iran deal may now feel that as opponents of the Trump administration, they are required to oppose any change to the Iran Nuclear Agreement Review Act	What outcome would justify another U.S. war in a region where all the previous wars in this century have left us bleeding, bankrupt, divided and disillusioned?
FC	*Cersei Lannister. She subsequently appeared in A Clash of Kings (1998) and A Storm of Swords (2000). A Clash of Kings. A Clash of Kings is the second novel in A Song of Ice and Fire, an epic fantasy series by American author George R. R. Martin expected to consist of seven volumes. → Cersei Lannister appears in a series that was written by an author from the United States*	*David Bowie. During his lifetime, his record sales, estimated at 140 million worldwide, made him one of the world's best-selling music artists. → David Bowie only sold records in Jamaica*
RD	BREAKING: Three gunmen involved in attack on Charlie Hebdo magazine, French Interior Minister Bernard Cazeneuve says. http://t.co/ak9mTVfJdR @cnni the Islamic leaders should do something about the image of Islam by speaking out against the terrorists @cnni expel Muslims from European soil and destroy all the mosques. @cnni it's not the religion. But how the people interpret the writings and that's what causes them to do bad things. @cnni terrorism needs concerted efforts from every citizen to fight it,religion is going beyond boundaries if it can cause terror attacks	Reports: #CharlieHebdo suspects killed http://t.co/rsl4203bcQ Damn, this is like a movie RT @HuffingtonPost Reports: #CharlieHebdo suspects killed http://t.co/zCuZD1cure ?@HuffingtonPost: Reports: #CharlieHebdo suspects killed http://t.co/mWCSjh3CkH? superb simultaneous response by the French tactics unit. @HuffingtonPost great news! No trial, no taxpayer money spent to support them. @HuffingtonPost Good news !!! Alah Akbar !! @HuffingtonPost damnit!!! That's what those fuckers wanted!! Now they will be hailed as martyrs. . .. @HuffingtonPost Can you confirm the reports that those suspects were killed by French police? (. . .)

4.3 FC: fact checking

Fact checking is the most advanced way human experts can verify credibility of a given text: by assessing the veracity of the claims it includes with respect to a knowledge base (drawing from memory, reliable sources and common sense). Implementing this workflow in AI systems as computational fact checking (Graves 2018) is a promising direction for credibility assessment. However, it involves many challenges—choosing check-worthy statements (Nakov *et al.* 2022), finding reliable sources (Przybyła *et al.* 2022), extracting relevant passages (Karpukhin *et al.* 2020) etc. Here we focus on the claim verification stage. The input of the task is a pair of texts—target

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

claim and relevant evidence—and the output label indicates whether the evidence supports the claim or refutes it. It essentially is Natural Language Inference (NLI) (MacCartney 2009) in the domain of encyclopaedic knowledge and newsworthy events.

We use the data[i] from FEVER shared task (Thorne *et al.* 2018a), aimed to evaluate fact-checking solutions through a manually created set of evidence-claim pairs. Each pair connects a one-sentence claim with a set of sentences from Wikipedia articles, including a label of SUPPORTS (the evidence justifies the claim), REFUTES (the evidence demonstrates the claim to be false) or NOT ENOUGH INFO (the evidence is not sufficient to verify the claim). For the purpose of BODEGA, we take the claims from the first two categories,[j] concatenating all the evidence text.[k] The labels for the test set are not openly available, so we use the development set in this role.

See the examples in the third row of Table 2: the credible instance, where combined evidence from two articles (titles underlined) supports the claim (after the arrow); and non-credible one, where the evidence refutes the claim.

4.4 RD: rumour detection

A rumour is an information spreading between people despite not having a reliable source. In the online misinformation context, the term is used to refer to content shared between users of social media that comes from an unreliable origin, for example an anonymous account. Not every rumour is untrue as some of them can be later confirmed by established sources. Rumours can be detected by a variety of signals (Al-Sarem *et al.* 2019), but here we focus on the textual content of the original post and follow-ups from other social media users.

In BODEGA we use the Augmented dataset of rumours and non-rumours for rumour detection (Han, Gao, and Ciravegna, 2019), created from Twitter threads relevant to six real-world events (2013 Boston marathon bombings, 2014 Ottawa shooting, 2014 Sydney siege, 2015 Charlie Hebdo Attack, 2014 Ferguson unrest, 2015 Germanwings plane crash). The authors of the dataset started with the core threads annotated manually as rumours and non-rumours, then automatically augmented them with other threads based on textual similarity. We followed this by converting each thread to a flat feed of concatenated text fragments, including the initial post and subsequent responses. We set aside one of the events (Charlie Hebdo attack) for attack and development subsets, while others are included in the training subset.

See the last row of Table 2 for examples, both regarding the *Charlie Hebdo* shooting, but only the credible one is based on information from a credible source.

5. Attack scenario

The adversarial attack scenarios are often classified according to what information is available to the attacker. The *black-box* scenarios assume that no information is given on the inner workings of the targeted model and only system outputs for a given input can be observed. In *white-box* scenarios, the model is openly available to the attacker, allowing them to observe its internal structure and understand how predictions are made.

We argue neither of these scenarios is realistic in the practical misinformation detection setting, for example a content filter deployed in a social network. We cannot assume a model is available to the attacker since such information is usually not shared publicly; moreover, the model likely gets updated often to keep up with the current topics. On the other hand, the black-box scenario is too restrictive, as it assumes no information about the model is ever revealed. Also, once a certain design approach is popularised as the best performing in the NLP community, it tends to

[i] https://fever.ai/dataset/fever.html

[j] NOT ENOUGH INFO was excluded to cast the task as binary classification, in line with the other ones.

[k] Including the titles, which are often an essential part of the context in case of encyclopaedic articles.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

be applied to very many, if not most, solutions to related problems (Church and Kordoni 2022)—this is especially noticeable in case of large language models, such as BERT (Devlin *et al.* 2018) or GPT (Radford *et al.* 2018) and their successors.

For these reasons, in BODEGA we use the *grey-box* approach. The following information is considered available to an attacker preparing AEs:

- A "hidden" classifier f that for any arbitrary input returns $f(x) \in \{0, 1\}$ and a likelihood score $s_f(x)$, that is a numerical representation on how likely a given example x is to be assigned a positive class. This information is more helpful to attackers than only $f(x)$, which is typically set by applying a threshold t_f, for example $f(x) = 1 \iff s_f(x) > t_f$. The threshold expresses the minimum value of the score necessary for the classifier to assign a positive label to the instance. Typically, this value is set to 0.5.
- The general description of an architecture of classifier f, for example "a BERT encoder followed by a dense layer and softmax normalisation."
- The training X_{train}, the development X_{dev}, and the evaluation X_{attack} subsets.

This setup allows users of BODEGA to exploit weaknesses of classifiers without using the complete knowledge of the model, while maintaining some resemblance of practical scenarios.

Note that the grey-box setup is significantly more challenging to attack compared to the white-box scenario. In the latter, the attacker can directly see how the input features affect the output decision and modify those with the highest influence. Mathematically, this approach can be expressed in terms of computing a gradient of the decision variable and following it—thus the gradient-based methods (Zhang *et al.* 2020b). However, this is not possible to do in grey-box approach, where internal model weights, necessary for such procedure, are not revealed.

Another choice that needs to be made concerns the goal of the attacker. Generally, adversarial actions are divided into *untargeted* attacks, where any change in the victim's predictions is considered a success and *targeted* attacks, which seek to obtain a specific response, aligned with the attacker's goals (Zhang *et al.* 2020b).

Consider a classifier f that for a given instance x_i, with true value y_i, outputs class $f(x_i)$, which may be correct or incorrect. An *untargeted* attack involves perturbing x_i into x_i^*, such that $f(x_i) \neq f(x_i^*)$. A successful attack would undoubtedly show the brittleness of the classifier, but may not be necessarily helpful for a malicious user, for example if y_i corresponded to malicious content, but the original response $f(x_i)$ was incorrect.

Taking into account the misinformation scenario, we consider the *targeted* attack to satisfy the following criteria:

- The true class corresponds to non-credible content, that is $y_i = 1$,
- The original classifier response was correct, that is $f(x_i) = y_i$.

Success in this attack corresponds to a scenario of the attacker preparing a piece of non-credible content that is falsely recognised as credible thanks to the adversarial modification. We therefore use only a portion of the evaluation X_{attack} subset for this kind of attack.

By *non-credible content,* we mean:

- In case of hyperpartisan news, an article from a hyperpartisan source,
- In case of propaganda recognition, a sentence with a propaganda technique,
- In case of fact checking, a statement refuted by the provided evidence,
- In case of rumour detection, a message feed starting from a post including a rumour.

In BODEGA, both untargeted and targeted attacks can be evaluated.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

All of the text forming an instance can be modified to make an adversarial attack. In case of fact checking, this includes both the claim and the evidence. Similarly for rumour detection, not only the original rumour but also any of the follow-up messages in the thread are included in the text instance. This corresponds to the real-life scenario, where all of the above content is user-generated and can to some degree be influenced by an attacker (see further discussion on this matter in Section 10.1).

Finally, note that BODEGA imposes no restriction on the number of queries sent to the victim, that is the number of variants an attacker is allowed to test for each instance before providing the final modification. This number would typically be limited, especially in a security-oriented application (Chen *et al.* 2022). However, the constraints might be very different depending on a particular application scenarios. Some services might impose very strict limits on a number of submissions a client can make within a specified time, while others might allow many more attempts. If an attacker knows the data the victim classifier was trained on, they can even train a surrogate classifier and issue as many queries as needed. Thus, in order to provide a comprehensive evaluation, the number of queries is not limited in BODEGA, but it is recorded as an evaluation metric (see the next section).

6. Evaluation

Preparing adversarial examples involves balancing two goals in the adversarial attack (see Fig. 1):

1. Maximising $\text{diff}(f(x_i), f(x_i^*))$—difference between the classes predicted by the classifier for the original and perturbed instance,
2. Maximising $\text{sim}(x_i, x_i^*)$—similarity between the original and perturbed instance.

If (1) is too small, the attack has failed, since the classifier preserved the correct prediction. If (2) is too small, the attack has failed, since the necessary perturbation was so large it defeated the original purpose of the text.

This makes the evaluation multi-criterion and challenging since neither of these factors measured in isolation reflects the quality of AEs. The conundrum is usually resolved by setting the minimum similarity (2) to a fixed threshold (known as *perturbation constraint*) and measuring the reduction in classification performance, that is accuracy reduction (Zhang *et al.* 2020b). This can be problematic as there are no easy ways to decide the value of the threshold that will guarantee that the class remains valid. The issue is especially relevant for a task as subtle as credibility analysis—for example how many word swaps can we do on a real news piece before it loses credibility?

In BODEGA, we avoid this problem by inverting the approach. Instead of imposing constraints on goal (2) and using (1) as the evaluation measure, we impose constraints on (1) and use (2) for evaluation. Specifically, we only count the instances when the modification was sufficient to change the classifier's decision (1) and treat text similarity (2) as the quality evaluation measure.

We define an adversarial modification quality score, called *BODEGA score*. BODEGA score always lies within 0-1 and a high value indicates good-quality modification preserving the original meaning (with score = 1 corresponding to no visible change), while low value indicates poor modification, altering the meaning (with score = 0 corresponding to completely different text).

In the remainder of this section, we discuss the similarity measurement techniques we employ and outline how they are combined to form a final measure of attack success.

6.1 Semantic score

The first element used to measure meaning preservation is based on *BLEURT* (Sellam, Das, and Parikh 2020). BLEURT was designed to compute the similarity between a candidate and reference

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

sentences in evaluating solutions for natural language generation tasks (e.g. machine translation). The underlying model is trained to return values between 1 (identical text) and 0 (no similarity).

BLEURT helps to properly assess semantic similarity; for example, replacing a single word with its close synonym will yield high score value, while using a completely different one will not. However, BLEURT is trained to interpret multi-word modifications (i.e. paraphrases) as well, leading to better correlation with human judgement than other popular measures, for example BLEU or BERTScore. This is possible thanks to fine-tuning using synthetic data covering various types of semantic differences, for example *contradiction* as understood in the NLI (Natural Language Inference) task. This is especially important for our usecase, helping to properly handle the situations where otherwise small modifications completely change the meaning of the text (e.g. a negation), rendering an AE unusable.

In BODEGA, we use the pyTorch implementation of BLEURT,[l] choosing the recommended[m] `BLEURT-20` variant. Since the score is only *calibrated* to the 0-1 range, other numbers can be produced as well. Thus, our semantic score is equal to BLEURT (clipped to 0-1 if necessary). Finally, since BLEURT is a sentence-level measure and our tasks involve longer text fragments,[n] we (1) split the text into sentences[o] using LAMBO (Przybyła, 2022), (2) find the pairs of sentences from the original and modified text that are most similar using Levenshtein distance and (3) compute semantic similarities between sentence pairs, returning its average as semantic score.

6.2 Character score

Levenshtein distance is used to express how different one string of characters is from another. Specifically, it computes the minimum number of elementary modifications (character additions, removals, replacements) it would take to transform one sequence into another (Levenshtein 1966).

Levenshtein is a simple measure that does not take into account the meaning of the words. However, it is helpful to properly assess modifications that rely on graphical resemblance. For example, one family of adversarial attacks relies on replacing individual characters in text (e.g. *call* to *ca||*), altering the attacked classifier's output. The low value of Levenshtein distance in this case represents the fact that such modification may be imperceptible for a human reader.

In order to turn Levenshtein distance $lev\_dist(a, b)$ into a character similarity score, we compute the following:

$$\text{Char\_score}(a, b) = 1 - \frac{lev\_dist(a, b)}{\max(|a|, |b|)}$$

Char_score is between 0 and 1, with higher values corresponding to larger similarity, with $\text{Char\_score}(a, b) = 1$ if a and b are the same and $\text{Char\_score}(a, b) = 0$ if they have no common characters at all.

6.3 BODEGA score

The BODEGA score for a pair of original text x_i and modified text x_i^* is defined as follows:

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

where $\text{Sem\_score}(x_i, x_i^*)$ is semantic score; $\text{Char\_score}(x_i, x_i^*)$ is character score; and $\text{Con\_score}(x_i, x_i^*)$ is confusion score, which takes value 1 when an adversarial example is produced and succeeds in changing the victim's decision (i.e. $f(x_i) \neq f(x_i^*)$) and 0 otherwise.

[l] https://github.com/lucadiliello/bleurt-pytorch
[m] https://github.com/google-research/bleurt
[n] Except propaganda detection, where input is a single sentence.
[o] Except fact-checking, where we simply split evidence from the claim.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

The overall attack success measure is computed as an average over BODEGA scores for all instances in the attack set available in a given scenario (targeted or untargeted). The success measure reaches 0 when the AEs bear no similarity to the originals, or they were not created at all. The value of 1 corresponds to the situation, unachievable in practice, when AEs change the victim model's output with immeasurably small perturbation.

Many adversarial attack methods include tokenisation that does not preserve the word case or spacing between them. Our implementation of the scoring disregards such discrepancies between input and output, as they are not part of the intended adversarial modifications.

Apart from BODEGA score, expressing the overall success, the intermediate measures can paint a fuller picture of the strengths and weaknesses of a particular solution:

- Confusion score—in how many of the test cases the victim's decision was changed,
- Semantic score—an average over the cases with changed decision,
- Character score—an average over the cases with changed decision.

We also report the number of queries made to the victim, averaged over all instances.

7. Victim classifiers

A victim classifier is necessary to perform an evaluation of an AE generation solution. We include implementations of text classifier based on various common architectures: a recurrent neural network (BiLSTM) trained from scratch; and fine-tuned language models: small masked model (BERT), large generative model (GEMMA2B) and a very large generative model (GEMMA7B), delivering state-of-the-art results in the established benchmarks.

This component of BODEGA could be easily replaced by newer implementations, either to test a robustness of a specific classifier architecture or to have a better understanding of applicability of a given AE generation solution.

7.1 BiLSTM

The recurrent network is implemented using the following layers:

- An embedding layer, representing each token as vector of length 32,
- Two LSTM (Hochreiter and Schmidhuber 1997) layers (forwards and backwards), using hidden representation of length 128, returned from the edge cells and concatenated as document representation of length 256,
- A dense linear layer, computing two scores representing the two classes, normalised to probabilities through softmax.

The input is tokenised using BERT uncased tokeniser (see below). The maximum allowed input length is 512, with padding as necessary. For each of the tasks, a model instance is trained from scratch for 10 epochs by using Adam optimiser (Kingma and Ba 2015), a learning rate of 0.001 and batches of 32 examples each. The implementation uses *PyTorch*.

7.2 BERT

As a baseline pretrained language model, we use BERT in the base variant (Devlin *et al.* 2018). The model is fine-tuned for sequence classification using Adam optimiser with linear weight decay (Loshchilov and Hutter 2019), starting from 0.00005, for 5 epochs. We use maximum input length of 512 characters and a batch size of 16. The training is implemented using the *Hugging Face Transformers* library (Wolf *et al.* 2020) (`bert-base-uncased` model).

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

7.3 Gemma

In order to assess the vulnerability of the large language models to AEs, we include *Gemma* (Gemma Team and Google DeepMind 2024). Gemma is a recent generative language model, derived from Google's *Gemini* models and following the same design principles as the GPT family (Radford *et al.* 2018). We include both the smaller variant with 2 billion parameters, as well as the full 7-billion model, loaded through *Hugging Face Transformers*. They have been evaluated in multiple benchmarks and the latter has shown the best performance among the openly available large language models (Gemma Team and Google DeepMind 2024).

The fine-tuning was performed using the same procedure as for BERT. However, in order to keep the computing requirements under control, we applied parameter-efficient fine-tuning (Lialin, Deshpande, and Rumshisky 2023). Namely, we used QLoRA optimisation (Dettmers *et al.* 2023), based on Low Rank Adaptation (LoRA) (Hu *et al.* 2021) with reduced numerical precision. These are implemented using the *Hugging Face*'s libraries `peft` and `bitsandbytes`, respectively.

8. AE generation solutions

Within BODEGA, we include the AE generation solutions implemented in the *OpenAttack* framework. We exclude the approaches for white-box scenario (gradient-based) and those that yielded poor performance in preliminary tests. We test 8 approaches:

- **BAE** (Garg and Ramakrishnan 2020) uses BERT (Devlin *et al.* 2018) as a masked language model to generate word candidates that are likely in a given context. This includes both replacing existing tokens as well as inserting new ones.
- **BERT-ATTACK** (Li *et al.* 2020) is a very similar approach, which starts with finding out if a word is vulnerable by checking victim's response to its masking. The chosen words are replaced using BERT candidates, but unlike in BAE, no new words are inserted.
- **DeepWordBug** (Gao *et al.* 2018) works at the character level, seeking modifications that are barely perceptible for humans, but will modify an important word into one unknown to the attacked model. The options include character substitutions, removal, insertion and reordering.
- **Genetic** (Alzantot *et al.* 2018) is using the genetic algorithm framework. A *population* includes variants of text built by word replacements (using GloVe representation to ensure meaning preservation), the most promising of which can replicate and combine until a successful AE is found.
- **SememePSO** (Zang *et al.* 2020) employs a related framework, namely Particle Swarm Optimisation (PSO). A group of *particles*, each representing a text modification with a certain probability of further changes (*velocity*), moves through the feature space until an optimal position is found.
- **PWWS** (Ren *et al.* 2019) is a classical greedy word replacement approach. However, it differs from the majority of the solutions by using *WordNet*, instead of vector representations, to obtain synonym candidates.
- **SCPN** (Iyyer *et al.* 2018) performs paraphrasing of the whole text through a bespoke encoder-decoder model. In order to train this model, the authors generate a dataset of paraphrases through backtranslation from English to Czech.
- **TextFooler** (Jin *et al.* 2020) is a greedy word-substitution solution. Unlike other similar approaches, it takes into account the syntax of the attacked text, making sure the replacement is a valid word that agrees with the original regarding its part of speech. This helps to make sure the AE is fluent and grammatically correct.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 3. Performance of the victim classifiers, expressed as F-score over the attack subset

	BiLSTM	BERT	GEMMA2B	GEMMA7B
HN	0.7076	0.7544	**0.7792**	0.7603
PR	0.4857	0.6410	0.6271	**0.6840**
FC	0.7532	0.9360	0.9701	**0.9727**
RD	0.6234	0.7547	**0.7609**	0.7229
Parameters	1 M	340 M	2B	7B

The main problem the presented solutions try to solve is essentially maximising a goal function (victim's decision) in a vast space of possible modifications to input text, which is further complicated by its discrete nature. Direct optimisation is not computationally feasible here, giving way to methods that are greedy (performing the change that improves the goal the most) or maintain a population of varied candidate solutions (PSO and evolutionary algorithms). The majority of the solutions operate on word level, seeking replacements that would influence the classification result without modifying the meaning. The exceptions are sentence-level SCPN, performing paraphrasing of entire sentences, and character-level DeepWordBug, replacing individual characters in text to preserve superficial similarity to the original. They all use victims' scores to look for most promising modifications, except for SCPN, which operates blindly, simply generating numerous possible paraphrases.

All of the attackers are executed with their default functionality, except for BERT-ATTACK, that we use without the generation of subword permutations, which is prohibitively slow for longer documents. Just like the victim classifier, the AE solution interface in BODEGA allows for new solutions to be added and tested as the field progresses.

8.1 Classification performance

Table 3 shows the performance of the victim classifiers, computed as F-score over the test data (combined development and attack subsets). As expected, BERT easily outperforms a neural network trained from scratch. The credibility assessment tasks are subtle and the amount of data available for training severely limits the performance. Thus, the BERT model has an advantage by relying on knowledge gathered during pretraining. This is demonstrated by the performance gap being the largest for the dataset with the least data available (propaganda detection) and the smallest for the most abundant corpus (hyperpartisan news). The Gemma models perform even better than BERT in all tasks. However, the improvement is not as spectacular (a few per cent) and GEMMA7B does not provide uniformly better results than the 2 billion model.

9. Experiments

The purpose of the experiments is to test the BODEGA framework in action and improve our understanding of the vulnerability of content-filtering solutions to adversarial actions. This will also establish a baseline for systematic evaluation of future classifiers and AE generators. To that end, we test the attack performance for:

- four tasks (HN, PR, FC, RD),
- eight attackers (BAE, BERT-ATTACK, DeepWordBug, Genetic, SememePSO, PWWS, SCPN, textFooler),

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

- four victims (BiLSTM, BERT, GEMMA2B, GEMMA7B),
- two scenarios (untargeted and targeted).

In total, $4 \times 8 \times 4 \times 2 = 256$ experiments are performed, each evaluated using the measures introduced in Section 6.

The full results are shown in the appendix. Here we present an analysis focused on key questions:

- Q1: Which attack method delivers the best performance?
- Q2: Are the modern large language models less vulnerable to attacks than their predecessors?
- Q3: How many queries are needed to find adversarial examples?
- Q4: Does targeting make a difference in attack difficulty?

Moreover, we perform a manual analysis of the most promising AEs (Section 9.5).

9.1 Q1: attack methods

Table 4 compares the performance of the untargeted attack methods in various tasks, averaged over victim models.

The hyperpartisan news detection task is relatively easy for generating AEs. BERT-ATTACK achieves the best BODEGA score of 0.56, which is possible due to changing the decision on 90 per cent of the instances while preserving high similarity, both in terms of semantics and characters. However, DeepWordBug (a character-level method) provides the best results in terms of semantic similarity, changing less than 1 per cent of characters on average. The only drawback of this method is that it works in 25 per cent of the cases, failing to change the victim's decision in the remaining ones.

The propaganda recognition task significantly differs from the previous task in terms of text length, including individual sentences rather than full articles. As a result, every word is more important and it becomes much harder to make the changes imperceptible, resulting in lower character similarity scores. This setup appears to favour the Genetic method, obtaining the best BODEGA score: 0.49. This approach performs well across the board, but it comes at a high cost in terms of model queries. Even for the short sentences in propaganda recognition, a victim model is queried over 800 times, compared to less than 150 for all other methods.

Fact checking resembles the propaganda recognition in terms of relatively short text fragments, but the best-performing method is BERT-ATTACK. As for hyperpartisan news, DeepWordBug achieves high similarity, but succeeds in finding an AE relatively rarely—26 per cent of times.

Finally, the rumour detection task in the untargeted scenario appears to be the hardest problem to attack. Here the best methods reach BODEGA score of 0.25, indicating low usability, mostly due to low confusion rates—barely above 60 per cent . This may be because rumour threads consist of numerous posts, each having some indication on the credibility of the news, forcing an attacker to make many modifications to change the victim's decision. The text of Twitter messages is also far from regular language, making the challenge harder for methods using models pretrained on well-formed text (e.g. BERT-ATTACK). It has to be noted however that this setup is equally problematic to the meaning preservation measurement (semantic score), thus suggesting these results should be taken cautiously.

Regarding the performance of the included attack methods, we can observe the following:

- Approaches relying on local changes (e.g. BERT-ATTACK, DeepWordBug) work better than global rephrasers (SCPN), because they are able to deliver more candidates for AEs and thus have more chances for success.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 4. The results of adversarial attacks, averaged over all victim classifiers, in four misinformation detection tasks (untargeted). Evaluation measures include BODEGA score, confusion score, semantic score, character score and number of queries to the attacked model per example. The best score in each task is in boldface

Task	Method	BODEGA	Confusion	Semantic	Character	Queries
HN	BAE	0.36	0.60	0.61	0.97	589.39
	BERT-ATTACK	**0.56**	**0.90**	0.63	0.97	910.00
	DeepWordBug	0.25	0.33	**0.78**	**1.00**	390.95
	Genetic	0.38	0.81	0.48	0.98	1740.55
	SememePSO	0.19	0.40	0.50	0.99	309.10
	PWWS	0.37	0.79	0.48	0.98	2051.62
	SCPN	0.00	0.82	0.09	0.02	11.75
	TextFooler	0.34	0.77	0.46	0.96	792.91
PR	BAE	0.14	0.21	0.71	0.94	33.31
	BERT-ATTACK	0.46	0.72	0.69	0.91	76.40
	DeepWordBug	0.20	0.26	**0.79**	**0.96**	27.33
	Genetic	**0.49**	**0.84**	0.65	0.89	886.55
	SememePSO	0.41	0.68	0.67	0.89	99.51
	PWWS	0.46	0.74	0.67	0.90	132.20
	SCPN	0.11	0.54	0.38	0.48	11.54
	TextFooler	0.41	0.72	0.65	0.87	62.26
FC	BAE	0.35	0.53	0.69	0.96	78.58
	BERT-ATTACK	**0.57**	**0.83**	0.72	0.95	153.37
	DeepWordBug	0.26	0.31	**0.83**	**0.98**	54.10
	Genetic	0.52	0.77	0.70	0.95	846.25
	SememePSO	0.44	0.65	0.71	0.96	145.06
	PWWS	0.48	0.69	0.72	0.96	225.98
	SCPN	0.07	0.66	0.30	0.33	11.66
	TextFooler	0.46	0.70	0.70	0.94	109.77
RD	BAE	0.10	0.24	0.42	0.98	310.71
	BERT-ATTACK	**0.25**	**0.62**	0.42	0.94	860.04
	DeepWordBug	0.13	0.19	**0.70**	**0.99**	235.85
	Genetic	0.24	0.53	0.46	0.96	2605.13
	SememePSO	0.12	0.26	0.47	0.97	330.20
	PWWS	0.21	0.46	0.46	0.96	1107.09
	SCPN	0.01	0.41	0.17	0.11	11.40
	TextFooler	0.18	0.46	0.44	0.92	654.20

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

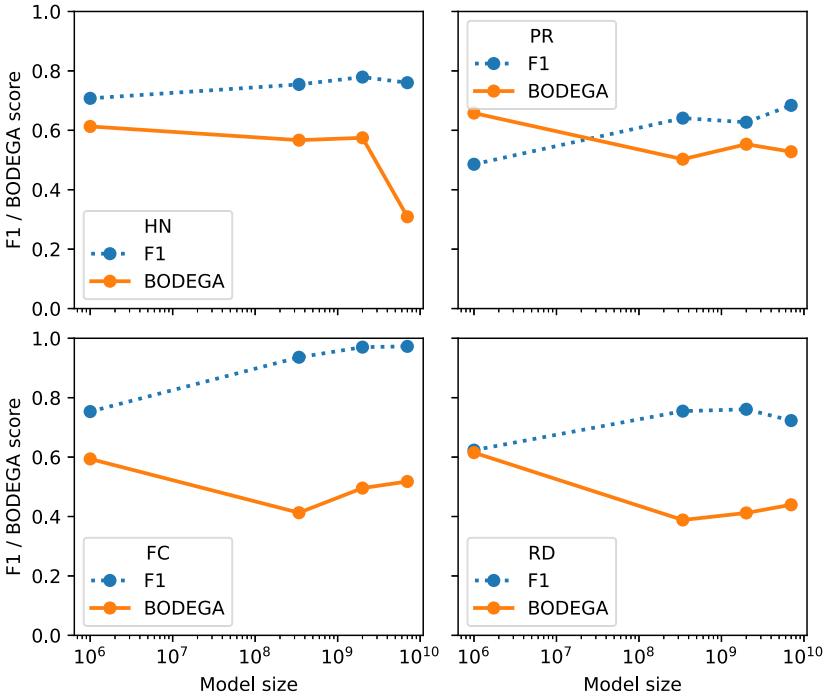

Figure 2. Classification performance (F1 score) and vulnerability to targeted attacks (BODEGA score) of models according to their size (parameter count, logarithmic scale), for different tasks.

- Character-replacing solutions (e.g. DeepWordBug) maintain high similarity, both in semantic and Levenstain measures, but suffer in terms of confusion rate. Clearly, sometimes changing a whole word is necessary to trigger a decision change.
- Methods relying on language models for meaning representation (esp. BERT-ATTACK) obtain better results than those relying on GloVe (Genetic) or WordNet (PWWS). This is likely because the older methods are not context-sensitive, resulting in less appropriate replacements, visible as reduced semantic scores.
- Solutions performing a very extensive search (esp. Genetic) find good AEs only for short text: propaganda and fact-checking. They become unfeasible for longer content, for example news.
- Even solutions with apparently similar designs (BAE and BERT-ATTACK) can deliver vastly different performance due to smaller details in their implementation.

9.2 Q2: victim size and vulnerability

Fig. 2 plots the performance and vulnerability to targeted attacks (BODEGA score of the most successful method) of models of increasing size: BiLSTM, BERT, GEMMA2B, GEMMA7B. We can see that while the classification scores almost universally improve with larger models (albeit with diminishing returns), the robustness assessment paints a more complex picture.

BiLSTM, which is by far the smallest model, is also clearly the most vulnerable to attacks. However, the results for the large pretrained models are surprising: the smallest of them (BERT) appears to be the most robust, except for one task (HN). This effect is the strongest for the FC task, where the best attacker on the GEMMA7B model achieves a score 27% higher than in the attack against BERT. For two of the tasks (FC and RD), this pattern holds even within the same model family, with the smaller GEMMA model showing lower vulnerability.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

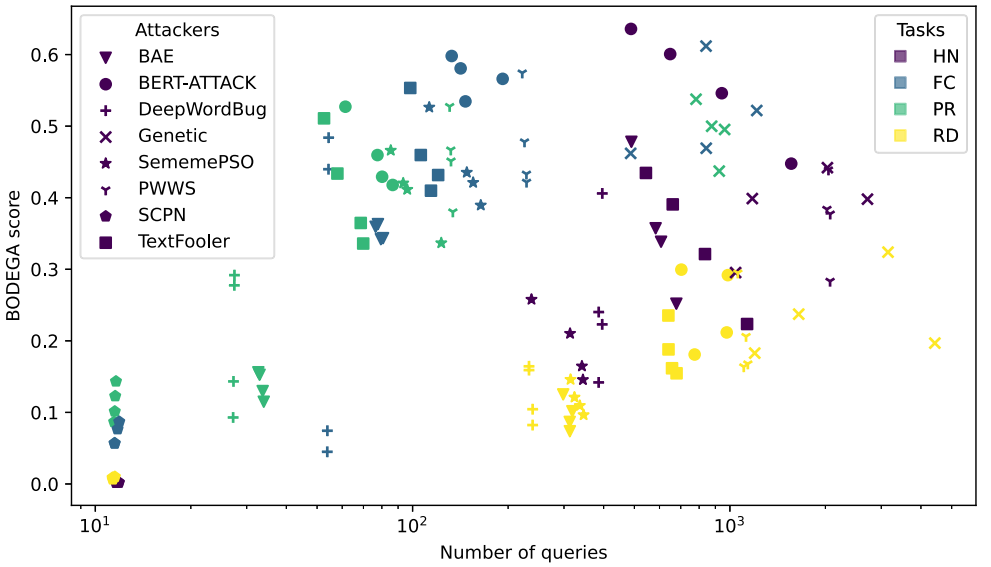

Figure 3. Results of the targeted attacks (y axis, BODEGA score) plotted against the number of queries necessary (x axis, logarithmic) for various attack methods (symbols) and tasks (colours).

Overall, new and more accurate language models are not less vulnerable to attacks, as one would hope. In the application scenarios involving adversarial actors, such as credibility assessment, smaller solutions may thus be a more appropriate choice. This observation is a contribution to the wider question of vulnerability of LLMs to adversarial actions (Yao *et al.* 2024; Goto, Ono, and Morita 2024). While this is a new research area, preliminary results are concordant with ours, namely showing larger models as not necessarily increasing robustness over the smaller predecessors (Liu *et al.* 2024). Our results do not explain why the robustness does not increase with model size as classification performance does, and we leave this problem as an interesting question for future research.

9.3 Q3: number of queries

Fig. 3 illustrates the number of queries necessary to perform attacks with various levels of success. Primarily, we can see the results are grouped according to the task being attacked. The tasks involving long text (HN and RD) both require many queries: for each attacked example, from several hundred to several thousand attempts are needed to find an adversarial variant. These two tasks differ in terms of success, with hyperpartisan news obtaining some of the highest BODEGA scores and rumour detection: the lowest. The tasks involving shorter text (FC and PR) have similarly high success rate, but good attacks require much less queries: from just over 100 (FC) to less than 60 (PR).

In terms of attack methods, BERT-ATTACK clearly achieves the best BODEGA score for most tasks. However, it requires many queries—even though not as many as the Genetic approach. Among the methods that work with less queries, often with little cost in terms of performance loss, we can distinguish TextFooler and DeepWordBug.

9.4 Q4: targeting

Table 5 compares targeted and untargeted scenarios in terms of performance—the best BODEGA score and the number of queries needed to achieve it. Interestingly, the individual score differences

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 5. A comparison of the results—highest BODEGA score and corresponding number of queries—in the untargeted (U) and targeted (T) scenario for various tasks and victims. The better values (higher BODEGA scores and lower number of queries) are highlighted

		BiLSTM		BERT		GEMMA2B		GEMMA7B	
		U	T	U	T	U	T	U	T
HN	B. score	**0.64**	0.61	**0.60**	0.57	0.55	**0.57**	**0.45**	0.31
	Queries	**487.85**	565.05	**648.41**	753.91	942.98	**761.53**	**1560.76**	2313.33
PR	B. score	0.54	**0.66**	0.50	**0.50**	0.50	**0.55**	0.44	**0.53**
	Queries	782.15	**50.14**	962.40	**99.95**	876.06	**94.05**	925.58	**110.32**
FC	B. score	**0.61**	0.59	**0.53**	0.41	**0.57**	0.50	**0.58**	0.52
	Queries	840.99	**123.24**	**146.73**	207.23	**192.25**	254.22	**141.70**	173.86
RD	B. score	0.32	**0.62**	0.20	**0.39**	0.30	**0.41**	0.21	**0.44**
	Queries	3150.24	**153.61**	4425.11	**174.03**	**703.07**	1108.21	977.27	**202.18**

can be quite high, but the pattern depends on the classification task. The targeted task is always harder for news bias assessment (except BERT) and fact checking. The untargeted one is always much more challenging for propaganda recognition and rumour detection.

9.5 Manual analysis

In order to better understand how a successful attack might look like, we manually analyse some of them. This allows us observe what types of adversarial modifications are the weakest point of the classifier, as well as verify if attack success scoring using automatic measures is aligned with the human judgement.

For that purpose, we select 20 instances with the highest BODEGA score from the untargeted interactions between a relatively strong attacker (BERT-ATTACK) and a relatively weak victim (BiLSTM), within all tasks. Next, we label the AEs according to the degree they differ from the original text:[P]

1. **Synonymous**: the text is identical in meaning to the original.
2. **Typographic**: change of individual characters, for example resembling sloppy punctuation or typos, likely imperceptible.
3. **Grammatical**: change of the syntax of the sentence, for example replacing a verb with a noun with the same root, possibly making the text grammatically incorrect,
4. **Semantic-small**: changes affecting the overall meaning of the text, but to a limited degree, unlikely to affect the credibility label,
5. **Semantic-large**: significant changes in the meaning of the text, indicating the original credibility may not apply,
6. **Local**: changes of any degree higher than Synonymous, but present only in a few non-crucial sentences of a longer text, leaving others to carry the original meaning (applies to tasks with many sentences, i.e. RD and HN).

[P]Note that while these categories might overlap, e.g. a typographic replacement significantly affecting the overall meaning, such cases were not encountered in practice during the analysis.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 6. Number of AEs using different modifications among the best 20 instances (according to BODEGA score) in each task, using BiLSTM as victim and BERT-ATTACK as attacker

	Number of instances				
AE degree	HN	PR	FC	RD	% of all
Synonymous	6	10	2	5	29%
Typographic	0	5	8	0	16%
Grammatical	0	4	3	2	11%
Semantic-small	0	1	2	3	7%
Local	13	-	-	2	19%
Semantic-large	1	0	5	8	17%

The changes labelled as Semantic-large indicate attack failure, while others denote success with varying visibility of the modification.

Table 6 shows the quantitative results of the manual analysis, while Table 7 includes some examples. Generally, a large majority of these attacks (82.5 per cent) were successful in maintaining the original meaning, confirming the high BODEGA score assigned to them. However, significant differences between the tasks are visible.

Consistently with the results of automatic analysis, rumour detection appears to be the most robust, resulting in many attacks changing the original meaning. Even though oftentimes only a word or two is changed, it affects the meaning of the whole Twitter thread, since the follow-up messages do not repeat the content, but often deviate from the topic (see EX4 in Table 7). The opposite happens for hyperpartisan news: a singular change does not affect the overall message, as the news article are typically redundant and maintain their sentiment throughout (see EX6). As a result, the HR task is one of the most vulnerable to attacks.

It is also interesting to compare the two tasks with shorter text: fact checking and propaganda recognition. While the FC classifier shows a large vulnerability to typographic changes (esp. in punctuation, see EX2), many of the changes performed by the attackers affect important aspects of the content (e.g. names or numbers, see EX5), making the AE futile. The propaganda recognition, on the other hand, appears to rely on stylistic features, allowing the AE generation while preserving full synonymy (see EX1) or just introducing grammatical issues (see EX3).

10. Discussion

10.1 Reality check for credibility assessment

While one of the principles guiding the design of BODEGA has been a realistic simulation of the misinformation detection scenarios, this is possible only to an extent. Among the obstacles are low transparency of content management platforms (Gorwa, Binns, and Katzenbach 2020) and the vigorous growth of the methods of attack and defence in the NLP field.

Firstly, we have included only four victim models in our tests: BiLSTM, BERT and two Gemma variants, while in reality dozens of architectures for text classification are presented at every NLP conference, with a significant share specifically devoted to credibility assessment. However, the field has recently become surprisingly homogeneous, with the ambition to achieve the state-of-the-art pushing researchers to reuse the common pretrained language models in virtually every application (Church and Kordoni 2022). But these lookalike approaches share not only good

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Table 7. Some examples of adversarial modifications that were successful (i.e. resulted in changed classifier decision), performed by BERT-ATTACK against BiLSTM, including identifier (mentions in text), task and type of modification. Changes are highlighted in boldface

Id., task, type	Original example	Adversarial example
EX1 PR Synonymous	Puerto Rico's housing secretary, Fernando Gil, says the number of **homes** destroyed by the hurricane totals about 70,000 so far, and homes with major damage have amounted to 250,000 across the island	Puerto Rico's housing secretary, Fernando Gil, says the number of **houses** destroyed by the hurricane totals about 70,000 so far, and homes with major damage have amounted to 250,000 across the island
EX2 FC Typographic	Sabbir Khan. Sabbir's second movie, Heropanti starring Tiger Shroff & Kriti Sanon**,** released on 23 May 2014. → Sabbir Khan directed a movie	Sabbir Khan. Sabbir's second movie, Heropanti starring Tiger Shroff & Kriti Sanon**?** released on 23 May 2014. → Sabbir Khan directed a movie
EX3 PR Grammatical	Fastiggi and Goldstein have managed to make the problem even worse in their attempt to **explain** it away	Fastiggi and Goldstein have managed to make the problem even worse in their attempt to **explained** it away
EX4 RD Semantic-small	A few of the best cartoons **drawn & shared in solidarity** with #charliehebdo after yesterday's massacre #jesuischarlie http://t.co/87et0xpnwr @theinquisitr war profiteers x'd #princessdiana & dodifayed in #paris. pushing #france to join war on terror video >> http://t.co/tysy8ys49w @theinquisitr l'amérique se tient avec la france. #jesuischarlie	A few of the best cartoons **contributed & held in friendship** with #charliehebdo after yesterday's massacre #jesuischarlie http://t.co/87et0xpnwr @theinquisitr war profiteers x'd #princessdiana & dodifayed in #paris. pushing #france to join war on terror video >> http://t.co/tysy8ys49w @theinquisitr l'amérique se tient avec la france. #jesuischarlie
EX5 FC Semantic-large	Hannah and Her Sisters. Hannah and Her Sisters is a 1986 American comedy—drama film which tells the intertwined stories of an extended family over two years that begins and ends with a family thanksgiving dinner. → Hannah and Her Sisters is an American **1986** film	Hannah and Her Sisters. Hannah and Her Sisters is a 1986 American comedy—drama film which tells the intertwined stories of an extended family over two years that begins and ends with a family thanksgiving dinner. → Hannah and Her Sisters is an American **1987** film
EX6 HN Local	Aleppo completely back under government control (GPA) Aleppo—the Syrian Arab Army (SAA) has reported today that the entirety of **east** Aleppo is fully back under government control, meaning the city is now completely liberated. The SAA has completed the evacuations of anti-government fighters and civilians looking to flee with these groups as of today. This is a major victory for the Syrian forces in Aleppo coming after almost 4 years of fighting in the city. Thousands of people have already taken to the streets to celebrate the last of the terrorists inside the city leaving. [347 words more]	Aleppo completely back under government control (GPA) Aleppo—the Syrian Arab Army (SAA) has reported today that the entirety of **south** Aleppo is fully back under government control, meaning the city is now completely liberated. The SAA has completed the evacuations of anti-government fighters and civilians looking to flee with these groups as of today. This is a major victory for the Syrian forces in Aleppo coming after almost 4 years of fighting in the city. Thousands of people have already taken to the streets to celebrate the last of the terrorists inside the city leaving. [347 words more]

performance but also weaknesses. Thus we expect that, for example, the results of attacks on fine-tuned BERT will also apply to other solutions that use BERT as a representation layer. Moreover, the current architecture of BODEGA supports binary text classification models only. This means it can be extended to other similar tasks with a binary label output, for example sentiment analysis or detecting machine-generated text. But it cannot be used to assess robustness of models for machine translation or other language generation tasks—these would require a different approach.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Secondly, we have re-used the attacks implemented in OpenAttack to have a comprehensive view of performance of different approaches. However, the field of AEs for NLP is relatively new, with the majority of publications emerging in the recent years, which makes it very likely that subsequent solutions will provide superior performance. With the creation of BODEGA as a universal evaluation framework, such comparisons become possible.

Thirdly, we need to consider the realism of evaluation measures. The AE evaluation framework assumes that if a modified text is very similar to the original, then the label (credible or not) still applies. Without this assumption, every evaluation would need to include manual re-annotation of the AEs. Fortunately, assessing semantic similarity between two fragments of text is a necessary component of evaluation in many other NLP tasks, for example machine translation (Lee *et al.* 2023), and we can draw from that work. Apart from BLEURT, we have experimented with SBERT cross-encoders (Thakur *et al.* 2021) and unsupervised BERT Score (Zhang *et al.* 2020a), but haven't found decisive evidence for the superiority of any approach. However, the problem remains open. The investigation on how subtle changes in text can invert its meaning and subvert credibility assessment is particularly vivid in the fact-checking field (Jaime *et al.* 2022), but it is less explored for tasks involving multi-sentence inputs, for example news credibility. An ideal measure of AE quality would take into account the characteristics of a text domain, assigning different impact to a given change depending on the nature of the text. This could be expressed by modifying the BODEGA score into a weighted score of the included factors and calibrating it by setting the weights for each text genre. However, to find the parameter values that accurately capture the human perception of acceptable changes, an annotation study would be necessary. We see this as a promising direction for future research. Moreover, the measures focusing on performance loss, for example computing the reduction in accuracy of the victim model under a specified modification might be worth investigating. However, an annotation study would be necessary as well, namely in order to establish the acceptable modification threshold for each task.

Fourthly, we also assume that an attacker has a certain level of access to the victim classifier, being able to send unlimited queries and receive numerical scores reflecting its confidence, rather than a final decision. In practice, this is currently not the case, with platforms revealing almost nothing regarding their automatic content moderation processes. However, this may change in future due to regulatory pressure from the government organisations; cf., for example, the recently agreed EU *Digital Services Act*.[q]

Finally, we need to examine how realistic is that an attacker could freely modify any text included in our tasks. While this is trivial in the case of hyperpartisan news and propaganda recognition, where the entire input comes from a malicious actor, the other tasks require closer consideration. In case of rumour detection, the text includes, apart from the initial information, replies from other social media users. These can indeed be manipulated by sending replies from anonymous accounts and this scenario has been already explored in the AE literature (Le *et al.* 2020). In the case of fact checking, the text includes, apart from the verified claim, also the relevant snippets from the knowledge base. However, it can be modified as well, when (as is usually the case) the knowledge is based on Wikipedia, which is often a subject of malicious alterations, from vandalism (Kiesel *et al.* 2017) to the generation of entire hoax articles (Kumar, West, and Leskovec 2016).

To sum up, we argue that despite certain assumptions, the setup of a BODEGA framework is close enough to real-life conditions to give insights about the robustness of popular classifiers in this scenario. BODEGA is already being used as a benchmark for new solutions that advance foundational AE generation methods tested here. Within the CheckThat! evaluation lab organised at CLEF 2024 (Barrón-Cedeño *et al.* 2024), focused on misinformation detection, Task 6 is devoted to measuring the robustness of credibility assessment. The evaluation of the AEs submitted by the

[q]https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

task participants is based on the framework described here, with certain expansions (Przybyła *et al*. 2024).[r]

10.2 Looking forward

We see this study as a step towards the directions recognised in the ML literature beyond NLP. For example, in security-oriented applications, there is the need to bring the evaluation of AEs closer to realistic conditions (Chen *et al.* 2022). Some limitations, esp. number of queries to the model, make attacks much harder. Even beyond the security field, assessing robustness is crucial for ML models that are distributed as massively-used products. This exposes them to unexpected examples, even if not generated with explicit adversarial motive. Individual spectacular failures are expected to be disproportionately influential on public opinion of technology, including AI (Mannes 2020), emphasising the importance of research on AEs.

Our work emphasises the need for taking into account the adversarial attacks when deploying text classifiers in adversarial scenarios, such as content filtering in social media. In many cases, changing just a few words in text can alter the decision of the models. We can recommend three ways to mitigate the associated risks.

Firstly, the vulnerability of ML models to adversarial examples indicates their output cannot be the only criterion in content-filtering systems. However, many AEs are quite transparent to humans, and the manipulation could be easily noticed. This suggests that the sensitive scenarios could benefit from a cooperation between a human operator and a ML model. For example, a system that uses ML models for prioritising work of human operators instead of making final decision is likely to be more robust than the ML model alone. Secondly, our work shows that the attack performance depends on the variety of factors, including dataset size, text length, victim architecture, etc. This makes it crucial to test every content-filtering solution before its deployment using real-world data and state-of-the-art attackers. Thirdly, taking into account adversarial environment in the classifier design, for example through adversarial training, can limit the amount of adversarial examples it is vulnerable to.

Finally, we need to acknowledge that the idea of using ML models for automatic moderation of user-generated content is not universally accepted, with some rejecting it as equivalent to censorship (Llansó 2020), and calling for regulations in this area (Meyer and Marsden 2019). Moreover, the recent changes in Twitter have served as an illustration of how relying on the automatic moderation to reduce operation costs (Paul and Dang 2022) can result in more prevalent misinformation (Graham and FitzGerald 2023).

10.3 Using BODEGA

Beyond the exploration of the current situation, we hope BODEGA will be useful for assessing the robustness of future classifiers and the effectiveness of new attacks. Towards this end, we make the software available openly,[s] allowing the replication of our experiments and evaluation of other solutions, both on the attack and the defence. Here, we also provide a handful of practical hints on how to use the software to perform such analysis in practice.

In order to **measure the robustness** of a classifier implemented in a particular scenario, the following is necessary:

1. Preparing a victim classifier. It can be based on the code in `runs/train_victims.py`, which provides training of baseline classifiers—BiLSTM, BERT or GEMMA—and only requires providing task-specific data. Otherwise, a completely different classifier can be

[r]Note that the works cited here are in print at the time of writing.
[s]https://github.com/piotrmp/BODEGA

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

included, as long as it implements the `OpenAttack. Classifier` interface. Note that both the classifier algorithm and training data will influence the robustness.

2. Choosing an attacker. For this purpose, the results in Table 4 can be helpful, as they show the quality of the AEs as well as the number of queries. If the tested classifiers are deployed in a service that only allows a limited number of queries, this should be taken into account in simulating an attack.
3. Evaluating an attack. This is performed by using the `runs/attack.py` script. Note that many of the attack methods consume significant computational resources and thus using a GPU device for both the victim and the attacker is recommended.
4. Analysing the results. BODEGA will output both the overall evaluation results and all of the successful AEs, with the changes highlighted. It is recommended to analyse these manually, as the automatic meaning preservation methods have their limits, especially in specialised text domains.

In order to **evaluate a new attack**, one needs to go through the following:

1. Implement an attacker. It needs to satisfy the `OpenAttack.attackers. ClassificationAttacker` interface, which sets out the procedure for finding AEs.
2. Choosing a victim. For the tasks and architectures tested here, the models are available for download from the BODEGA website. However, the `victims/transformer.py` script uses the *HuggingFace* library, so a user can train a model with a newer architecture, as long as it is available through `AutoModelForSequenceClassification` interface.
3. Evaluating an attack and analysing the results, as above.

These are the most obvious usages of BODEGA, but other scenarios are possible as well, such as modifying the evaluation measure (BODEGA score) by improving the semantic similarity assessment, adding a different text classification task, linguistic inquiry into the generated AEs, cybersecurity-focused analyses, etc.

11. Conclusion

Through this work, we have demonstrated that popular text classifiers, when applied for the purposes of misinformation detection, are vulnerable to manipulation through adversarial examples. We have discovered numerous cases where making a single barely perceptible change is enough to prevent a classifier from spotting non-credible information. Among the risk factors are large input lengths and the possibility of making numerous queries. Surprisingly, the classifiers trained on the basis of new state-of-the-art large language models are usually more vulnerable than their predecessors.

Nevertheless, the attack is never successful for every single instance and often entails changes that make text suspiciously malformed or ill-suited for the misinformation goal. This emphasises the need for thorough testing of the robustness of text classifiers at various stages of their development: from the initial design and experiments to the preparation for deployment, taking into account likely attack scenarios. We hope the BODEGA benchmark we contribute here, providing an environment for comprehensive and systematic tests, will be a useful tool in performing such analyses.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/nlp.2024.54

Acknowledgements. This work 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. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We also acknowledge the support from Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021) and the Spanish State Research Agency under the

Maria de Maeztu Units of Excellence Programme (CEX2021-001195-M). The computation for this study was made possible by the *Google Cloud Platform* through research credits.

Competing interests. The author(s) declare none.

References

Akers J., Bansal G., Cadamuro G., Chen C., Chen Q., Lin L., Mulcaire P., Nandakumar R., Rockett M., Simko L., Toman J., Wu T., Zeng E., Zorn B. and Roesner F. (2018). Technology-Enabled Disinformation: Summary, Lessons, and Recommendations. Technical report, University of Washington.

Al-Sarem M., Boulila W., Al-Harby M., Qadir J. and Alsaeedi A. (2019). Deep learning-based rumor detection on microblogging platforms: a systematic review. *IEEE Access* **7**, 152788–152812.

Ali H., Khan M. S., AlGhadhban A., Alazmi M., Alzamil A., Al-utaibi K. and Qadir J. (2021). All your fake detector are belong to us: evaluating adversarial robustness of fake-news detectors under black-box settings. *IEEE Access* **9**, 81678–81692.

Allcott H. and Gentzkow M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives* **31**(2), 211–236.

Alsmadi I., Ahmad K., Nazzal M., Alam F., Al-Fuqaha A., Khreishah A. and Algosaibi A. (2022). Adversarial NLP for social network applications: attacks, defenses, and research directions. *IEEE Transactions on Computational Social Systems*.

Alzantot M., Sharma Y., Elgohary A., Ho B.-J., Srivastava M. and Chang K.-W. (2018). Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics, pp. 2890–2896.

Bagdasaryan E. and Shmatikov V. (2022). Spinning language models: risks of propaganda-as-a-service and countermeasures. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, pp. 769–786.

Bakir V. and McStay A. (2017). Fake news and the economy of emotions: problems, causes, solutions. *Digital Journalism* **6**(2), 154–175.

Barrón-Cedeño A., Alam F., Struß J. M., Nakov P., Chakraborty T., Elsayed T., Przybyła P., Caselli T., Da San Martino G., Haouari F., Li C., Piskorski J., Ruggeri F., Song X. and Suwaileh R. (2024). Overview of the CLEF-2024 CheckThat! lab: check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*.

Brown B., Richardson A., Smith M., Dozier G. and King M.C. (2020). The adversarial UFP/UFN attack: a new threat to ML-based fake news detection systems? In *2020 IEEE Symposium Series on Computational Intelligence, SSCI*. IEEE, pp. 1523–1527.

Carter M., Tsikerdekis M. and Zeadally S. (2021). Approaches for fake content detection: strengths and weaknesses to adversarial attacks. *IEEE Internet Computing* **25**(2), 73–83.

Chen Y., Gao H., Cui G., Qi F., Huang L., Liu Z. and Sun M. (2022). Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11222–11237.

Church K.W. and Kordoni V. (2022). Emerging trends: SOTA-chasing. *Natural Language Engineering* **28**(2), 249–269.

Ciampaglia G.L., Mantzarlis A., Maus G. and Menczer F. (2018). Research challenges of digital misinformation: toward a trustworthy web. *AI Magazine* **39**(1), 65–74.

da San Martino G., Barrón-Cedeño A., Wachsmuth H., Petrov R. and Nakov P. (2020). Task 11: detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pp. 1377–1414.

Dalvi N., Domingos P., Mausam S.S. and Verma D. (2004). Adversarial classification. In *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (ACM), pp. 99–108.

Dettmers T., Pagnoni A., Holtzman A. and Zettlemoyer L. (2023). QLoRA: efficient finetuning of quantized LLMs. In Oh A., Neumann T., Globerson A., Saenko K., Hardt M. and Levine S. (eds), *Advances in Neural Information Processing Systems 36*. Curran Associates, Inc, pp. 10088–10115.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2018). 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, pp. 4171–4186.

Eger S., Şahin G. G., Rücklé A., Lee J.-U., Schulz C., Mesgar M., Swarnkar K., Simpson E. and Gurevych I. (2019). Text processing like humans do: visually attacking and shielding NLP systems. 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)*, Minneapolis, Minnesota. Association for Computational Linguistics, pp.1634–1647.

Ettinger A., Rao S., H. D. and Bender E.M. (2017). Towards linguistically generalizable NLP systems: a workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*. Association for Computational Linguistics (ACL), pp. 1–10.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Gao J., **Lanchantin J.**, **Soffa M.L. and Qi Y.** (2018). Black-box generation of adversarial text sequences to evade deep learning classifiers. In *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW*. IEEE, pp. 50–56.

Garg S. and Ramakrishnan G. (2020). BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 6174–6181.

Gong Z., **Wang W.**, **Li B.**, **Song D. and Ku W.-S.** (2018). Adversarial Texts with Gradient Methods. arXiv:1801.07175.

Gorwa R., **Binns R. and Katzenbach C.** (2020). Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society* **7**(1), 205395171989794.

Goto T., **Ono K. and Morita A.** (2024). A Comparative Analysis of Large Language Models to Evaluate Robustness and Reliability in Adversarial Conditions. techrxiv:171173447.70655950.

Graham T. and FitzGerald K.M. (2023). Bots, Fake News and Election Conspiracies: Disinformation During the Republican Primary Debate and the Trump Interview. Technical report, Digital Media Research Centre, Queensland University of Technology, Brisbane, Australia.

Graves L. (2018). Understanding the Promise and Limits of Automated Fact-Checking. Technical report, Reuters Institute, University of Oxford.

Guo C., **Sablayrolles A.**, **Jégou H. and Kiela D.** (2021). Gradient-based adversarial attacks against text transformers. In *EMNLP. 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), pp. 5747–5757.

Han S., **Gao J. and Ciravegna F.** (2019). 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*. Association for Computing Machinery, Inc, pp. 105–112.

Hidey C., **Chakrabarty T.**, **Alhindi T.**, **Varia S.**, **Krstovski K.**, **Diab M. and Muresan S.** (2020). DeSePtion: dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), pp. 8593–8606.

Hochreiter S. and Schmidhuber J. (1997). Long short-term memory. *Neural Computation* **9**(8), 1735–1780.

Horne B.D. and Adali S. (2017). 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*.

Hu E.J., **Shen Y.**, **Wallis P.**, **Allen-Zhu Z.**, **Li Y.**, **Wang S.**, **Wang L. and Chen W.** (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

Iyyer M., **Wieting J.**, **Gimpel K. and Zettlemoyer L.** (2018). Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL HLT. 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Association for Computational Linguistics (ACL), vol. 1, pp. 1875–1885.

Jaime L., **Flores Y. and Hao Y.** (2022). An adversarial benchmark for fake news detection models. In *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*.

Jin D., **Jin Z.**, **Zhou J.T. and Szolovits P.** (2020). 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*. AAAI Press, pp. 8018–8025.

Kantartopoulos P., **Pitropakis N.**, **Mylonas A. and Kylilis N.** (2020). Exploring adversarial attacks and defences for fake twitter account detection. *Technologies* **8**(4), 64.

Karpukhin V., **Oguz B.**, **Min S.**, **Lewis P.**, **Wu L.**, **Edunov S.**, **Chen D. and Yih W.-t.** (2020). Dense passage retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics, pp. 6769–6781.

Kiesel J., **Mestre M.**, **Shukla R.**, **Vincent E.**, **Adineh P.**, **Corney D.**, **Stein B. and Potthast M.** (2019). Task 4: hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA. Association for Computational Linguistics, pp. 829–839.

Kiesel J., **Potthast M.**, **Hagen M. and Stein B.** (2017). Spatio-temporal analysis of reverted wikipedia edits. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 11, pp. 122–131.

Kingma D.P. and Ba J.L. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR. 2015 - Conference Track Proceedings*, San Diego, USA. ICLR.

Koenders C., **Filla J.**, **Schneider N. and Woloszyn V.** (2021). How Vulnerable Are Automatic Fake News Detection Methods to Adversarial Attacks? arXiv:2107.07970.

Kumar S., **West R. and Leskovec J.** (2016). Disinformation on the web: impact, characteristics, and detection of wikipedia hoaxes. In *25th International World Wide Web Conference, WWW 2016*. International World Wide Web Conferences Steering Committee, pp. 591–602.

Le T., **Wang S. and Lee D.** (2020). MALCOM: generating malicious comments to attack neural fake news detection models. In *Proceedings - IEEE International Conference on Data Mining, ICDM*. IEEE, pp. 282–291.

Lee S., **Lee J.**, **Moon H.**, **Park C.**, **Seo J.**, **Eo S.**, **Koo S. and Lim H.** (2023). A survey on evaluation metrics for machine translation. *Mathematics* **11**(4), 1006

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

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

Lewandowsky S., **Ecker U.K. and Cook J.** (2017). Beyond misinformation: understanding and coping with the "Post-truth" era. *Journal of Applied Research in Memory and Cognition* **6**(4), 353–369.

Li L., **Ma R.**, **Guo Q.**, **Xue X. and Qiu X.** (2020). 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, pp. 6193–6202.

Lialin V., **Deshpande V. and Rumshisky A.** (2023). Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning. arXiv preprint arXiv:2303.15647.

Liu Y., **Cong T.**, **Zhao Z.**, **Backes M.**, **Shen Y. and Zhang Y.** (2024). Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models .

Liu Y. and Wu Y.F.B. (2020). FNED: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)* **38**(3), 1–33.

Llansó E.J. (2020). No amount of "AI" in content moderation will solve filtering's prior-restraint problem. *Big Data and Society* **7**(1), 205395172092068.

Loshchilov I. and Hutter F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR. 2019*, New Orleans, LA, USA.

MacCartney B. (2009). Natural Language Inference. Ph. d. thesis, Stanford University.

Mannes A. (2020). Governance, risk, and artificial intelligence. *AI Magazine* **41**(1), 61–69.

Meyer T. and Marsden C. (2019). Regulating disinformation with artificial intelligence: Effects of disinformation initiatives on freedom of expression and media pluralism. Technical report, European Parliament.

Mierzyńska A. (2020). Chmura znad Czarnobyla - kolejna dezinformacja, która straszono Polaków. Wiemy, skad się wzięła.

Morris J., **Lifland E.**, **Yoo J.Y.**, **Grigsby J.**, **Jin D. and Qi Y.** (2020). 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*, Online. Association for Computational Linguistics, pp. 119–126.

Nakov P., **Barrón-Cedeño A.**, **Da San Martino G.**, **Alam F.**, **Míguez R.**, **Caselli T.**, **Kutlu M.**, **Zaghouani W.**, **Li C.**, **Shaar S.**, **Mubarak H.**, **Nikolov A. and Kartal Y.S.** (2022). Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *CLEF 2022: Conference and Labs of the Evaluation Forum*, Bologna, Italy, vol. 3180, pp. 368–392. CEUR Workshop Proceedings (CEUR-WS.org).

Neekhara P., **Hussain S.**, **Dubnov S. and Koushanfar F.** (2019). Adversarial reprogramming of text classification neural networks. In *EMNLP-IJCNLP. 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp. 5216–5225.

Paul K. and Dang S. (2022). Exclusive: twitter leans on automation to moderate content as harmful speech surges.

Potthast M., **Kiesel J.**, **Reinartz K.**, **Bevendorff J. and Stein B.** (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 231–240.

Przybyła P. (2020). Capturing the style of fake news. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, New York, USA. AAAI Press, vol. 34, pp. 490–497.

Przybyła P. (2022). LAMBO: Layered Approach to Multi-level BOundary identification.

Przybyła P., **Borkowski P. and Kaczyński K.** (2022). Countering disinformation by finding reliable sources: a citation-based approach. In *Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Przybyła P., **Wu B.**, **Shvets A.**, **Mu Y.**, **Sheang K.C.**, **Song X. and Saggion H.** (2024). Overview of the CLEF-2024 checkThat! lab Task 6 on robustness of credibility assessment with adversarial examples (InCrediblAE). In Faggioli G., Ferro N., Galuščáková P. and García Seco de Herrera A. (eds), *Working Notes of CLEF. 2024 - Conference and Labs of the Evaluation Forum, CLEF 2024, Grenoble, France*

Radford A., **Wu J.**, **Child R.**, **Luan D.**, **Amodei D. and Sutskever I.** (2018). Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Ren S., **Deng Y.**, **He K. and Che W.** (2019). Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics, pp. 1085–1097.

Ribeiro M.T., **Singh S. and Guestrin C.** (2018). Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics, pp. 856–865.

Sellam T., **Das D. and Parikh A.** (2020). BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 7881–7892.

Shu K., **Wang S. and Liu H.** (2019). Beyond news contents: the role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, New York, NY, USA. ACM, vol. 9.

https://doi.org/10.1017/nlp.2024.54 Published online by Cambridge University Press

Singhal M., Ling C., Paudel P., Thota P., Kumarswamy N., Stringhini G. and Nilizadeh S. (2022). 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.

Smith T.J. (1989). *Propaganda: A pluralistic perspective*. Praeger.

Smith M., Brown B., Dozier G. and King M. (2021). Mitigating attacks on fake news detection systems using genetic-based adversarial training. In *2021 IEEE Congress on Evolutionary Computation, CEC. 2021 - Proceedings*. IEEE, pp. 1265–1271.

Srivastava B., Lakkaraju K., Bernagozzi M. and Valtorta M. (2023). Advances in automatically rating the trustworthiness of text processing services. In *Spring Symposium on AI Trustworthiness Assessment*.

Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I. and Fergus R. (2013). Intriguing properties of neural networks. arXiv: 1312.6199.

Team Gemma and DeepMind Google (2024). Gemma: Open Models Based on Gemini Research and Technology. Technical report, Google DeepMind.

Thakur N., Reimers N., Daxenberger J. and Gurevych I. (2021). Augmented SBERT: data augmentation method for improving Bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics, pp. 296–310.

Thorne J., Vlachos A., Christodoulopoulos C. and Mittal A. (2019). Evaluating adversarial attacks against multiple fact verification systems. In *EMNLP-IJCNLP. 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics, pp. 2944–2953.

Thorne J., Vlachos A., Cocarascu O., Christodoulopoulos C. and Mittal A. (2018a). The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*.

Thorne J., Vlachos A., Cocarascu O., Christodoulopoulos C. and Mittal A. (2018b). The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*.

Tucker J.A., Guess A., Barberá P., Vaccari C., Siegel A., Sanovich S., Stukal D. and Nyhan B. (2018). Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Technical report, Hewlett Foundation.

van der Linden S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* **28**(3), 460–467.

Vlachos A. and Riedel S. (2014). Fact checking: task definition and dataset construction. In *Proceedings of the ACL. 2014 Workshop on Language Technologies and Computational Social Science*, pp. 18–22.

Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., von Platen P., Ma C., Jernite Y., Plu J., Xu C., Scao T.L., Gugger S., Drame M., Lhoest Q. and Rush A.M. (2020). Transformers: state-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online. Association for Computational Linguistics, pp. 38–45.

Yao Y., Duan J., Xu K., Cai Y., Sun Z. and Zhang Y. (2024). A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. *High-Confidence Computing* **4**(2), 100211.

Yoo K., Kim J., Jang J. and Kwak N. (2022). Detection of adversarial examples in text classification: benchmark and baseline via robust density estimation. In Muresan S., Nakov P. and Villavicencio A. (eds), *Findings of the Association for Computational Linguistics: ACL, Dublin, Ireland*. Association for Computational Linguistics, pp. 3656–3672.

Zang Y., Qi F., Yang C., Liu Z., Zhang M., Liu Q. and Sun M. (2020). Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics, pp. 6066–6080.

Zeng G., Qi F., Zhou Q., Zhang T., Ma Z., Hou B., Zang Y., Liu Z. and Sun M. (2021). 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), pp. 363–371.

Zhang T., Kishore V., Wu F., Weinberger K.Q. and Artzi Y. (2020a). BERTScore: evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*, Addis Ababa, Ethiopia.

Zhang W.E., Sheng Q.Z., Alhazmi A. and Li C. (2020b). Adversarial attacks on deep-learning models in natural language processing. *ACM Transactions on Intelligent Systems and Technology (TIST)* **11**(3), 1–41.

Zhou Z., Guan H., Bhat M.M. and Hsu J. (2019). Fake news detection via NLP is vulnerable to adversarial attacks. In *ICAART. 2019 - Proceedings of the 11th International Conference on Agents and Artificial Intelligence*. SciTePress, vol. 2, pp. 794–800.

Cite this article: Przybyła P, Shvets A and Saggion H. Verifying the robustness of automatic credibility assessment. *Natural Language Processing* https://doi.org/10.1017/nlp.2024.54