Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Experiments with adversarial attacks on text genres

Mikhail Lepekhin
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
lepehin.mn@phystech.edu

Serge Sharoff
University of Leeds
Leeds, UK
s.sharoff@leeds.ac.uk

Abstract

Neural models based on pre-trained transformers, such as BERT or XLM-RoBERTa, demonstrate SOTA results in many NLP tasks, including non-topical classification, such as genre identification. However, often these approaches exhibit low reliability to minor alterations of the test texts. One of the problems concerns topical biases in the training corpus, for example, the prevalence of words on a specific topic in a specific genre can trick the genre classifier to recognise any text on this topic in this genre. In order to mitigate this problem, we investigate techniques for attacking genre classifiers to understand the limitations of the transformer models and to improve their performance. While simple text attacks, such as those based on word replacement using keywords extracted by tf-idf, are not capable of deceiving powerful models like XLM-RoBERTa, we show that embedding-based algorithms which can replace some of the most "significant" words with words similar to them, for example, TextFooler, have the ability to influence model predictions in a significant proportion of cases.
Keywords: textfooler, genre classification, non-topical classification, bert
DOI: 10.28995/2075-7182-2022-21-1097-1108

Эксперименты с состязательными атаками на текстовые жанры

Михаил Лепехин
Московский Физико-Технический Институт
Долгопрудный, Россия
lepehin.mn@phystech.edu

Сергей Шаров
Университет Лидса
Лидс, Великобритания
s.sharoff@leeds.ac.uk

Аннотация

Нейронные модели, основанные на предобученных трансформаторах, таких как BERT или XLM-RoBERTa, демонстрируют результаты SOTA во многих задачах NLP, включая задачи нетематической классификации текстов. Однако часто эти подходы демонстрируют низкую надежность при незначительных изменениях текстов в тестовом датасете. Одна из проблем связана с тематическими сдвигами в учебном корпусе, например, преобладание слов по определенной теме в определенном жанре может обмануть классификатор жанров, и привести к тому, что классификатор научится распознавать преобладающую тему вместо заданного жанра. Чтобы смягчить эту проблему, мы исследуем методы атаки на классификаторы жанров, чтобы понять ограничения моделей transformer и улучшить их производительность. В то время как простые текстовые атаки, основанные на замене ключевых слов, извлеченных tf-idf, не способны обмануть мощные модели, такие как XLM-RoBERTa, мы показываем, что алгоритмы на основе эмбеддингов текстов, которые могут заменить некоторые из наиболее "значимых"слов словами, похожими на них, например, TextFooler, обладают способностью влиять на прогнозы моделей в значительной доле случаев.
Ключевые слова: textfooler, жанровая классификация, нетематическая классификация, bert

1 Introduction

Non-topical text classification concerns a wide range of problems that are aimed at predicting a text property that is not connected directly to the text topic, for example, at predicting its genre, difficulty level, the age or the first language of its author, etc. Unlike topical text classification, non-topical text classification needs a model that predicts a label on the basis of its stylistic properties. Automatic genre

identification is one of the standard problems of non-topical text classification, as it is useful in many areas such as information retrieval, language teaching or basic linguistic research [Santini et al.2010].

An early comparison of various datasets, models and linguistic features for genre classification [Sharoff et al.2010] shows that traditional machine learning models, for example, SVM, can be very accurate in genre classification on their native dataset, but suffer from a dataset shift. Since then, many new approaches for text classification have emerged. In particular, BERT (Bidirectional Encoder Representations from Transformers) is an efficient pre-trained model based on the Transformer architecture [Devlin et al.2018]. It achieves the state-of-the-art results for various NLP tasks, including text classification. XLM-RoBERTa [Conneau et al.2019] is an improved variant of BERT. It has the same architecture, but uses bigger and more genre diverse corpora and an updated pre-training procedure. In addition, XLM-RoBERTa is multilingual. Therefore, we choose XLM-RoBERTa as the classifier for the experiments in this study.

One of the most significant problems in genre classification is topical shifts [Petrenz and Webber2010]. If a specific topic is more frequent in the training corpus for a specific genre than many classification models can be biased towards indicating this genre by the keywords of this topic. This becomes especially problematic in the case of data shift [Petrenz and Webber2010]. For this reason, they check reliability of their genre classifiers via testing on the datasets from different domains, and so do we in our work.

There are numerous attempts to attack various NLP models by making minor changes to a text which lead to different predictions. An overview of different methods is presented in [Huq and Pervin2020]. These techniques help to reveal the flaws of the NLP models and to find out what are the features in the texts that are taken into account by the models. TextFooler [Jin et al.2019] sorts the lexicon of the texts in the order of the impact on the target class probability and tries to replace the most important words with one of the most similar words to it where similarity is defined as the dot product between the corresponding word embeddings. BertAttack [Li et al.2020] has a similar algorithm, but instead of using word embeddings it relies on Bert token embeddings. Because of this, BertAttack processes the whole words and subword tokens in different ways, while trying to find suitable words to replace subword tokens.

Until now, there were no reports of successful attempts on attacking genre classifiers or non-topical classification in general using neural methods, even though it is important to understand their reliability and to find ways for improving their robustness. In this study, we test two methods to attack text genre classifiers. The first method is based on swapping the keywords which are found with tf-df extraction, while the second method applies a modified TextFooler algorithm. Moreover, we try to improve the performance of the original classifiers by adding a set of texts broken by TextFooler to the training corpus.

In this paper we perform the following steps to investigate attacking techniques and to improve the reliability of the genre classifier:

1. training a baseline classifier using XLM-RoBERTa (Section 2);
2. attacking the XLM-RoBERTa classifier by swapping topical keywords between the genres (Section 3.1);
3. attacking the XLM-RoBERTa classifier with TextFooler (Section 3.2);
4. performing targeted attacks on the XLM-RoBERTa classifier (Section 3.3);
5. training a new XLM-RoBERTa classifier by using the original training corpus combined with the successfully attacked texts (Section 4);

Data and scripts to replicate our experiments are available.[1]

2 Baseline

2.1 Training data

For training, we use existing FTD datasets in English and in Russian [Sharoff2018]. Each of them contains nearly 2000 texts from a wide range of sources annotated with 10 genre labels, see Table 1. The

[1] `https://github.com/MikeLepekhin/TextGenresAttack`

Genre label	Prototypes	FTD EN		FTD RU		Homogeneous EN		LJ RU
		Train	Val	Train	Val	Test	Sources	Test
Argument	Expressing opinions, editorials	276	77	207	77	400	[Kiesel et al.2019]	481
Fiction	Novels, songs, film plots	69	28	62	23	400	BNC	199
Instruction	Tutorials, FAQs, manuals	141	50	59	17	400	StackExchange	384
News	Reporting newswires	114	37	379	103	400	Giga News	1518
Legal	Laws, contracts, T&C	56	17	69	13	400	Legal codes	14
Personal	Diary entries, travel blogs	72	19	126	49	400	ICWSM	513
Promotion	Adverts, promotional postings	218	66	222	85	400	websites	68
Academic	Academic research papers	59	23	144	49	400	arxiv.org	20
Information	Encyclopedic articles	131	38	72	33	400	Wikipedia	171
Review	Product reviews	48	22	107	34	400	Amazon	185
	Total	1184	377	1447	483	4000		3553

Table 1: Training and testing corpora

dataset is relatively balanced with the most common categories being Argumentation and Promotion. For validation of the success of attacking models at the last stage (see the next section) we reserve a small dataset obtained by stratified sampling (columns Val in Table 1), which is not used in the training and attacking pipelines.

It is known that genre classifiers are often not robust when applied to a different corpus with the same labels [Sharoff et al.2010], therefore we use independently produced test sets to simulate out-of-domain performance on large collections coming from a smaller number of sources. This is in comparison to the training datasets, which came from a much wider range of sources.

For the Russian test set we use 3,500 posts from LiveJournal, a social media platform popular in Russia. Since LiveJournal is a social media platform, the distribution of its texts significantly differs from that in the FTD corpora. It contains less of Legal, Academic and Promotion texts. But it has more News, Personal and Instruction texts. The most popular genres in the Russian test corpus are News and Promotion (Table 1).

As we lack an independent test set for English, we use "natural annotation" in the sense of using a text collection from sources relatively homogenous with respect to their genres, such as StackExchange which mainly contains instructive texts, Wikipedia which mainly contains texts for reference information, see more details in the Sources column in Table 1.

2.2 Training genre classifiers

We fine-tune the baseline XLM-RoBERTa classifier following the same archine as [Sun et al.2019]. This model has around 279M parameters. We concatenate the English and the Russian texts from the training part of the FTD corpus, since XLM-RoBERTa is a multilingual model which has shown better results on concatenation in our initial experiments. To establish the accuracy of the respective classifier models, we keep the record of the source text language in the five-fold cross-validation procedure. The classifier is trained during 10 epochs with the Adam optimiser with learning rate $= 5 \cdot 10^{-5}$ since these hyperparameters are used in the original papers for several BERT-like models [Devlin et al.2018, Liu et al.2019]. We use one TITAN RTX-based GPU for each experiment. Available GPU capacity: 24 GB.

3 Genre attacks

The genre attack task is to make minimal alterations to a target text to change its prediction. If a test text can be altered to change its label and this can be achieved within a set limit of alterations, the text is counted as "broken". We can try untargeted and targeted attacks

Replaced	10%	50%	100%
EN	14 (1,1%)	31 (2,5%)	196 (15,5%)
RU	22 (1,5%)	44 (3,0%)	148 (10,0%)

Table 2: Successful attacks with tf-idf keyword replacement

untargeted the attacks that intend to force the classifier to change its correct prediction of test set text to produce any incorrect label from our set of labels without considering a specific label;

targeted the opposite direction of attack when we attack texts for which the classifier makes a mistake by making alterations to force the classifier to predict the correct label.

The genre attacks are conducted to achieve cross-validation for attacks without leaking information about the target texts to the classifier: we randomly shuffle the train dataset and make 5 iterations of the cross-validation mechanism: For every i the texts with numbers from $0.2i|X|$ to $0.2(i+1)|X|-1$ are used to attack a classifier which has been trained on the remaining texts from the training corpus. Thus, we get 5 architecturally identical classifier models with slightly different weights, as well as a set of successfully attacked texts we use for the following analysis.

Unlike the classifier training step, we train text attack models separately for each language as this helps in achieving successful attacks on more texts.

3.1 Attacking by swapping genre keywords

First, we test a simple text attack generator which is based on replacing the keywords extracted for each genre, with the keywords defined by their tf-idf scores within the genre texts. We collect keywords individually for each genre from the list of Noun, Adjective, Verb, Adverb. For each genre, we take all the texts corresponding to this genre, from them we select the words of the corresponding part of speech and concatenate all these words into one document. Having thus obtained one document for each genre, we count the tf-idf signs for the corresponding words. The resulting keywords generally turn out to be quite relevant to the corresponding genre labels. But sometimes, it reveals topical shifts. For instance, such words as *united*, *nations*, *international* show the predominance of specific political texts in the genre of argumentation, but do not seem to be specific to the genre. However, for the Legal and Academic genres, the keywords seem to be much more genre-specific. Words *system*, *quantum*, *software*, and *node* correspond to texts that describe mathematics or technological details. shall, article, *paragraph*, *court* are quite specific to legal documents.

Then the attack generator replaces a certain percentage of the keywords for every genre to a keyword of a different genre. We choose the following percentage of the keywords for replacement: 10%, 50%, 100%. Contrary to our expectations concerning the prevalence of topic-specific keywords, our XLM-R classifier is reasonably robust to attacks on both English and Russian texts, as the rate of successfully broken texts is fairly low, see Table 2.

3.2 Attacking with untargeted TextFooler

The original TextFooler algorithm has the following stages. First, we order the words w_i (after excluding the stop-words) by the descending order of word importance scores I_w, that defined in the following way:

$$\mathrm{I}_{w_i} = \left\{ \begin{array}{l} F_Y(X) - F_Y(X_{\setminus w_i}), \text{ if } F(X) = F(X \setminus w_i) \\ (F_Y(X) - F_Y(X_{\setminus w_i})) + (F_{\widehat{Y}}(X_{\setminus w_i}) - F_{\widehat{Y}}(X)), \\ \text{if } F(X) = Y, F(X \setminus w_i) = \widehat{Y}, Y \neq \widehat{Y} \end{array} \right\},$$

where $F(X)$ is predicted label for the text X, and $F_C(X)$ is predicted probability of the genre C for the text X. The intuition of the importance is that the more is distortion of the predicted probability distribution after removal of a word, the more important the word is.

Then for every word in the attacked text the k closest words are chosen by maximising the dot product of their embeddings with the embedding of the original word. These words are the candidates for replacing the original word. We iterate through the words w_i and try to replace it with one of the candidates following a set of filters. If we succeeded to do that, then the text replacement is considered as successful. Otherwise, we continue to iterate through the list of candidates. If we could not find a candidate $_i$ for replacing the word w_i, we take the word that the classifier will give the minimal probability of the original class for the text with this replacement. If we have iterated all over the words w_i, but the classifier still predicts the original label for the text, the attack is unsuccessful.

The filters for choosing a suitable replacement can vary. First, we can keep the same part-of-speech tag. Second, we can vary the lower limit threshold for the word similarity score for each candidate. In the original TextFooler algorithm, it is fixed at 0.5. In our study the cosine embedding similarities between each word and its closest neighbour lie between 0.61 (0.2-percentile) and 0.82 (0.8-percentile) for the English words, and between 0.67 (0.2-percentile) and (0.82-percentile) for the Russian ones. For both languages, the minimum cosine similarity among the pairs of most similar embeddings is 0.34. If we take into account the top-15 most similar embeddings for each word embedding, the 20–80 percentile range for English is 0.49–0.66, for Russian it is 0.52–0.68. This limits the range for the selection threshold.

Finally, to preserve the meaning and the grammatical correctness of the attacked text, we estimate the similarity between the original sentence and its attacked version with the Universal Sentence Encoder [Cer et al.2018]. The original TextFooler paper fixed the minimal dot product between the USE embedding of the original text and the attacked one to the threshold to 0.84, we tried varying it in our study.

In our experiments when attacking the datasets with TextFooler we use the same cross-validation mechanism.

Our experiments with applying TextFooler to genre classification produced convincing replacements which preserved the meaning at the word level for both English and Russian. However, we found that preserving the grammar is trickier, especially for Russian. Probably, due to the richness of the Russian morphology, the grammatical cases, noun genders and plural forms are not coherent in many broken texts. The same phenomenon is also common in English: Table 6 shows an example of alteration that makes a text ungrammatical.

Genre	English	Russian
Argument	18.3	13
Fiction	21.8	14
Instruction	29.8	25
News	30.5	24
Legal	24.3	17.5
Personal	8.4	6
Promotion	31.3	26.5
Academic	20.6	18
Information	11	5.5
Review	8.9	3.5

Table 3: The median number of words per text for successful genre attacks

We also made two experiments when the replacement of the stop-words is allowed and not. We find that there is no big difference in the number of broken texts in either case. Furthermore, we experimented with various values of k and the minimal USE score to find out how they affected the number of the attacked texts and the robustness of the XLM-RoBERTa model trained on them. Since the original TextFooler implementation in the TextAttack framework [Morris et al.2020] does not contain embeddings for Russian, we used FastText embeddings for both English and Russian to make the experiments with both languages identical.

Table 3 lists the results for untargeted attack for both English and Russian FTD corpora in terms of the number of words needed for a successful change of genre predictions for a text.

Table 4 shows that the number of the successfully attacked texts is practically independent from the

USE	**Language**	**k=15**	**k=30**	**k=50**
0.84	EN	416 (32,9%)	438 (34,7%)	453 (35,8%)
0.84	RU	686 (47,4%)	718 (49,6%)	744 (51,4%)
0.6	EN	424 (33,5%)	444 (35,1%)	457 (36,2%)
0.6	RU	687 (47,5%)	720 (49,8%)	744 (51,4%)
0	EN	424 (33,5%)	444 (35,1%)	457 (36,2%)
0	RU	687 (47,5%)	720 (49,8%)	744 (51,4%)

Table 4: Successful untargeted attacks with different USE thresholds

Original	**Attacked**
As a Company Limited by Guarantee **this** charity is owned not by any shareholders but by its members . Only members can vote at Annual General Meetings to elect officers and Directors or become Directors of the charity . So if you would like to help us in this way , contributing at least £ 5 per year and in return receive regular updates and an invitation to the AGM please complete a membership form Company Membership Form Friends Membership Form There is also the option to make a monthly donation towards our work . As little as £ 2 a month can make a real difference to Emmaus Projects .	As a Company Limited by Guarantee **that** charity is owned not by any shareholders but by its members . Only members can vote at Annual General Meetings to elect officers and Directors or become Directors of the charity . So if you would like to help us in this way , contributing at least £ 5 per year and in return receive regular updates and an invitation to the AGM please complete a membership form Company Membership Form Friends Membership Form There is also the option to make a monthly donation towards our work . As little as £ 2 a month can make a real difference to Emmaus Projects .
label: Promotion	**label: Argument**

Table 5: Example of an untargeted attack

USE threshold when it varies from the default 0.84 to 0. At the same time, as expected the proportion of the broken texts increases when more variants for attack are considered (the value of k, the number of nearest neighbours to consider).

Besides, TextFooler turned out to be more efficient for the Russian texts, about 15% difference in the proportion of broken texts. However, we should note that TextFooler tries to attack only texts which the model classifies correctly. As the XLM-RoBERTa classifier performed better on the Russian texts, we make more attacks on Russian texts in general.

Table 5 represents an example of a text, successfully broken by our mechanism. It shows that a replacement of just one word to its synonyms is able to change the classifier prediction.

3.3 Targeted attacks with TextFooler

For targeted attacks we use the same mechanism with TextFooler, but we choose the replacement candidate that maximises the probability of the true class.

Table 7 lists for how many texts the classifier predictions can be improved by the attack mechanism. Targeted attacks are harder that the untargeted ones.

4 Genre classifiers trained on the attacked texts

Table 9 lists the robust classifier performance on the test corpora. It shows that the XLM-RoBERTa classifier trained on the attacked texts attains higher accuracy than the baseline classifier. Table 10 shows, that for most genres the robust classifier achieves higher f1-score. The same is true for precision and recall. There is no genre for which both precision and recall with the base model are higher than those with the robust one. But some genres still have worse f1-scores with robust classifiers. This happens because making classifiers more robust to the topical shifts does not benefit the score on test data if it has the same topical shift as the training data. For example, politics is the prevailing topic in News and Argument texts, and most Review texts are devoted to movies in the English FTD corpus.

Table 11 shows an example of a successful targeted attack for an English text. The number of replaced words is low, and the grammatical correctness is preserved.

Original	**Attacked**
In addition to the internet connection , you **should** also **try** to have at least 100 MB of free space available on your drive when you install Titan Poker .	In addition to the internet connection , you **need** also **trying** to have at least 100 MB of free space available on your drive when you install Titan Poker .
label: Instruct	**label: Eval**

Table 6: Example of deterioration of grammar in untargeted attack

Language	**k=15**	**k=30**	**k=50**
RU	317 (57,3%)	326 (59,0%)	328 (59,3%)
EN	233 (34,2%)	248 (36,4%)	254 (37,2%)

Table 7: The number of the texts broken by the targeted attack, USE threshold = 0.84

Genre	F1		Prec		Rec	
	Base	Robust	Base	Robust	Base	Robust
Argument	**0.585**	0.550	0.514	**0.612**	**0.678**	0.499
Fiction	**0.685**	0.677	**0.902**	0.697	0.553	**0.658**
Instruction	0.651	**0.738**	**0.891**	0.762	**0.813**	0.716
News	**0.940**	0.937	0.917	**0.943**	**0.965**	0.931
Legal	0.585	**0.615**	0.444	**0.480**	**0.857**	**0.857**
Personal	**0.742**	0.723	**0.747**	0.657	0.737	**0.805**
Promotion	0.333	**0.408**	0.316	**0.369**	0.353	**0.456**
Academic	0.273	**0.489**	0.250	**0.440**	0.300	**0.550**
Information	**0.586**	0.578	**0.690**	0.596	0.509	**0.561**
Review	**0.571**	0.535	0.550	**0.559**	**0.595**	0.514

Table 8: Comparison of the Base and the Robust XLM-RoBERTa results for English

Training XLM-RoBERTa on concatenation of the original and broken texts does not improve the classifier performance on the LiveJournal corpus but significantly increases the accuracy on the English genre corpus with natural annotation. Besides, the best result is attained when hyper-parameter value $k = 15$ is used. It shows that the quality of attack is more important than the number of the successfully attacked texts for boosting the classifier performance. In the Table 8 we can see that the robust classifier performs better for most genres. In the Table 12 the improvement in terms of the F1 score is limited, since for many genres improving recall implies deterioration of precision.

Besides, we conduct a mechanism for attacking the XLM-RoBERTa classifier trained on the texts broken by the targeted attacks. And we do the same thing for the texts made by the untargeted attacks. We show that the classifier trained on the broken texts from the targeted attacks is significantly more reliable than the original one. The difference from the original mechanism is that here we use the train subset for training classifiers, but the TextFooler attacks are performed on the validation subset Table 1 that makes 25% of the train corpus. In other words, we do not use here the cross-validation mechanism we used before.

Table 14 and Table 13 list the number of the Russian and the English texts, successfully attacked by the target attack mechanism. As for the untargeted attacks, the number of the broken Russian texts is higher than that for the English ones. It could be caused by the morphological richness of the Russian language. It has numerous suffices and word endings many of which are represented by a single XLM-RoBERTa token. Therefore, frequently changing of morphological forms of some words causes change to the genre predicted by the classifier.

5 Related Work

Genre classification is not a new task, since non-topical classification is needed for many applications. There have been experiments with various architectures from linear discriminant analysis [Karlgren

Corpus	no attacked	k=15	k=30	k=50
Ru, LiveJournal	0.76 ± 0.003	0.756 ± 0.008	0.755 ± 0.009	0.756 ± 0.005
En, Genre-homogeneous texts	0.747 ± 0.026	0.796 ± 0.011	0.771 ± 0.01	0.776 ± 0.029

Table 9: Accuracy of the XLM-RoBERTa classifier trained on the attacked texts

Genre	F1		Prec		Rec	
	Base	Robust	Base	Robust	Base	Robust
Argument	0.566	**0.732**	0.534	**0.724**	0.603	**0.740**
Fiction	0.914	**0.929**	**0.951**	0.913	0.88	**0.945**
Instruction	0.448	**0.621**	0.613	**0.636**	0.353	**0.608**
News	0.689	**0.856**	0.529	**0.784**	**0.988**	0.943
Legal	**0.798**	0.652	0.985	**0.995**	**0.670**	0.485
Personal	0.658	**0.702**	0.580	**0.681**	**0.760**	0.725
Promotion	0.502	**0.885**	0.802	**0.915**	0.365	**0.858**
Academic	**0.910**	0.888	**0.883**	0.820	0.940	**0.968**
Information	**0.944**	0.847	**0.917**	0.753	**0.973**	0.968
Review	0.752	**0.777**	**0.933**	0.865	0.630	**0.705**

Table 10: Comparison of the Base and the Robust XLM-RoBERTa results for the English homogenous corpus

and Cutting1994] to SVM [Dewdney et al.2001] to recurrent neural networks [Kunilovskaya and Sharoff2019]. Early work on detection of topical shifts in genre classification [Sharoff et al.2010] reveals the problem of topic shifts in the genre corpora. In this paper we try to solve the problem indirectly. [Petrenz and Webber2010] investigate a very important idea concerning estimation of the reliability of genre classifiers based via its validation on a corpus with different topical domains but with the same genre labels. Our study continues this line of research when we use the datasets from natural annotation and LiveJournal to estimate the model accuracy on an out-of-domain testing corpus.

Our experiments on using adversarial attacks for genre classification are novel. The most efficient adversarial attack techniques for classifiers [Jin et al.2019, Li et al.2020] are based on usage of word-level embeddings and finding for each word a fixed number of the most similar words as candidates for replacing with. Our genre attacks are based on the TextFooler [Jin et al.2019] with a modification that we allow replacing of the stop-words and vary the USE threshold. TextFooler [Jin et al.2019] was chosen as the basis for genre attacks in this study due to its efficiency and flexibility as it can be applied to various neural models. We also experimented with BertAttack, that differs from the TestFooler algorithm in its usage of BERT token embeddings instead of pre-trained word-level embeddings. In our initial experiments we found it to be much slower than TextFooler and also somewhat less efficient for the genre attack task. The percent of the texts successfully broken by BertAttack is lower than 15% for the English language. Therefore, we only report the results with TextFooler here.

6 Conclusions

In our experiments we show that XLM-RoBERTa genre classifier is resistant to simple attack methods, such as replacement of genre keywords. At the same time it can be easily deceived by word-based adversarial attacks, such as TextFooler. In the case of the baseline classifier, more than 35% of English texts in the training corpus can be successfully broken, raising to more than 50% for Russian. In addition, the number of successfully attacked texts is an important metric for estimating the robustness of the classifiers. The lower the number of broken texts, the more difficult it is to break the classifier which implies higher robustness. In the future, we plan to work on correcting grammar when attacking text classifiers. In particular, we plan to use the pymorphy library to bring words after attack to the correct form. It can help to cut off cases when successful attack occurs by replacing the original word with the same word, but in a different grammatical form.

Original	**Attacked**
CVC Capital Partners , the UK private equity **firm** , which is currently in the process of making an offer to purchase Forbo , the Swiss flooring and drive belts **manufacturer** , would consider selling Forbo's Swift adhesives division , if it is successful in its bid , a source close to the situation	CVC Capital Partners , the UK private equity **company** , which is currently in the process of making an offer to purchase Forbo , the Swiss flooring and drive belts **maker** , would consider selling Forbo's Swift adhesives division , if it is successful in its bid , a source close to the situation
label: News	**label: Promotion**

Table 11: Example of a targeted attack

Genre	F1		Prec		Rec	
	Base	Robust	Base	Robust	Base	Robust
Argument	0.584	**0.728**	0.487	**0.734**	**0.723**	**0.723**
Fiction	0.913	**0.928**	**0.977**	0.907	0.858	**0.950**
Instruction	0.535	**0.617**	**0.708**	0.635	0.430	**0.600**
News	0.816	**0.848**	0.710	**0.767**	**0.960**	0.948
Legal	**0.860**	0.652	0.990	**0.995**	**0.760**	0.485
Personal	0.682	**0.707**	**0.719**	0.685	0.648	**0.730**
Promotion	0.819	**0.881**	**0.914**	0.912	0.742	**0.853**
Academic	**0.892**	0.886	**0.823**	0.816	**0.975**	0.968
Information	**0.942**	0.845	**0.923**	0.753	**0.963**	**0.963**
Review	**0.812**	0.774	**0.892**	0.857	**0.745**	0.705

Table 12: Comparison of the Base and the Robust XLM-RoBERTa results for Russian

We also tried targeted attacks, but the classifiers trained on the targeted attacked texts performed worse than those trained on the untargeted attacked ones. Our experiments demonstrate the effectiveness of TextFooler at generating targeted adversarial texts for genre classification. Also we find some important patterns in the attack results:

1. the threshold for USE almost does not affect the number of the attacked texts;
2. attacks are more efficient for the Russian language;
3. the higher the number of replacing candidates, the less the difference between reliability of the original and the robust classifier fine-tuned on the attacked texts.

In addition, adding broken texts improves the overall accuracy. It happens because texts in the new collection cannot be broken by the same set of adversarial attacks, thus implying a more robust classifier. It could be caused by the initial bias of our genre classifiers, but in practice it is difficult to find out. New methods for developing more robust genre classifiers is one of the important practical applications of the adversarial attacks on genres.

Model	**k=15**	**k=30**	**k=50**
base	234	247	252
targeted	254	269	272
robust	209	234	244

Table 13: Targeted attack on English texts, USE threshold=0.84

Model	**k=15**	**k=30**	**k=50**
base	363	373	375
targeted	332	343	355
robust	292	329	350

Table 14: Targeted attack on Russian texts, USE threshold=0.84

References

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Chaudhary Vishrav, Guillaume Wenzek, Edouard Grave Francisco Guzman´, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv*, arXiv: 1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. 2001. The form is the substance: classification of genres in text. // *Proc. Human Language Technology and Knowledge Management*, P 1–8.

Aminul Huq and Mst. Tasnim Pervin. 2020. Adversarial attacks and defense on texts: A survey. *arXiv*, arXiv: 2005.14108.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. *arXiv*, arXiv: 1907.11932.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. // *COLING '94: Proc. of the 15th. International Conference on Computational Linguistics*, P 1071 – 1075, Kyoto, Japan.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. // *Proceedings of the 13th International Workshop on Semantic Evaluation*, P 829–839, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Maria Kunilovskaya and Serge Sharoff. 2019. Building functionally similar corpus resources for translation studies. // *Proc RANLP*, Varna, September.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv*, arXiv: 2004.09984.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. *arXiv*, arXiv: 2005.05909.

Philipp Petrenz and Bonnie Webber. 2010. Stable classification of text genres. *Computational Linguistics*, 34(4):285–293.

Marina Santini, Alexander Mehler, and Serge Sharoff. 2010. Riding the rough waves of genre on the web. // Alexander Mehler, Serge Sharoff, and Marina Santini, *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. // *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Serge Sharoff. 2018. Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification? *arXiv preprint arXiv:1905.05583*.

Genre	Keywords
Argument	united, nations, reconciliation, international, development, people, security, countries
Fiction	said, would, one, could, little, man, came, like, went, upon
Instruction	tap, device, screen, email, tab, select, settings, menu, contact, message
News	said, million, committee, disarmament, kongo, report, program, also, budget, democratic
Legal	shall, article, may, paragraph, court, person, order, department, party, state
Personal	church, one, like, people, could, really, congo, time, years, would
Promotion	viagra, cialis, online, writing, posted, service, levitra, business, buy, essay
Academic	system, quantum, fault, data, software, image, node, faults, application, fig
Information	committee, convention, parties, secretariat, iran, meeting, shall, mines, states, conference
Review	google, home, new, like, star, paul, one, shoes, pro, art

Table 15: Examples of English keywords extracted with tf-idf

A Examples of extracted keywords

A.1 Extraction with tf-idf

Table 15 lists the most significant tf-idf keywords. Some keywords correspond to their genres quite reasonably, for example, Fiction or Legal texts. However, most genres have fairly genre keywords, which indicates the prevalence of specific topics in the training corpus. For example, both Argument and News contain a lot of texts about international politics, many Instruction texts mostly refer to internet services or electronic devices.

A.2 Extraction with TextFooler

The Fiction and Legal texts turn out to be harder to attack. This is likely that it is because their training sets are less affected by a topical bias. In contrast, News, Information, and Review texts are more affected by topical biases, such as politics, see also the keywords in Table 15 and the most salient words in Table 17. However, the difference is that TextFooler amends more frequent English verbs, when the words chosen by the tf-df mechanism are more genre-specific. It is caused by the fact that many successful TextFooler attacks do not replace the original words to the words specific to the new genre, predicted by the classifier.

Frequently changing of morphological forms of some words causes change to the genre predicted by the classifier. It can be seen in the pairs of the most frequent replacement pairs Table 16.

Table 17 lists the most common words which were replaced with untargeted attacks for each genre for English. For some genres (Promotion, Academic, Legal), the replaced words represent the according genre. But for genres (Argument, Information), there are clear topical shifts. Many Argument keywords (people, nations, world, government) are related to politics, the Information keywords are connected with science (telescope, astronomy, energy, plants, chemical). The words on which the original words are replaced in order to change the genre prediction does not make the text look visually like the genre the classifier actually predicts.

Genre	Words
Argument	(('people', 'residents'), 14), (('have', 'be'), 13), (('have', 'has'), 12), (('world', 'worldwide'), 8), (('be', 'have'), 8), (('social', 'societal'), 8), (('do', 'know'), 7), (('children', 'infants'), 7), (('people', 'individuals'), 7), (('nuclear', 'fissile'), 7)
Fiction	(('had', 'has'), 12), (('had', 'have'), 10), (('will', 'wants'), 10), (('have', 'has'), 6), (('king', 'monarch'), 5), (('each', 'every'), 4), (('did', 'does'), 4), (('came', 'coming'), 4), (('come', 'happen'), 4), (('have', 'be'), 4)
Instruction	(('do', 'know'), 18), (('will', 'wants'), 12), (('be', 'have'), 10), (('have', 'be'), 10), (('should', 'ought'), 10), (('click', 'clicking'), 6), (('choose', 'choices'), 5), (('based', 'inspired'), 4), (('try', 'trying'), 4), (('example', 'examples'), 4)
News	(('will', 'want'), 13), (('has', 'maintains'), 7), (('has', 'have'), 6), (('be', 'have'), 5), (('will', 'wants'), 5), (('have', 'be'), 5), (('said', 'stating'), 5), (('year', 'olds'), 4), (('new', 'ny'), 4), (('week', 'days'), 4)
Legal	(('be', 'have'), 22), (('shall', 'hereof'), 18), (('shall', 'howsoever'), 11), (('terms', 'terminology'), 8), (('order', 'ordering'), 8), (('person', 'somebody'), 8), (('conditions', 'situations'), 5), (('contract', 'agreement'), 5), (('agreement', 'agreed'), 5), (('time', 'hour'), 5)
Personal	(('life', 'lives'), 6), (('do', 'know'), 5), (('think', 'suppose'), 5), (('wanted', 'want'), 5), (('felt', 'knew'), 4), (('people', 'individuals'), 3), (('started', 'begin'), 3), (('went', 'going'), 3), (('design', 'styling'), 2), (('so', 'because'), 2)
Promotion	(('be', 'have'), 6), (('new', 'ny'), 5), (('business', 'commerce'), 5), (('company', 'corporation'), 5), (('have', 'be'), 5), (('products', 'byproducts'), 4), (('opportunity', 'opportunities'), 4), (('help', 'aid'), 4), (('company', 'venture'), 4), (('model', 'models'), 4)
Academic	(('scattering', 'scatter'), 8), (('have', 'be'), 5), (('findings', 'confirmatory'), 3), (('mathematical', 'dynamical'), 3), (('analysis', 'analyzed'), 3), (('show', 'showcase'), 3), (('idea', 'thought'), 3), (('computation', 'computing'), 3), (('be', 'have'), 3), (('bone', 'bones'), 3)
Information	(('system', 'integrator'), 4), (('number', 'numbering'), 4), (('has', 'have'), 3), (('system', 'mechanism'), 3), (('each', 'every'), 3), (('had', 'has'), 2), (('person', 'someone'), 2), (('little', 'scant'), 2), (('astronomy', 'ephemeris'), 2), (('ehc', 'liga'), 2)
Review	(('google', 'yahoo'), 3), (('quality', 'dependability'), 2), (('review', 'reassessment'), 2), (('synth', 'synths'), 2), (('movie', 'movies'), 1), (('company', 'corporation'), 1), (('rescue', 'rescued'), 1), (('get', 'got'), 1), (('engadget', 'wired'), 1), (('users', 'irc'), 1)

Table 16: English word pairs amended with untargeted attack

Genre	Words
Argument	people, have, children, nations, nuclear, world, human, government, be, many
Fiction	had, will, have, king, each, began, think, wife, little, says
Instruction	do, should, will, click, be, have, need, use, mailbox, choose
News	will, has, said, last, week, new, have, be, says, pay
Legal	shall, be, terms, act, contract, person, agreement, site, order, conditions
Personal	think, life, work, wanted, started, do, really, people, felt, years
Promotion	company, business, work, have, new, be, help, information, customer, team
Academic	scattering, have, idea, cells, analysis, shown, kinase, nuclear, equations, time
Information	system, number, has, telescope, astronomy, energy, sun, plants, chemical, each
Review	review, google, company, engadget, good, quality, polar, synth, movie, rescue

Table 17: English words amended with untargeted attack