**National Research University Higher School of Economics**
**Faculty of Computer Science**
**Programme 'Master of Data Science'**

**MASTER'S THESIS**

# Adversarial attacks on NLP models

**Student** Mansurov Ilya

**Supervisor:** Sarkisyan Veronica

**Moscow, 2021**

## Abstract

Despite impressive advances in NLP in recent years, the technology still needs improvements, such as making models more robust and interpretable. In my work, I will study several types of adversarial attacks on a text classification model. Attacks that distort text at the character level (imitation of typos), replace words with synonyms based on embeddings and masked language models. The aim of the work is to test the applicability of methods to Russian language and compare the results with the English-language attacks.

## Introduction

The first mentions of adversarial attack refer to attacks on computer vision models; later, a number of works have appeared that study attacks on NLP models. Adversarial examples are examples practically indistinguishable from the original ones, but forcing the attacked model to give an incorrect answer. Attacks on the CV model usually consist of adding noise to the image. The nature of attacks on textual models is more complex due to the discreteness of the space of input parameters and the need for grammatical correctness and consistency in meaning between adversarial and the original data. With the help of adversarial attack, we can identify vulnerabilities in models, expand training data to achieve better robustness of the model, and such attacks can bring insight into the interpretability of the model.

## Background

The first articles suggesting an adversarial attack used the simplest attack methods - symbolic distortion (Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou). Such attacks are easy to implement, but they are not very effective and, generally, are easily recognized by humans. In later works, researchers proposed word-level perturbation (Li, J., Ji, S., Du, T., Li, B., and Wang, T) - replacing words using synonyms, this method increases the effectiveness of attacks, however, is still often human-discoverable because the synonym substitution is based on the proximity of word embeddings and does not take context into account.

Recent works suggest using the masked language model to replace insertion or deletion of words in a context-sensitive manner (Siddhant Garg, Goutham Ramakrishnan); currently, this type of attack gives the best results

## Methodology

### Datasets, task and models

In the study, I wanted to compare the work of attacks on English and Russian-language datasets, so I chose two English-language and two Russian-language datasets of similar topics. One common domain dataset in each language, these are datasets with news headings ag_news and lenta_ru. And for one specific domain dataset, unfortunately, it is quite difficult to find special data, for example, medical or pharmacological, in the public domain, especially in Russian, therefore, film reviews from rotten_tomatoes and kinopoisk were chosen as the specific domain. News datasets contain 4-5 classes, reviews for 2 classes, classes are balanced.

| **Dataset** | **Language** | **Domain** | **Classes** | **Average num. words** |
|---|---|---|---|---|
| ag_news https://huggingface.co/datasets/ag_news | English | News | 4 | 42 |
| rotten_tomatoes https://huggingface.co/datasets/rotten_tomatoes | English | Movie reviews | 2 | 20 |
| lenta_ru https://huggingface.co/datasets/zloelias/lenta-ru | Russian | News | 5 | 182 |
| kinopoisk_reviews https://huggingface.co/datasets/zloelias/kinopoisk-reviews | Russian | Movie reviews | 2 | 330 |

Table 1. Datasets

As the attack target I chose text classification models, classification by rubrics for ag_news and lenta_ru datasets, and qualification for review sentiment is positive or negative for rotten_tomatoes and kinopoisk datasets. For English-language datasets, I used pre-trained classification models based on bert-base-uncased, provided as examples in the TextAttack framework. For Russian-language datasets, I chose cointegrated/rubert-tiny2, then models were fine-tuned for the classification problem on selected datasets.

For the job, I took the TextAttack framework, designed to standardize attacks and facilitate research, providing a handy tool for constructing, measuring, and ensuring reproducibility of experiments. It's design allows in a few lines of code to easily construct new attacks from combinations of novel and existing components (John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, Yanjun Qi). Attacks are built of four components.

- **Goal function.** Determines whether the attack was successful in terms of the output of the model being attacked. For example untargeted classification, minimum BLEU score
- **Constraints.** A set of restrictions imposed on perturbation. For example, the proportion of perturbed words, minimum sentence encoding cosine similarity, part-of-speech consistency
- **Transformations.** Defines a set of transformations applied to sources. For example character permutation, masked language model word swap
- **Search method.** Algorithm for choosing the most successful set of transformations. For example greedy search, beam search

Thus, the problem of the attack is reduced to combinatorial search within all potential transformations to find a sequence of transformations that generate a successful adversarial example.

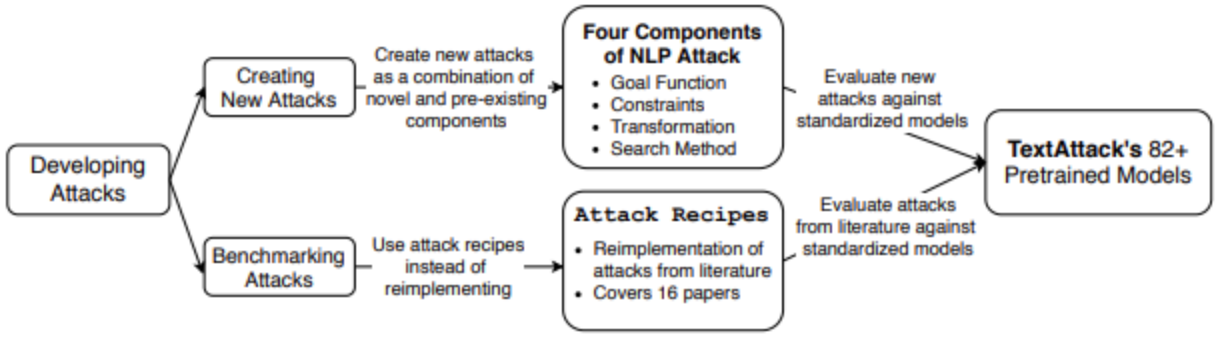


Figure 1. Developing attacks with TextAttack [(John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, Yanjun Qi)]

The attacks were carried out on test datasets. The same constraints were installed for all attacks.

- Cosine distance between the original and perturbed texts greater than threshold value of 0.8, calculated using the Universal Sentence Encoder for English texts and Multilingual Universal Sentence Encoder for Russian [(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, Ray Kurzweil)]
- Same part-of-speech word replacements
- Do not perturbe stopwords (nltk)
- Fraction of possible perturbed words within the range from 0.1 to 0.9

Accuracy was used as the metric. An increase in the proportion of possible perturbed words increased the power of the attack and significantly affected the degradation of accuracy; however, the presence of other constraints did not allow the number of really perturbed words to reach the indicated values. Due to this, accuracy did not fall to 0, but to some threshold values. As is often the case, the framework and the main attacks were developed for the English-language model, so some components had to be improved. I have improved the following components:

- Character level attack. Added transformation imitating typos in Russian text typed on the QWERTY keyboard

- Character level attack. Added transformation to replace Cyrillic characters with homoglyphs
- Fixed a bug in the MultilingualUniversalSentenceEncoder module
- Alternative word embeddings was used via the standard module

## Experiments

### Character level attack

This attack is a combination of skipping letters, permutation of adjacent letters, inserting extra characters, replacing letters that simulates typos according to keyboard layouts (qwerty, йцукен), and replacing characters with homoglyphs (O -> 0, З -> 3, etc)

| Dataset | Original Accuracy | Best Attack Accuracy |
|---|---:|---:|
| ag_news | 0.965 | 0.690 |
| kinopoisk | 0.950 | 0.470 |
| lenta | 0.980 | 0.810 |
| rotten_tomatoes | 0.840 | 0.188 |

Table 2. Character level attack results

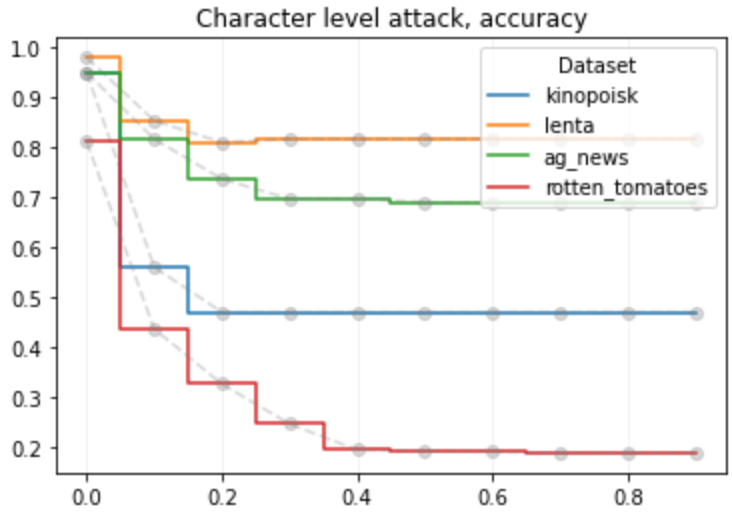


Figure 2. Character level attack results

| Original text | Perturbed text |
|---|---|
| [[Потрясающий]] фильм. Легкий, не нагружающий, но [[вместе]] с тем абсолютно ненаигранный. [[Жизненные]] ситуации, естественные для девочек-подростков поступки. Красивые актеры и блестящий юмор. [[Молодцы]] проказницы!<br><br>Positive | [[Потрясающий]] фильм. Легкий, не нагружающий, но [[вместе]] с тем абсолютно ненаигранный. [[Жизеенные]] ситуации, естественные для девочек-подростков поступки. Красивые актеры и блестящий юмор. [[Молодцы]] проказницы!<br><br>Negative |
| 'Место встречи изменить нельзя' - [[чушь]] собачья, написанная двумя 'мальчиками постаревшими в своей песочнице' и совершенно не знающими того послевоенного воровского мира и, ещё более не простого, официального общества и общественного строя. Фильм спасли своим мастерством и популярностью великие актёры.<br><br>Negative | 'Место встречи изменить нельзя' - [[чущь]] собачья, написанная двумя 'мальчиками постаревшими в своей песочнице' и совершенно не знающими того послевоенного воровского мира и, ещё более не простого, официального общества и общественного строя. Фильм спасли своим мастерством и популярностью великие актёры.<br><br>Positive |
| Владельцы японских [[мобильников]] смогут найти на карте ближайший общественный туалет, пишет The Daily Telegraph. Соответствующее приложение для сотовых телефонов разработало японское подразделение компании Access Co. B | Владельцы японских [[мобильников]] смогут найти на карте ближайший общественный туалет, пишет The Daily Telegraph. Соответствующее приложение для сотовых телефонов разработало японское подразделение компании Access Co. B |

| | |
|---|---|
| программу, [[которая]] [[называется]] [[Check]] A Toilet, [[заложено]] расположение всех общественных уборных Токио и [[других]] японских городов. Приложение также сообщает о дополнительном [[оборудовании]] туалетов. Например, Check A Toilet указывает, есть ли в уборной специальное место, чтобы перепеленать ребенка. Мобильные телефоны в Японии становятся все более многофункциональными. Большинство из них, например, оснащено бесконтактным электронным кошельком, который позволяет оплачивать покупки без кредитной карты, а также удобнее совершать микроплатежи.<br><br>Наука и Техника | программу, [[которая]] [[называется]] [[Chec*k*]] A Toilet, [[заложено]] расположение всех общественных уборных Токио и [[других]] японских городов. Приложение также сообщает о дополнительном [[оброудовании]] туалетов. Например, Check A Toilet указывает, есть ли в уборной специальное место, чтобы перепеленать ребенка. Мобильные телефоны в Японии становятся все более многофункциональными. Большинство из них, например, оснащено бесконтактным электронным кошельком, который позволяет оплачивать покупки без кредитной карты, а также удобнее совершать микроплатежи.<br><br>Экономика |
| Vodafone [[hires]] Citi for Cesky bid (TheDeal.com) [[TheDeal]].com - The U.K. [[mobile]] giant wants to find a way to disentagle the Czech [[wireless]] and [[fixed-line]] businesses.<br><br>Sci/Tech | Vodafone [[hires]] Citi for Cesky bid (TheDeal.com) [[TheDeɑl]].com - The U.K. [[mobi ı e]] giant wants to find a way to disentagle the Czech [[wireless]] and [[fixed- ı ine]] businesses.<br><br>Business |
| Sister of man who died in Vancouver police custody slams chief ([[Canadian]] [[Press]]) [[Canadian]] Press - VANCOUVER (CP) - The sister of a [[man]] who [[died]] after a violent confrontation with police has demanded the [[city's]] [[chief]] [[constable]] resign for [[defending]] the officer [[involved]].<br><br>World | Sister of man who died in Vancouver police custody slams chief ([[Canadian]] [[Press]]) [[Cɑnadian]] Press - VANCOUVER (CP) - The sister of a [[ᴍan]] who [[died]] after a violent confrontation with police has demanded the [[city`s]] [[chief]] [[constable]] resign for [[defeding]] the officer [[involved]].<br><br>Sci/Tech |
| the [[wonderfully]] lush morvern callar is pure punk existentialism , and ms . ramsay and her co-writer , liana dognini , have [[dramatized]] the alan warner novel , which itself felt like an answer to irvine welsh's book trainspotting .<br><br>Positive | the [[wpnderfully]] lush morvern callar is pure punk existentialism , and ms . ramsay and her co-writer , liana dognini , have [[dramaᵵized]] the alan warner novel , which itself felt like an answer to irvine welsh's book trainspotting .<br><br>Negative |
| [[brutally]] [[honest]] and [[told]] with humor and [[poignancy]] , which makes its message [[resonate]] .<br><br>Positive | [[brutɑlly]] [[honest]] and [[t]] [[old]] with humor and [[pignancy]] , which makes its message [[resonate]] .<br><br>Negative |

Table 3. Character level attack examples

### Word swap based on embeddings

This attack is an attack of replacing words with synonyms; the closest words in the space of embeddings are selected to perturbe the example. For English I used Gensim pre-trained on part of Google News dataset (about 100 billion words) 300-dimensional vectors for 3 million words and phrases (https://code.google.com/archive/p/word2vec/). For attacks on Russian-language datasets, I used 300-dimensional fasttext CBOW vectors pre-trained on the Araneum dataset (about 10 billion words) (https://rusvectores.org/en/models/).

| Dataset | Original Accuracy | Best Attack Accuracy |
|---|---|---|
| ag_news | 0.965 | 0.220 |
| kinopoisk | 0.950 | 0.145 |
| lenta | 0.980 | 0.790 |
| rotten_tomatoes | 0.840 | 0.045 |

Table 4. Word embeddings word swap attack results

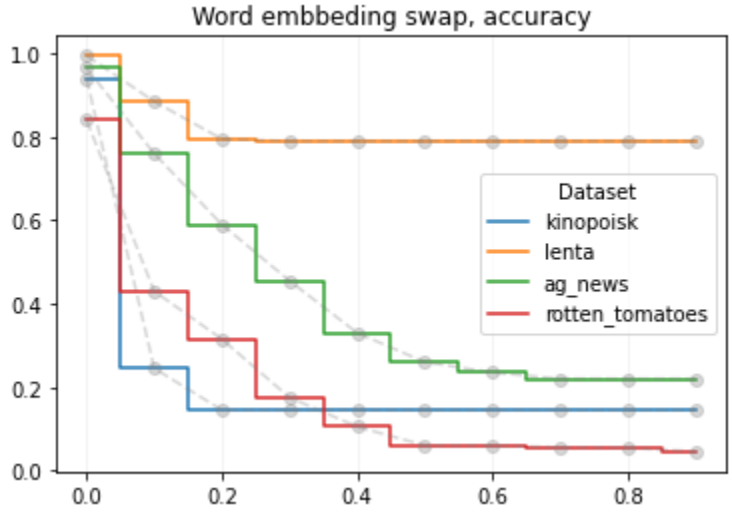


Figure 3. Word embeddings word swap attack results

| Original text | Perturbed text |
|---|---|
| [[Обалденный]] исторический фильм. Прямо захватывает, как будто попадаешь в то жестокое, даже кошмарное, время когда всё происходило. Так всё [[живо]] показано. [[Именно]] так всё и представляется, когда читаешь историю британских монархов.<br><br>Positive | [[Умопомрачительный]] исторический фильм. Прямо захватывает, как будто попадаешь в то жестокое, даже кошмарное, время когда всё происходило. Так всё [[душевно]] показано. [[Собственно]] так всё и представляется, когда читаешь историю британских монархов.<br><br>Negative |
| Участники опроса, проведенного британским Советом по культуре признали [[Ливерпуль]] "самым музыкальным городом" страны, сообщает [[NME]]. Культурной столице Европы - 2008 удалось обойти по итогам опросов город Шеффилд, также набравший значительное число голосов. Третье место в опросе занял Манчестер. Ливерпуль известен прежде всего как родина легендарной группы The [[Beatles]]. Шеффилд известен как место рождения регулярных обладателей всевозможных музыкальных наград Arctic Monkeys, ну а родом из Манчестера - братья Галлахеры из Oasis.<br><br>Культура | Участники опроса, проведенного британским Советом по культуре признали [[Тоттенхэм]] "самым музыкальным городом" страны, сообщает [[TOD]]. Культурной столице Европы - 2008 удалось обойти по итогам опросов город Шеффилд, также набравший значительное число голосов. Третье место в опросе занял Манчестер. Ливерпуль известен прежде всего как родина легендарной группы The [[Битлз]]. Шеффилд известен как место рождения регулярных обладателей всевозможных музыкальных наград Arctic Monkeys, ну а родом из Манчестера - братья Галлахеры из Oasis.<br><br>Спорт |
| Rocking the Cradle of Life When did life begin? One evidential clue stems from the fossil records in Western [[Australia]], although [[whether]] these layered sediments are biological or chemical has spawned a spirited [[debate]]. Oxford researcher, [[Nicola]] [[McLoughlin]], [[describes]] some of the issues in [[contention]].<br><br>Sci/Tech | Rocking the Cradle of Life When did life begin? One evidential clue stems from the fossil records in Western [[O]], although [[unless]] these layered sediments are biological or chemical has spawned a spirited [[discussing]]. Oxford researcher, [[Nikolaus]] [[molloy]], [[contours]] some of the issues in [[wrangle]].<br><br>Sports |
| the [[wonderfully]] lush morvern callar is pure punk existentialism , and ms . ramsay and her co-writer , liana dognini , have [[dramatized]] the alan warner novel , which itself felt like an answer to irvine welsh's book trainspotting .<br><br>Positive | the [[supremely]] lush morvern callar is pure punk existentialism , and ms . ramsay and her co-writer , liana dognini , have [[fictionalized]] the alan warner novel , which itself felt like an answer to irvine welsh's book trainspotting .<br><br>Negative |

Table 5. Word embeddings word swap attack examples

### Word swap based on masked language model

In this attack, word replacement is done using a pre-trained masked language model. Thanks to the transformer architecture, word replacement is context-sensitive, unlike the previous type of attack. This type of attack is the most recent and promising

| Dataset | Original Accuracy | Best Attack Accuracy |
|---|---|---|
| ag_news | 0.965 | 0.775 |
| kinopoisk | 0.950 | 0.370 |
| lenta | 0.980 | 0.800 |
| rotten_tomatoes | 0.840 | 0.260 |

Table 6. Masked language model word swap attack results

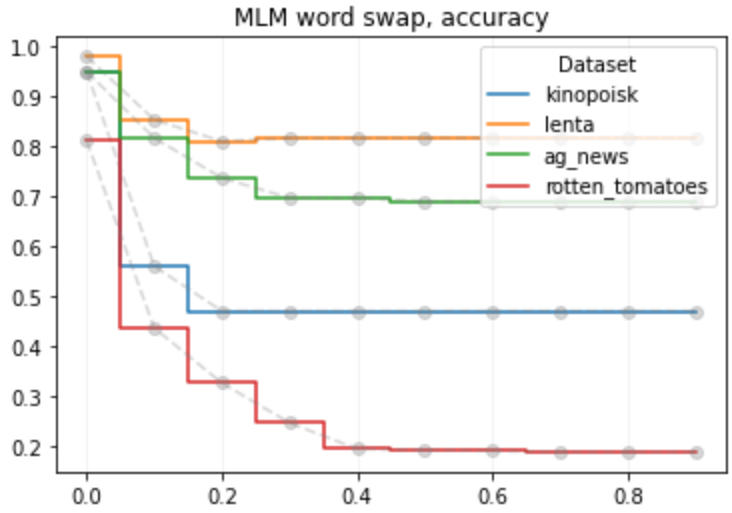


Figure 4. Masked language model word swap attack results

| Original text | Perturbed text |
|---|---|
| Один из [[лучших]] мультфильмов Бронзита. Схематичная зарисовка на тему типичных американских боевиков, очень умело их высмеивающая. Про поджог шутка меня вообще свалила.<br><br>Positive | Один из [[интересных]] мультфильмов Бронзита. Схематичная зарисовка на тему типичных американских боевиков, очень умело их высмеивающая. Про поджог шутка меня вообще свалила.<br><br>Negative |
| Главный приз Каннского [[кинофестиваля]] - "Золотую пальмовую ветвь" - получил фильм румынского режиссера Кристиана Мунгиу "Четыре месяца, три недели и два дня", сообщает AFP. [[Картина]], [[рассказывающая]] историю двух девушек, одна из которых делает криминальный аборт, считалась основным претендентом на победу на 60-м Каннском кинофестивале. Гран-при кинофестиваля - вторая по [[значимости]] награда - [[достался]] японской [[картине]] "Mogari No Mori" ("[[Лес]] скорби"), снятой женщиной-кинорежиссером Наоми Кавасе. [[Приз]] жюри - третья по [[степени]] престижности награда фестиваля в Каннах - в 2007 [[году]] вручен сразу [[двум]] картинам: "Персеполису" режиссера Марджаны Сатрапи и "Мягкому свету" ("[[Silent]] Light") - Карлоса Рейгадаса. Специальный приз в честь 60-го юбилея Каннского фестиваля [[достался]] [[режиссеру]] Гасу Ван [[Сенту]] за фильм "Параноидальный парк". В 2004 году он [[завоевал]] "[[Золотую]] пальмовую [[ветвь]]" с [[фильмом]] "Слон".<br><br>Бизнес | Главный приз Каннского [[кубка]] - "Золотую пальмовую ветвь" - получил фильм румынского режиссера Кристиана Мунгиу "Четыре месяца, три недели и два дня", сообщает AFP. [[kz]], [[представила]] историю двух девушек, одна из которых делает криминальный аборт, считалась основным претендентом на победу на 60-м Каннском кинофестивале. Гран-при кинофестиваля - вторая по [[итогам]] награда - [[приз]] японской [[марки]] "Mogari No Mori" ("[[зеленой]] скорби"), снятой женщиной-кинорежиссером Наоми Кавасе. [[Сборная]] жюри - третья по [[лучшей]] престижности награда фестиваля в Каннах - в 2007 [[был]] вручен сразу [[вторым]] картинам: "Персеполису" режиссера Марджаны Сатрапи и "Мягкому свету" ("[[Green]] Light") - Карлоса Рейгадаса. Специальный приз в честь 60-го юбилея Каннского фестиваля [[состоялась]] [[в]] Гасу Ван [[ьон]] за фильм "Параноидальный парк". В 2004 году он [[завоевала]] "[[бронзовую]] пальмовую [[ель]]" с [[рисунком]] "Слон".<br><br>Спорт |
| US fighter squadron to be deployed in South Korea next month (AFP) [[AFP]] - A [[squadron]] of US Air Force F-15E [[fighters]] based in Alaska will fly to [[South]] Korea next month for temporary deployment [[aimed]] at enhancing US firepower on the Korean [[peninsula]], US authorities said.<br><br>World | US fighter squadron to be deployed in South Korea next month (AFP) [[july]] - A [[total]] of US Air Force F-15E [[fighter]] based in Alaska will fly to [[may]] Korea next month for temporary deployment [[s]] at enhancing US firepower on the Korean [[defense]], US authorities said.<br><br>Sport |
| a [[powerful]] , chilling , and affecting [[study]] of one [[man's]] [[dying]] fall .<br><br>Positive | a [[simple]] , chilling , and affecting [[scent]] of one [[thing]] [[would]] fall .<br><br>Negative |

Table 7. Masked language model word swap attack examples

## Conclusion

The TextAttack framework and the proposed attack methods are applicable to Russian-language datasets. Despite the bugs found in the framework code during the experiments, it turned out to be a rather convenient tool for research work. The proposed attack methods showed comparable results on Russian and English datasets.

The attacks most strongly influenced the tasks of binary classification on datasets with UGC content, movie's reviews and comments, for which an 80% drop in accuracy was recorded. On multi-class classification and datasets with news, the results are more modest, but also significant - accuracy drops by 15% in the worst case. Attacks on Russian-language datasets turned out to be weaker than on English-language ones, which is most likely due to the fact that more compact and lightweight models were always chosen for pre-trained Russian models.

For production use, it makes sense to use combinations of different types of attacks in data augmentation. Character level attacks show the weakest results, but they are the fastest and require less computing resources. In addition, in some domains (chat bots, antispam) they will represent real errors made by users. This type of attack will also be useful for increasing the robustness of models in relation to intentionally distorted texts (for example, replacing Cyrillic symbols with their homoglyphs from the Latin alphabet to hide text from search engines).

## Future work

For further research, it is worth trying not only word replacements for the masked language model, but also word deletion, additional word insertion, and a combination of these approaches (Siddhant Garg, Goutham Ramakrishnan). And also investigate the work of sentence based attacks. Prepare an attack recipe for Russian-language datasets and publish it in the framework code.

## References


Javid Ebrahimi, Anyi Rao, Daniel Lowd, Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification https://arxiv.org/abs/1712.06751

Siddhant Garg, Goutham Ramakrishnan. 2020. BAE: BERT-based Adversarial Examples for Text Classification. https://arxiv.org/abs/2004.01970

Li, J., Ji, S., Du, T., Li, B., and Wang, T. 2018. TextBugger: Generating Adversarial Text Against Real-world Applications. https://arxiv.org/abs/1812.05271

Jin, D., Jin, Z., Zhou, J.T., & Szolovits, P. 2019. Is BERT Really Robust? Natural Language Attack on Text Classification and Entailment. https://arxiv.org/abs/1907.11932

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, Ray Kurzweil. 2018. Universal Sentence Encoder. https://arxiv.org/abs/1803.11175

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. https://arxiv.org/abs/2005.05909