BBAEG: Towards BERT-based Biomedical Adversarial Example Generation for Text Classification

Ishani Mondal
Microsoft Research
Lavelle Road, Bangalore, India
ishani340@gmail.com

Abstract

Healthcare predictive analytics aids medical decision-making, diagnosis prediction and drug review analysis. Therefore, prediction accuracy is an important criteria which also necessitates robust predictive language models. However, the models using deep learning have been proven vulnerable towards insignificantly perturbed input instances which are less likely to be misclassified by humans. Recent efforts of generating adversaries using rule-based synonyms and BERT-MLMs have been witnessed in general domain, but the ever-increasing biomedical literature poses unique challenges. We propose **BBAEG** (Biomedical BERT-based Adversarial Example Generation), a black-box attack algorithm for biomedical text classification, leveraging the strengths of both domain-specific synonym replacement for biomedical named entities and BERT-MLM predictions, spelling variation and number replacement. Through automatic and human evaluation on two datasets, we demonstrate that BBAEG performs stronger attack with better language fluency, semantic coherence as compared to prior work.

1 Introduction

Recent studies have exposed the importance of biomedical NLP in the well-being of human-beings, analyzing the critical process of medical decision-making. However, the dialogue managing tools targeted for medical conversations (Zhang et al., 2020), (Campillos Llanos et al., 2017), (Kazi and Kahanda, 2019) between patients and healthcare providers in assisting diagnosis may generate certain insignificant perturbations (spelling errors, paraphrasing), which when fed to the classifier to determine the type of diagnosis required/detecting adverse drug effects/drug recommendation, might provide unreasonable performance. Insignificant perturbations might also creep in from the casual language expressed in the tweets (Zilio et al., 2020). Thus, the classifier needs to be robust towards these perturbations.

Generating adversarial examples in text is challenging compared to computer vision tasks because of (i) discrete nature of input space and (ii) preservation of semantic coherence with original text. Initial works for attacking text models relied on introducing errors at the character level or manipulating words (Feng et al., 2018) to generate adversarial examples. But due to grammatical disfluency, these seem very unnatural. Some rule-based synonym replacement strategies (Alzantot et al., 2018), (Ren et al., 2019) have lead to more natural looking examples. (Jin et al., 2019) proposed TextFooler, as a baseline to generate adversaries for text classification models. But, the adversarial examples created by TextFooler rely heavily on word-embedding based word similarity replacement technique, and not overall sentence semantics. Recently, (Garg and Ramakrishnan, 2020) proposed BERT-MLM-based (Devlin et al., 2019) word replacements to create adversaries to better fit the overall context.

Despite these advancements, there is much less attention towards making robust predictions in critical domains like biomedical, which comes with its unique challenges. (Araujo et al., 2020) has proposed two types of rule-based adversarial attacks inspired by natural spelling errors and typos made by humans and synonym replacement in the biomedical domain. Some challenges include: 1) Biomedical named entities are usually *multi-word phrases* such as *colorectal adenoma*. During token replacement, we need the entire entity to be replaced, but the MLM model (token-level replacement) fails to generate correct synonym of entity fitting in the context. So, we need a BioNER+Entity Linker (Martins et al., 2019), (Mondal et al., 2019) to link entity to ontology for generating correct synonyms.

The work started when the author was a student at IIT Kharagpur, India.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5378–5384
June 6–11, 2021.

2) Due to several variations of representing medical entities such as *Type I Diabetes* could be expressed as *'Type One Diabetes'*, we explore *numeric entity expansion* strategies for generating adversaries. 3) Spelling variations (keyboard swap, modification). While we evaluate on two benchmark datasets, our method is general and is applicable for any biomedical classification datasets.

In this paper, we present **BBAEG (Biomedical BERT-based Adversarial Example Generation)**[1], a novel black-box attack algorithm for biomedical text classification task leveraging both the BERT-MLM model for non-named entity replacements combined with NER linked synonyms for named entities to better fit the overall context. In addition to replacing words with synonyms, we explore the mechanism of generating adversarial examples using *typographical variations and numeric entity modification*. Our BBAEG attack beats the existing baselines by a wide margin on both automatic and human evaluation across datasets and models. To the best of our knowledge, we are the first to introduce a novel algorithm for generating adversarial examples for biomedical text whose success attack is higher than the existing baselines like TextFooler and BAE (Garg and Ramakrishnan, 2020), (Li et al., 2020). *The overall contributions of the paper* include: **1)** We explore several challenges of biomedical adversarial example generation. **2)** We propose BBAEG, a biomedical adversarial example generation technique for text classification combining the power of several perturbation techniques. **3)** We introduce 3 type of attacks for this purpose on two biomedical text classification datasets. **4)** Through human evaluation, we show that BBAEG yields adversarial examples with improved naturalness.

Algorithm 1: BBAEG Algorithm

Input: D=[w_1, ... w_l], label = y, target classification model M
Output: Adversarial example of D = D_{adv}
1 **Initialization:** $D_{adv} \leftarrow$ **D**, Tag the entities in D, Named entities are in S_{NE} and the rest in S_{NNE} ;
2 Compute token importance $I_i \; \forall \; w_i \in D$;
3 **for** *i in descending order of I_i* **do**
4 L = { } ;
5 **if** *(w_i in S_{NE} and $(w_{i-t}..w_{i+t})$ is a NE)* **then**
6 Syns = synonyms of NE;
7 **for** *$s \in$ Syns* **do**
8 $L[s] = D_{adv[1:i-t-1]}$[s]$D_{adv[i+t+1:l]}$
9 **end for**;
10 **else if** *(w_i in S_{NNE})* **then**
11 $D_{adv} = D_{adv[1:i-1]}$[M]$D_{adv[i+1:l]}$;
12 T = top-K filtered and semantically similar tokens for $M \in D_M$;
13 **for** $t \in T$ **do**
14 $L[t] = D_{adv[1:i-1]}$[t]$D_{adv[i+1:l]}$
15 **end for**;
16 **end if**;
17 **if** $\exists \; t \in T$ *such that* $M(L[t]) \neq y$ **then**
18 **Return:** $D_{adv} \leftarrow L[t']$ where $M(L[t]) \neq y$ and $L[t']$ has maximum similarity with D
19 **else**
20 N_1 = Rotate p characters in w_i ($p \leq l$);
21 N_2 = Random insertion of symbols before/end in w_i;
22 Noise = N_1 + N_2 ;
23 **for** $t \in Noise$ **do**
24 $L[t] = D_{adv[1:i-1]}$[t]$D_{adv[i+1:l]}$
25 **end for**;
26 **if** $\exists \; t \in T$ *such that* $M(L[t]) \neq y$ **then**
27 **Return:** $D_{adv} \leftarrow L[t']$ where $M(L[t]) \neq y$ and $L[t']$ has maximum similarity with D
28 **else if** *w_i contains numeric entity* **then**
29 t = Replace w_i by $num2words$;
30 $L[t] = D_{adv[1:i-1]}$[t]$D_{adv[i+1:l]}$;
31 **Return:** $D_{adv} \leftarrow L[t]$ if $M(L[t]) \neq y$
32 **else**
33 **Return:** $D_{adv} \leftarrow L[t']$ where $L[t']$ causes max reduction in y probability
34 **end if**;
35 **end if**;
36 **end for**;
37 **Return** $D_{adv} \leftarrow$ None

2 Methodology

Problem Definition: Given a set of n inputs (D, Y) = [(D_1, y_1), . . .(D_n, y_n)] and a trained classifier $M : D \rightarrow Y$, we assume the soft-label black-box setting where the attacker can only query the classifier for output probabilities on a given input, and has no access to the model parameters, gradients or training data. For an input of length l consisting of words w_i, where $1 \leq i \leq l$, $(D_i = [w_1, ..., w_l], y)$, we want to generate an adversarial example D_{adv} such that $M(D_{adv}) \neq y$. We would like D_{adv} to be grammatically correct, semantically similar to D ($Sim(D, D_{adv}) \geq \alpha$), where α denotes the similarity threshold.

BBAEG Algorithm:
Our proposed **BBAEG algorithm** consists of four steps: **1)** Tagging the biomedical entities on D and prepare two classes NE (named entities) and Non-NE (non-named entities) **2)** Ranking the important words for perturbation **3)** Choosing perturbation schemes **4)** Final adversaries generation.

1) Named Entity Tagging: For each input instance D_i (Line 1 in Algorithm), we apply

[1]https://github.com/Ishani-Mondal/BBAEG.git

sciSpacy[2] with *en-ner-bc5cdr-md* to extract biomedical named entities (drugs and diseases), followed by its Entity Linker (Drugs to DrugBank (Wishart et al., 2017), Disease to MESH[3])). After linking the NE to respective ontologies, we use pyMeshSim[4] (for disease) and DrugBank (for drugs) to obtain synonyms. In each D_i of size l $(w_1, w_2, ...[w_i...w_{i+2}], ...w_l)$, multi-word expressions $(w_i...w_{i+2})$ are named entities. We put them in Named Entities Set (S_{NE}) and other words in non-Named Entity set (S_{NNE}).

2) Ranking of important words: We estimate token importance I_i of each $w_i \in D$, by deleting w_i from D and computing the decrease in probability of predicting the correct label y (Line 2), similar to (Jin et al., 2019). Thus, we receive a set for each token which contains the tokens in decreasing order of their importance.

3) Choosing perturbation schemes: Consider the input D_i, we describe a *sieve-based approach* of perturbing D_i. Sieves are ordered by precision, with the most precise sieve appearing first.

Sieve 1 : In the first sieve, we propose to alter the synonyms of the tokens in S_{NE} (Line 5-9) using Ontology linking and the words in S_{NNE} (Line 10-15) using BERT-MLM predicted tokens. This stems from the fact that synonym replacement of the non-named entities using **BERT-MLM** generates reasonable predictions considering the surrounding context (Garg and Ramakrishnan, 2020). If the token is a part of S_{NE}, replace them with the domain-specific synonyms one by one, but if the token is part of S_{NNE}, then replace those words by the top-K BERT-MLM predictions. To achieve high semantic similarity with the original text, we filter the set of top K tokens (K is a pre-defined constant) (Line 12) predicted by BERT-MLM for the masked token, using a Sentence-Transformer (Reimers and Gurevych, 2019) based sentence similarity scorer. Additionally, we filter out predicted tokens that do not belong to the same part of speech as original token. If this sieve generates adversaries for D_i, then D_{adv} is being returned.

Sieve 2: (Line 20-28) If the first sieve does not generate adversary, we introduce two typographical noise in the input 1) **Spelling Noise-N1:** Rotating random p characters (Line 20) 2) **Spelling Noise-N2:** insertion of symbols to the beginning or end (Line 21). If this sieve generates adversaries for D_i, then D_{adv} is being returned.

Sieve 3: (Line 29-31) If Sieve 2 does not generate adversary, we replace the numeric entities by expanding the numeric digit. For example: *PMD1* can be rewritten as PMD One, *Covid19* as *Covid nineteen*. If this sieve generates adversaries for D_i, then D_{adv} is being returned.

4) Final adversaries generation: For each of the three sieves, among all the winning adversaries, the one which is the most similar to original text as measured by (Reimers and Gurevych, 2019) is returned. If the sieves do not generate adversaries, we return the perturbed example which causes maximum reduction in the probability of output.

3 Experimental setup

Datasets and Experimental Details: We evaluate BBAEG on two different biomedical text classification datasets: 1) Adverse Drug Event (ADE) Detection (Gurulingappa et al., 2012) and 2) Twitter ADE dataset (Rosenthal et al., 2017) for the task of classifying whether the sentence contains mention of ADE (binary).

We use 6 classification models as M: Hierarchical Attention Model (Yang et al., 2016), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BioBERT (Lee et al., 2019), Clinical-BERT (Huang et al., 2019), SciBERT (Beltagy et al., 2019). We fine-tune these models on the training data (of each corpus) using Adam Optimizer (Kingma and Ba, 2015) with learning rate of 0.00002, 10 epochs and perform adversarial attack on the test data. For the BBAEG non-NER synonym attacks, we use BERT-base-uncased MLM to predict the masked tokens. We consider top K=10 synonyms from the BERT-MLM predictions and set threshold α of 0.75 for cosine similarity between (Reimers and Gurevych, 2019) embeddings of the adversarial and input text, we set p=2 characters for rotation to introduce noise in input. For more details refer to the appendix.

[2]https://allenai.github.io/scispacy/
[3]https://meshb.nlm.nih.gov/
[4]https://github.com/luozhhub/pyMeSHSim

	Twitter ADE Corpus			ADE		
	Before-attack	**After-attack**	**%**	**Before-attack**	**After-attack**	**%**
HAN-TF	0.80	**0.33**	0.10	0.83	0.46	0.09
HAN-BAE	0.80	0.35	0.08	0.83	0.43	**0.06**
HAN-Ours	0.80	0.36	**0.05**	0.83	**0.31**	0.11
BERT-base-TF	0.83	0.52	**0.12**	0.85	0.59	**0.11**
BERT-base-BAE	0.83	0.50	0.16	0.85	0.60	0.15
BERT-base-BBAEG	0.83	**0.44**	**0.12**	0.85	**0.54**	0.13
RoBERTa-base-TF	0.82	0.66	0.26	0.86	0.75	0.28
RoBERTa-base-BAE	0.82	0.63	0.23	0.86	0.74	0.24
RoBERTa-base-BBAEG	0.82	**0.57**	**0.19**	0.86	**0.70**	**0.23**
SciBERT-TF	0.85	0.45	0.11	0.88	0.53	0.13
SciBERT-BAE	0.85	0.43	0.11	0.88	0.56	0.11
SciBERT-BBAEG	0.85	**0.38**	**0.10**	0.88	**0.50**	**0.08**
BioBERT-TF	0.86	0.51	0.18	0.87	0.51	0.09
BioBERT-BAE	0.86	0.48	**0.13**	0.87	0.48	0.13
BioBERT-BBAEG	0.86	**0.37**	**0.13**	0.87	**0.45**	**0.07**
ClinicalBERT-TF	0.81	0.47	0.17	0.81	0.54	**0.15**
ClinicalBERT-BAE	0.81	0.48	**0.16**	0.81	0.58	0.22
ClinicalBERT-BBAEG	0.81	**0.46**	0.17	0.81	**0.50**	0.19

Table 1: Before-attack and after-attack accuracies of the models along with the % of perturbed words in the input space. Best attack and least % of perturbations are shown in **bold** for each dataset.

Adverse Drug Event (ADE) Corpus (Adversaries : ADE Present → ADE Not present)	
Original:	Successful challenge with **clozapine** in a history of **pulmonary eosinophilia** ailment.
BAE (Using BERT-MLM):	Successful challenge with hydrochloride in a history of pulmonary disease ailment.
BBAEG (Best Combination):	Successful challenge with clozapinum in a history of Loeffler Syndrome ailment.
Original:	A 21-year-old patient developed rhabdomyolysis during 19th week of treatment with **clozapine** for **schizophrenia**.
BBAEG (Spelling Noise-N2):	A 21-year-old patient developed rhabdomyolysis during 19th week of treatment with inoclozapine for cdschizophrenia.
BBAEG (Spelling Noise-N1):	A 21-year-old patient developed rhabdomyolysis during 19th week of treatment with clpazoine for schizoerhpnia.
BBAEG (Synonyms):	A 21-year-old patient developed rhabdomyolysis during 19th week of treatment with Clozapinum for dementia Praecox.
BBAEG (Number Replacement):	A twenty-one-year-old patient developed rhabdomyolysis during nineteenth week of treatment with clozapine for schizophrenia.

Table 2: shows the adversaries generated by BBAEG on handpicked examples from test set of ADE corpus. The different adversaries generated by baselines and BBAEG are shown. Also, the adversaries generated using different ablation of sieves [Spellings in Blue and Number in green, synonyms by attack algorithms in red] are shown.

4 Results

Automatic Evaluation Results: We examine the success of adversarial attack using two criteria: **(1) Performance Drop** (Adrop): Difference between original (accuracy on original test set) and after-attack accuracy (accuracy on the perturbed test set) **(2) Perturbation of input (%):** Percentage of perturbed words in adversary generated. Success of attack is directly and indirectly proportional with criteria 1 and 2 respectively.

Effectiveness: Table 1 shows the results of BBAEG attack on two datasets across all the models. During our experiments with HAN (general deep learning model), we observe that the attack is the most successful compared to BERT-variants, RoBERTa and the existing baselines, in terms of both the criteria (1 and 2). Also, using BioBERT and Sci-BERT (35-45% and 40-50% accuracy drop respectively), the attack is the most successful. This stems from the fact that the vocabularies used in the datasets have already been explored during pre-training by the contextual embeddings, thus more sensitive towards small perturbations. Moreover, it has been clearly observed that unlike BERT and HAN, RoBERTa is very less susceptible to adversarial attacks (10-20% accuracy drop), perturbing 20-25% words in the input space. We also observe that BERT-MLM-based synonym replacement techniques for non-NER, combined with *multi-word NER* synonym replacement using entity linking outperforms TextFooler(TF) and BAE-based approaches in terms of accuracy drop.

Ablation Analysis: In Table 3, we perform an ablation analysis on the different perturbation schemes and the effect of the attack using each of the sieves by making use of two fine-tuned contextual embedding model as the target model for ADE classification. *Synonym replacement (S1)* (average 35% accuracy drop) and *character rotation (S2-1)* (average 38% accuracy drop) seems to be the most promising approach for success attacks on biomedical text classification. Moreover, we conduct a deeper analysis to gain an insight of how much the synonyms of *NER vs Non-NER entities* contribute towards prediction change. We have found that the multi-

	Twitter ADE Accuracy Drop (Semantic Similarity)	ADE Accuracy Drop (Semantic Similarity)
BioBERT-BBAEG (best variation)	0.43 (0.893)	0.42 (0.906)
- w/o Synonym Replacement (S1)	0.39 (0.899)	0.40 (0.919)
- w/o Spelling Noise N1 (S2-1)	0.37 (0.901)	0.35 (0.912)
- w/o Spelling Noise N2 (S2-2)	0.34 (0.913)	0.31 (0.891))
- w/o Number Replacement (S3)	0.30 (0.920)	0.27 (0.915)
SciBERT-BBAEG (best variation)	0.45 (0.879)	0.38 (0.881)
- w/o Synonym Replacement (S1)	0.42 (0.901)	0.35 (0.912)
- w/o Spelling Noise N1 (S2-1)	0.39 (0.915)	0.36 (0.901)
- w/o Spelling Noise N2 (S2-2)	0.31 (0.891)	0.31 (0.847)
- w/o Number Replacement (S3)	0.32 (0.911)	0.36 (0.903)

Table 3: Ablation analysis of the sieves (S1-S3) on accuracy drop and average semantic similarities between adversaries and original text.

	Twitter ADE		ADE	
	Accuracy	Naturalness	Accuracy	Naturalness
TextFooler (TF)	0.85	3.78	0.78	3.55
BAE Algorithm	0.88	3.95	0.84	3.89
BBAEG (Our Method)	**0.94**	**4.23**	**0.90**	**4.56**

Table 4: Human Evaluation on both the datasets.

word NERs during replacement generates natural-looking examples (compared to MLM-based entity replacement such as *pulmonary eosinophillia* is replaced by *Loeffler Syndrome* (for BBAEG) by normalizing to MESH vocabulary, while replaced by *disease* in BAE predictions as shown in Table 2 and they seem very unnatural. This proves that high semantic similarity does not always ensure generation of proper grammatical adversaries.

Human Evaluation: Apart from automatic evaluation, we also perform human evaluation of our BBAEG attacks on the BERT classifier. We perform similar kind of human evaluation by two biomedical domain-experts on randomly selected 100 generated adversarial examples (from each of the different attack algorithms) on each of the two datasets. For each sample, 50 annotations were collected. Similar setup was performed by (Garg and Ramakrishnan, 2020) during evaluation. The main two criteria for evaluation of the perturbed samples are as follows:

1) Naturalness : How much the adversaries generated is semantically similar to the original text content, preserving grammatical correctness on Likert Scale (1-5)? To evaluate the naturalness of the adversarial examples, we first present the annotators with 50 different set of original data samples to understand data distribution.

2) Accuracy of generated instances: on the binary classification of presence of *Adverse Drug Reaction (ADR)* on the adversarial examples. We enumerate the average scores of two annotators (for TextFooler (TF), BAE and our BBAEG) and present those in Table 4.

During ablation analysis, we observe that the synonym replaced perturbed samples looked more natural to the human evaluators compared to the spelling perturbed samples and number replaced entities. When considered jointly, the number replaced and synonym replaced samples seemed more natural to the annotators compared to spelling perturbed samples. This arises due to the fact that the number replaced entities when thrown to the annotators they could easily interpret the meaning correctly when given in combination with the original sample. For instance, in the examples shown in table 2, the number replaced samples (*21-year old* $\rightarrow$ *twenty-one-year old*) look more natural and easily interpretable compared to spelling perturbed samples (*clozapine* $\rightarrow$ *clpazoine*).

5 Conclusion and Future Work

In this paper, we propose a new technique for generating adversarial examples combining contextual perturbations based on BERT-MLM, synonym replacement of biomedical entities, typographical errors and numeric entity expansion. We explore several classification models to demonstrate the efficacy of our method. Experiments conducted on two benchmark biomedical datasets demonstrate the strength and effectiveness of our attack. As a future work, we would like to explore more about retraining the models with the perturbed samples in order to improve model robustness.

Acknowledgement

The author would like to thank the annotators for hard work, and also the anonymous reviewers for their insightful comments and feedback.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Vladimir Araujo, Andres Carvallo, C. Aspillaga, and Denis Parra. 2020. On adversarial examples for biomedical nlp tasks. *ArXiv*, abs/2004.11157.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Leonardo Campillos Llanos, Sophie Rosset, and Pierre Zweigenbaum. 2017. Automatic classification of doctor-patient questions for a virtual patient record query task. In *BioNLP 2017*, pages 333–341, Vancouver, Canada,. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Mohit Iyyer, Pedro Rodriguez, Alvin Grissom II, and Jordan L. Boyd-Graber. 2018. Right answer for the wrong reason: Discovery and mitigation. *CoRR*, abs/1804.07781.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. In *EMNLP*.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885 – 892. Text Mining and Natural Language Processing in Pharmacogenomics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is BERT really robust? natural language attack on text classification and entailment. *CoRR*, abs/1907.11932.

Nazmul Kazi and Indika Kahanda. 2019. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 140–148, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

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. *CoRR*, abs/1907.11692.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical entity linking using triplet network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 95–100, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

David Wishart, Yannick Djoumbou, An Chi Guo, Elvis Lo, Ana Marcu, Jason Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, and Michael Wilson. 2017. Drugbank 5.0: A major update to the drugbank database for 2018. *Nucleic acids research*, 46.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. pages 1480–1489.

Yuanzhe Zhang, Z. Jiang, T. Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. Mie: A medical information extractor towards medical dialogues. In *ACL*.

L. Zilio, Liana Braga Paraguassu, Luis Antonio Leiva Hercules, G. Ponomarenko, Laura Berwanger, and Maria José Bocorny Finatto. 2020. A lexical simplification tool for promoting health literacy. In *READI*.