BERT Rankers are Brittle

A Study using Adversarial Document Perturbations

Wang, Yumeng; Lyu, Lijun; Anand, Avishek

DOI
10.1145/3539813.3545122

Publication date
2022

Document Version
Final published version

Published in
ICTIR 2022 - Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval

BERT Rankers are Brittle: A Study using Adversarial Document Perturbations

Yumeng Wang
L3S Research Center
Hannover, Germany
wang@l3s.de

Lijun Lyu
L3S Research Center
Hannover, Germany
lyu@l3s.de

Avishek Anand
Delft University of Technology
Delft, Holland
Avishek.Anand@tudelft.nl

ABSTRACT

Contextual ranking models based on BERT are now well established for a wide range of passage and document ranking tasks. However, the robustness of BERT-based ranking models under adversarial inputs is under-explored. In this paper, we argue that BERT-rankers are not immune to adversarial attacks targeting retrieved documents given a query. Firstly, we propose algorithms for adversarial perturbation of both highly relevant and non-relevant documents using gradient-based optimization methods. The aim of our algorithms is to add/replace a small number of tokens to a highly relevant or non-relevant document to cause a large rank demotion or promotion. Our experiments show that a small number of tokens can already result in a large change in the rank of a document. Moreover, we find that BERT-rankers heavily rely on the document start/head for relevance prediction, making the initial part of the document more susceptible to adversarial attacks. More interestingly, we find a small set of recurring adversarial words that when added to documents result in successful rank demotion/promotion of any relevant/non-relevant document respectively. Finally, our adversarial tokens also show particular topic preferences within and across datasets, exposing potential biases from BERT pre-training or downstream datasets.

CCS CONCEPTS

• **Information systems → Adversarial retrieval**.

KEYWORDS

BERT, ranking, neural networks, adversarial attack, biases

ACM Reference Format:
Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. BERT Rankers are Brittle: a Study using Adversarial Document Perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539813.3545122

1 INTRODUCTION

Adversarial examples are small deliberate perturbations to an input instance that can lead to wrong predictions. There is a growing amount of work that has shown that over-parameterized neural models can easily be fooled/attacked for a variety of machine learning tasks [5, 8, 12]. Adversarial examples not only showcase the limitations of the underlying trained model by exposing non-intuitive and unreliable results, but they also expose the potential biases of the model or training corpus.

The ability to generate adversarial examples for ranking models is of substantial interest to search engines and e-commerce websites, that are increasingly using neural ranking models. Additionally search engine optimization (SEO) companies benefit from adversarial terms that can potentially improve the ranking of any arbitrary document. Finally, adversarial examples also provide an insight into the inner workings of the models in terms of spurious correlations resulting from pre-training and fine-tuning procedures.

So far, however, the question of adversarial perturbations for text ranking models has not been addressed in detail. Existing works on document perturbations for text ranking deal predominantly with black-box attacks with limited applicability [9], human-assisted adversarial examples [2] or for interpretability of rankers [11]. In this work, we propose white-box adversarial attacks on BERT-based rankers by perturbing text documents.

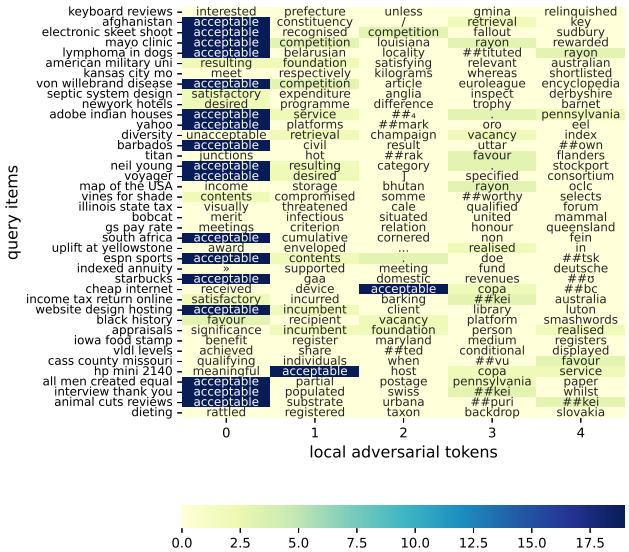

Figure 1: The five adversarial tokens added to the beginning of the highest-ranked document for 40 queries from ClueWeb09, selected by local ranking attack method to demote the document. Specific tokens frequently recur across queries. The frequency is denoted by the color.

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545122

Unlike adversarial perturbations on images where arbitrarily small changes are possible in the image space, text data is different and arguably more challenging. Due to the discrete nature of language, small changes in the input space in the form of word additions or replacements can easily cause a big difference in embedding vectors. In this paper, we define the problem of adversarial ranking attacks that generate adversarial terms/tokens that when added to a retrieved document greatly shift the rank of the document. Note that we choose to perturb long documents since it is hard to detect instead of perturbing queries.

We consider two attack scenarios – (a) where a highly relevant document is *demoted*, and (b) where a lowly ranked document is *promoted*. Additionally, we define the scope of adversarial ranking attacks on a per-query level and on an entire query workload. Corresponding to these scopes we design a *local ranking attack* and a *global ranking attack* that aim to maximize the rank shifts. Our document perturbation approaches are adapted from gradient-based token search algorithms for local [1] and the global [13] attack. Accessing the parameters of a real-life ranking model is hard. Instead, we choose the BERT-style ranker as the victim, because it's been intensively applied in many text-related tasks including ranking for its superior success. Such big gain of performance usually comes at the cost of vulnerability against adversarial attack, as well as the lack of interpretability. Figure 1 showcases the five adversarial tokens selected by the *local ranking attack* to demote the documents for 40 anecdotal queries. More details can be found in Section 4.3.

We conduct extensive experiments where we attack the BERT ranker on ClueWeb09 and TREC-DL datasets. Our results show that both local and global attacks can cause significant rank shifts, exposing the fragility of BERT rankers with the addition of as few as five tokens. Our further post-hoc analysis suggests the adversarial tokens from local ranking attack recur across different queries, sometimes transcending across datasets. Finally, we find a majority of adversarial tokens fall into a group of sensitive topics like *ethnics*, *diseases* or *natural disasters*, indicating potential dataset biases. The source code is publicly available [1].

2 RELATED WORK

Adversarial Attacks in NLP.

There are two major lines of attacking approaches in NLP tasks such as text classification and natural language inference. The first is without knowing model architectures and parameters, namely under the black-box setup, the attacker uses predefined heuristic rules to generate *natural* (in human perspective) substitutes of words or sentences, so that to fool the victim model. The prevalent heuristics include synonym replacement [4, 10], mask-and-fill by contextualized language models [6, 7], human-in-the-loop strategy [14], etc. On the other hand, white-box attack assumes full access to model parameters. Thus the attacker can use gradient signals to guide the searching process within a few rounds of forward and backward computation [1, 16]. Due to the discrete nature of human languages, gradient-based approaches are efficient yet lack of fluency compared to rule-based methods. For more related adversarial attacking methods, we point the reader to recent surveys [3, 17].

[1]https://github.com/menauwy/brittlebert

Universal Triggers. Unlike above methods, Wallace et al. [13] generate input-agnostic universal triggers using HotFlip [1] for NLP models and datasets. Triggers are tokens or phrases, injecting such trigger to any input text can mislead the victim model to a target prediction. For high-dimensional neural networks, it's unsurprising to find tokens from a large vocabulary that always result in a particular prediction. Moreover, Wallace et al. [15] also argue it's possible to plant any trigger by poisoning the training data with a small set of crafted instances containing such trigger. After retraining, the model will pick up the shortcuts induced by the trigger and make the same prediction whenever the trigger occurs in the input. We are more interested in this universal trigger because it exposes higher potential risks if a single phrase can not only impact the model prediction, but also apply to the whole dataset.

Adversarial Attacks in Text Ranking. There has been limited research regarding adversarial attacks on ranking models, especially text ranking models. Closest to our work is Raval and Verma [9], who propose a model-agnostic document-perturbation procedure for rank demotions. Our work is different in that (i) we use gradient-based method to search candidate tokens to replace and be replaced; (ii) we consider more scenarios such as document demotion and promotion for transformer models and (iii) more importantly, we focus on local as well as global adversarial tokens as an attempt to discover potential dataset and model biases.

3 ADVERSARIAL ATTACKS ON NEURAL RANKERS

We operate on the common retrieve and re-rank framework for document ranking and focus on adversarial perturbations in the re-ranking phase. Our aim is to generate adversarial examples for already trained neural ranking models by perturbing retrieved documents at user-specified positions in the document. We consider two attack scenarios: (1) **rank demotion** for high-ranked documents (e.g., ranked in top 50, since it makes no sense to demote the lowest-ranked document), and (2) **rank promotion** for low-ranked documents (e.g., the last 50 positions). The scope of an adversarial ranking attack can be either local (on a given query) or global (on an entire query workload). Towards this, we firstly envision the **local ranking attack** where we intend to perturb a retrieved document given a query. Secondly, we formulate the **global ranking attack** where we intend to generate adversarial tokens for an entire workload of queries. We describe our approach to realize both attacks in detail in the following subsections.

3.1 Local Ranking Attack

Given a query q and a text input d from dataset $\mathcal{D} = \{Q, D\}$ where $q \in Q$ and $d \in D$, the ranking model Φ computes the relevance score of d w.r.t q by $\Phi(q, d)$ and the ranking position Π_d is obtained by comparing all relevance scores of a list of pre-retrieved documents. We intend to craft an adversarial example $\hat{d}$ for d, by adding i tokens $\mathbf{x} = \{x_1, x_2, \cdots, x_i\}$ or replacing i tokens with $\mathbf{x}$ in d, so that the rank difference $|\Pi_d - \Pi_{\hat{d}}|$ is maximized. We initialize with i placeholder tokens (i.e. [MASK]) if the adversarial tokens are to be added and the position of perturbation is user-defined. We denote such perturbation by $\hat{d} = d \odot \mathbf{x}$. Specifically, we minimize

the relevance score to demote d (or maximize it to promote d):

$$\arg\min_{\mathbf{x}} \Phi(q, d \odot \mathbf{x}) \tag{1}$$

To ensure that number of document perturbations is small, we choose $i \leq 20$ in our experiments. To solve the above objective function, we adapt gradient-based token search algorithms like HotFlip [1] to find the adversarial tokens. More concretely, we compute the gradient of the relevance function and use the gradient signals to search for the tokens over the entire vocabulary $\mathcal{V}$. Due to the discrete nature of tokens, we approximate the relevance changes induced by tokens from the whole vocabulary $\mathcal{V}$ using the first-order Taylor expansion as following:

$$\arg\min_{v \in \mathcal{V}} [v - x_i]^\top \nabla_{x_i} \Phi(q, d \odot x_i) \tag{2}$$

Note that $v - x_i$ is element-wise subtraction on the embedding dimension. To search for multiple adversarial tokens, we then apply beam search to extract the top i candidate tokens similarly as [13]. The best $\mathbf{x}$ is obtained by repeating equation (2) until the relevance score decreases no more.

3.2 Global Ranking Attack

Rather than generating adversarial tokens for a particular query-document instance, in global ranking attacks, we aim to find tokens $\mathbf{x}$ that are adversarial to the entire query set. In other words, adding such tokens can promote or demote the respective highly relevant or non-relevant document for any query in the dataset. Towards this, we minimize the expected relevance score for all queries in the demotion scenario as:

$$\arg\min_{\mathbf{x}} \mathbb{E}_{(q,d)\sim\mathcal{D}}[\Phi(q, d \odot \mathbf{x})] \tag{3}$$

The global adversarial tokens are selected by using the same gradient-based search strategy as equation (2) and updating on all queries until the average relevance score decreases no more.

We choose to ensure imperceptible perturbation by the number of adversarial tokens. This sort of adversarial texts are inevitably not as natural as those generated by black-box methods, or with synonyms constrains. We take the trade-off because attacking is not the only purpose of the work, but also discovering the spurious recurrences of adversarial tokens.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets and Model This paper focuses on the widely successful BERT-based rankers for experiments. We fine-tune the BERT (bert-base-uncased) model on (i) ClueWeb09 and (ii) MSMARCO passage ranking datasets using a pairwise training loss. Specifically, we construct the input consists of a query and a document separated by the [SEP] token, with which the BERT ranker then predicts a relevance score. The ranker is trained to maximize the margin between the scores of a relevant and non-relevant input pair. We execute our local and global attacks on a query workload of 200 real-world ClueWeb09 queries. For MSMARCO, we randomly selected 1000 queries from the development set and the 200 queries from the TREC-DL test set (all denoted as TREC-DL for simplicity) to attack.

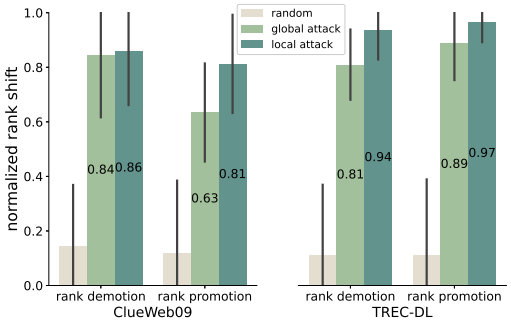

Figure 2: Attack effectiveness. Local vs. Global vs. Random

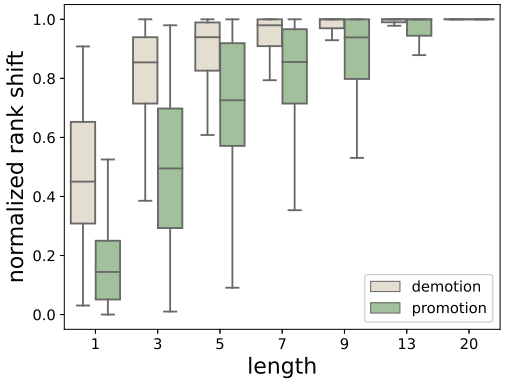

Figure 3: Impact of adversarial token lengths

For each query, we consider the top 50 as relevant and the rest as non-relevant, for each group we randomly sample 10 documents to commit rank demotion and promotion respectively. Thus, we deal with a retrieval depth of 100 for all our experiments.

Metric. We measure the absolute rank shifts normalized by the maximum shift distance (e.g., a document at rank 10 can only be demoted maximally 90 positions) as $\mathsf{NRS} = \frac{|\Pi_d - \Pi_{\hat{d}}|}{|\Pi| - \Pi_d}$ where $|\Pi|$ is the retrieval depth (i.e. $|\Pi| = 100$).

4.2 Adversary Attack Effectiveness

In this experiment, we measure the rank shifts of the relevant and non-relevant documents when a fixed small number of tokens (5 tokens for ClueWeb09 and 3 for TREC-DL) are added to the document's start (refer to Figure 2). We compare our two proposed approaches to a baseline where the same number of randomly selected tokens are used as adversarial tokens. As Figure 2 shows, for both datasets even a small number of tokens can cause significant rank shifts in comparison to the random baseline. We observe that TREC-DL shows significantly more pronounced rank shifts when compared to ClueWeb09. This observation can be attributed to the fact that TREC-DL contains shorter passages than ClueWeb09, and thus it is more sensitive to input perturbations. Specifically in TREC-DL, carefully selecting three words is sufficient (on average) to promote irrelevant passages that are ranked below 50-th position

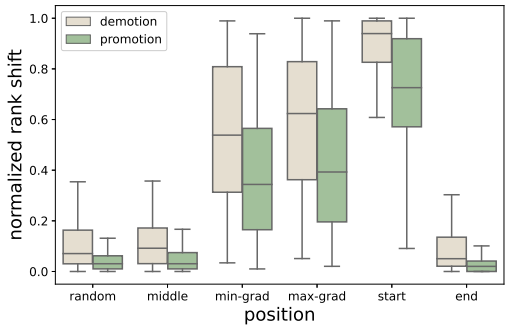

Figure 4: Impact of adversarial token positions

into the top-3. Similarly, selecting three words can demote the high-ranked passages to the bottom-six ranks (rank 94 – 100). Meanwhile, in ClueWeb09, adding five tokens to the retrieved documents results in relatively smaller yet significant rank shifts of around 80. It is also clear that the local ranking attack outperforms the global attack method. This observation is perhaps not surprising since the global adversarial tokens are updated based on the entire query workload, trading-off effectiveness for better generalization.

4.2.1 Effect of the number of adversarial tokens. A logical follow-up question is how many tokens are needed to cause a significant rank shift. Towards this, we conduct an experiment where we vary the number of allowable tokens added (1 – 20 tokens) to the start of the retrieved documents and measure the average rank shifts. In the following, whenever we mention rank shifts, we mean *average* rank shifts for ease of exposition. In Figure 3, we present the rank shifts for ClueWeb09. We omit TREC-DL due to space constraints.

The results indicate that as few as seven tokens are sufficient to cause a rank shift of above 90%. Also, increasing the number of allowable tokens monotonously increases rank shifts with noticeable marginal returns after 5 tokens. More stark is the observation that 20 tokens are sufficient to cause nearly the maximum rank shift. Note that for long ClueWeb09 documents (that typically contain more than 512-token BERT input limit), even a single token is sufficient to cause a rank shift of 46. Finally, we note that rank shifts in both directions are equally affected, indicating the fragility of BERT ranking models to minor document perturbations.

Insight 1: *We observe that both datasets show significant rank shifts by adding a small number (≤ 5) of tokens.*

4.2.2 Effect of Token Positions. Until now, we have focused on adding the tokens to the start of the document. We also conducted an ablation study for different attack positions within the document. We considered the following attack positions – (i) the **start**, (ii) the **end**, the positions where the original tokens have the (iii) **highest** and (iv) the **lowest** i (i=5) gradient scores. We also looked at (v) **random** positions and (vi) approximately the **middle** position in the document (computed from the text length divided by 2). Note that all the attack positions assume that the document fits in the 512 token-limit of the BERT input. In case document is longer than the input limit, only the truncated head of the document is considered in our experiments, a common experimental design choice for BERT-based rankers. The results for ClueWeb09 are reported in Figure 4, which indicates a significant sensitivity of ranking models to document perturbations on the start position. Namely, modifying the start of the texts causes much higher rank shifts than other positions. We hypothesize that BERT rankers might automatically associate higher relevance to the terms present at the beginning of the document. We leave the detailed analysis of this claim to future work.

Insight 2: *We observe a significant sensitivity of ranking models to document perturbations on the start position.*

4.3 Adversarial Token Analysis

Since the local attack approach generates i adversarial tokens for each query independently, we performed a post-hoc analysis of the local adversarial tokens to check for potential recurrence patterns. We analyzed both rank demotions and promotions with an expectation that seemingly disparate queries should have low overlap in adversarial tokens.

Rank Demotion. We randomly sample 40 queries from ClueWeb09 and list the 5 tokens selected to demote the top-ranked documents in Figure 1 (due to space limitations, we omit the rest of the queries). Surprisingly, we observe some particular tokens recur with a high frequency, such as the term "acceptable". In other words, the ranking model associates the term "acceptable" with a negative relevance signal irrespective of the query in the dataset. To avoid the position bias, we also show the results selected by the same algorithm to replace tokens with the highest gradient scores for the same documents in Figure 5. It suggests such negative relevance of particular terms also hold irrespective of the attacking position. On the other hand, we also find the term "acceptable" in the five global adversarial tokens for the same dataset. This suggests that both local as well as global ranking attack are capable of uncovering recurring terms like "acceptable", "competition," and "rayon", that negatively impact the relevance score of a document. Interestingly, our experiments also show that adding 5 of the most frequent tokens to all top-ranked documents causes, on average of a rank drop of 83 positions, compared to a rank drop of 79 positions by global adversarial tokens.

Rank Promotion. Unlike demoting relevant documents, adversarial tokens to promote partially relevant documents are expected to contain a high fraction of query terms. This is both understandable, expected, and in fact, serves as a sanity check for our attack algorithms. Unsurprisingly, we find that around half of the adversarial tokens are indeed query tokens. However, similar to rank demotions, we find that there are also frequent repetitions of particular tokens for rank promotions. Specifically, natural-disaster-relevant tokens such as "tornadoes" and "hurricane" are the most favored by the model for many different queries. Note that the 5 tokens selected by the global attack, are "hindusim", "earthquakes", "childbirth", "tornadoes" and "Wikipedia", showing high similarity to the frequent local adversarial tokens.

This recurrence of potentially non-relevant tokens is an important application of our study and reflects the potential biases of the BERT-ranking model. On further investigation, we find that these terms are not frequent in the corpus and do not have a high mutual

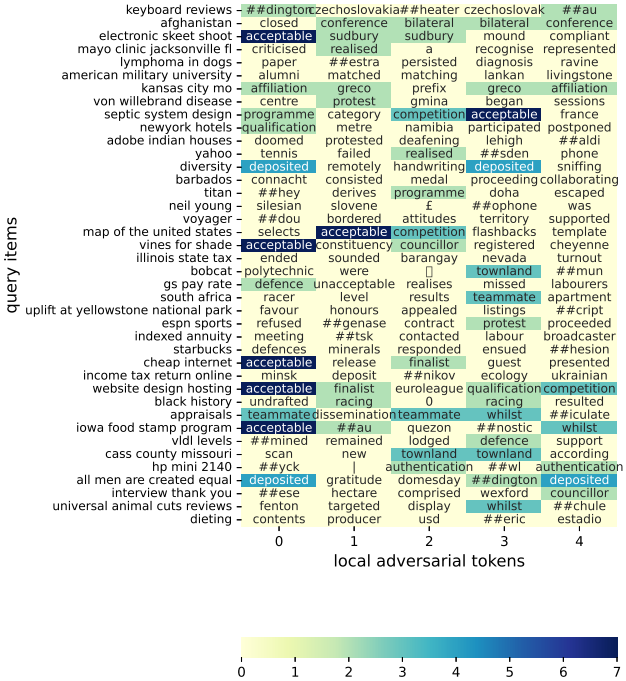

Figure 5: The five adversarial tokens replacing the tokens in documents at *max-grad* positions for 40 queries from ClueWeb09, selected by local ranking attack method to demote the document. Specific tokens frequently recur across queries. The frequency is denoted by the color.

information with the relevance label. The reason for the recurrence of the term is out of the scope of this work and is left for future investigations.

Insight 3: *We observe that local adversarial tokens recur across multiple queries and also in across datasets.*

4.4 Pre-Training Bias vs Dataset Bias

Generalizing from the token level, we also noticed the existence of additional recurrent topics from the local adversarial tokens. In Figure 6, We present the adversarial tokens added to the document start in a 2D visualization (using PCA) for both datasets. We omit rank demotion since the tokens do not show prominent topical patterns, and we consider only tokens with a minimum support of 2. Figure 6 indicates some tokens related to *nature*, *religion*, *ethnicity* and *medicine* are chose for both datasets. We omit presenting the selected tokens from diverse positions in addition to the document start since they show a similar topic preference as Figure 6. Since we use the same BERT model and fine-tune it on two different downstream datasets, we hypothesize that the model might exhibit some bias due to the pretraining process. On the other hand, it could also be possible that the datasets retain some topic preferences. Figure 6 shows that TREC-DL has a strong preference of *medicine* while for ClueWeb09, *religion* and *nature* are slightly more dominant. Via manual observation of the 200 queries from ClueWeb09, we found out 8 queries relevant to *nature*, 14 related

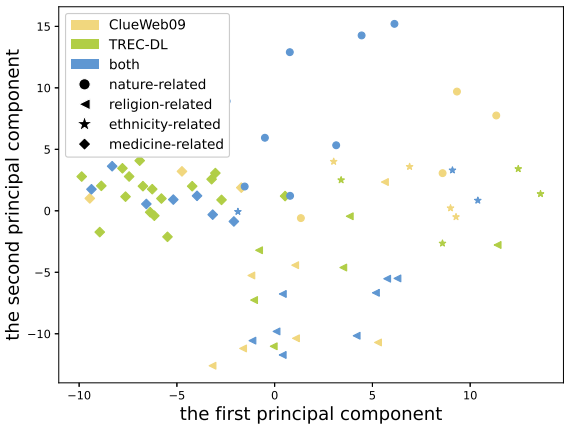

Figure 6: Frequently appeared adversarial tokens from local ranking attack for both datasets.

to *ethnicity*, the queries about diseases are as many as around 24. Additionally, some sporadic queries about cities or hotels resulted in many documents relevant to *nature* and *religion*. All of the relevant queries and documents caused the frequent occurrences of the topic specific adversarial tokens. For TREC-DL, we leave a more automatic method for topic extraction to future work due to the large training corpus.

5 CONCLUSION

In this paper, we adapt the gradient-based adversarial attack algorithm on BERT ranking models to deliberately demote and promote documents in per-query level and the entire query workload. Our experimental results suggest a minor perturbation on the text documents can shift the rank by a large margin, exposing the fragility of BERT ranking models. Moreover, we also show the BERT model is particularly sensitive to the perturbations on the document start. Finally, we conducted a post-hoc statistical analysis on the adversarial tokens and found out a set of specific tokens recurring across queries and datasets. Our findings implicate the potential biases of BERT pretraining and downstream datasets. We hope our study can raise more awareness of the risk of applying large pretrained models for ranking task. For future work, we plan to include more diverse models and datasets, endeavoring to uncover biases for large corpus and neural ranking models.

ACKNOWLEDGMENTS

This work is partially supported by DFG Project AN 996/1-1.

REFERENCES

[1] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751* (2017).

[2] Gregory Goren, Oren Kurland, Moshe Tennenholtz, and Fiana Raiber. 2018. Ranking robustness under adversarial document manipulations. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 395–404.

[3] Aminul Huq and Mst. Tasnim Pervin. 2020. Adversarial Attacks and Defense on Texts: A Survey. arXiv:2005.14108 [cs.CL]

[4] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932* 2 (2019).
[5] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236* (2016).
[6] Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502* (2020).
[7] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984* (2020).
[8] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725* (2016).
[9] Nisarg Raval and Manisha Verma. 2020. One word at a time: adversarial attacks on retrieval models. *arXiv preprint arXiv:2008.02197* (2020).
[10] 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*. 1085–1097.
[11] Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 618–628.
[12] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204* (2017).
[13] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125* (2019).
[14] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics* 7 (2019), 387–401.
[15] Eric Wallace, Tony Z. Zhao, Shi Feng, and Sameer Singh. 2021. Concealed Data Poisoning Attacks on NLP Models. In *North American Chapter of the Association for Computational Linguistics*.
[16] Jincheng Xu and Qingfeng Du. 2020. TextTricker: Loss-based and gradient-based adversarial attacks on text classification models. *Engineering Applications of Artificial Intelligence* 92 (2020), 103641. https://doi.org/10.1016/j.engappai.2020.103641
[17] Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 3 (2020), 1–41.