Attack on Unfair ToS Clause Detection: A Case Study using Universal Adversarial Triggers

Shanshan Xu and **Irina Broda** and **Rashid Haddad**
Marco Negrini and **Matthias Grabmair**
School of Computation, Information, and Technology; Technical University of Munich, Germany
{firstname.lastname}@tum.de

Abstract

Recent work has demonstrated that natural language processing techniques can support consumer protection by automatically detecting unfair clauses in the Terms of Service (ToS) Agreement. This work demonstrates that transformer-based ToS analysis systems are vulnerable to adversarial attacks. We conduct experiments attacking an unfair-clause detector with universal adversarial triggers. Experiments show that a minor perturbation of the text can considerably reduce the detection performance. Moreover, to measure the detectability of the triggers, we conduct a detailed human evaluation study by collecting both answer accuracy and response time from the participants. The results show that the naturalness of the triggers remains key to tricking readers.

1 Introduction

When using online platforms, users are asked to agree to the Terms of Service (ToS), which are often long and difficult to understand. According to (Obar and Oeldorf-Hirsch, 2020), it would take a user around 45 minutes on average to read a ToS properly. Most users accept the terms without reading them, including clauses which would be deemed unfair under consumer protection standards. Software applications that warn consumers about unfair clauses can support consumers' rights, and have been the subject of prior work (e.g., Lippi et al., 2019; Ruggeri et al., 2022). At the same time, their existence forms an incentive for drafters of ToS to formulate clauses with potentially unfair effects that bypass automated screening. In turn, developers of control systems seek to make their detectors robust against such 'adversarial attacks'. In this paper, we report on an experiment in discovering weaknesses of ToS analysis models.

Natural language processing (NLP) models for ToS analysis conduct binary classification of a given clause as fair/unfair. Previous studies have shown that state-of-the-art transformer-based classifiers are vulnerable to adversarial attacks (Belinkov and Bisk, 2017); even slight modifications to the input text (e.g., changing a few characters) can cause incorrect classifications (Ebrahimi et al., 2018). Numerous adversarial attack methods have been developed and demonstrate effective attack performance in various downstream NLP tasks such as sentiment analysis (Iyyer et al., 2018), question answering (Wang et al., 2020), machine-translation (Cheng et al., 2019) etc. One such method is the attack via a *universal adversarial trigger*, which is a sequence of tokens (words, sub-words, or characters) that can be injected into *any* text input from a dataset to mislead the victim model to a target prediction (see Table 1 for examples). These input-agnostic triggers, once generated, can be distributed to anyone, and do not need access to the victim model at the time of attack.

Adversarial attacks have, to the best of our knowledge, remained largely unaddressed in legal NLP. Our work extends the state of the art through the following contributions: (1) We conduct experiments attacking ToS unfair clause detectors trained on the public CLAUDETTE dataset with universal adversarial triggers. Our results show that a minor perturbation of the text can reduce the detection performance of transformer based models significantly. (2) We also use artifacts from the training data for universal trigger attacks. Our experiments demonstrate that such words can considerably reduce the victim model's accuracy, highlighting the potential threat of training data leakage. (3) We conduct a human evaluation study to measure the detectability of the generated triggers. The results show that suppressing sub tokens can make generated triggers more difficult to detect. [1]

[1] Our code is available at https://github.com/TUMLegalTech/ToS_attack_nllp22

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 238 - 245

ToS Clause (red = trigger)	Model Detection
Pinterest isn't liable for damages that result from a **may vote against** non-material breach of any other applicable duty of care.	Unfair → Fair
The English courts will have jurisdiction over any claim arising from **may vote against**, or related to , any use of our services.	Unfair → Fair

Table 1: The universal adversarial trigger can be injected into *any* input from a dataset to mislead the victim model. By inserting the displayed trigger can cause the trained unfair ToS detector to flip its correct unfair predictions to fair.

2 Related Work

Adversarial Attacks in NLP: Most adversarial attack methods in NLP are white-box, where the attacker has full access to the victim model (including architectures, parameters, and training data). Prevalent white-box attacks include HotFlip (Ebrahimi et al., 2018), a gradient-based method that generates adversarial examples on discrete text structure; PWWS (Ren et al., 2019), an importance-based method that substitutes words of high saliency. By contrast, black-box attacks assume no knowledge of the victim model's architectures and parameters. Example techniques include the use of generative adversarial networks (GANs) (Zhao et al., 2018) and human-in-the-loop heuristics (Wallace et al., 2019b)

Universal Triggers: Wallace et al. (2019a) generate universal attack triggers by using gradient signals to guide a search over the word embedding space. They are input-agnostic, which makes them more threatening in real-world scenarios. Despite being successful in confusing classification systems, universal triggers are often unnatural and can easily be detected by human readers. Song et al. (2021) generate attack triggers that appear closer to natural text by using a pre-trained GAN. Training a GAN in the ToS domain from scratch requires large datasets and GPU resources. In this work we try to generate natural triggers by simply skipping all the subword and special tokens during the search process; and leave the development and evaluation of a ToS-GAN to future work.

3 Universal Trigger Generation

We assume a text input x and its target label y from the dataset $D = \{X, Y\}$, a trained victim classifier model f that predicts $f(x) = \hat{y}$. While in a *non-universal* targeted attack the focus is on flipping the prediction of a single text input x, our goal is to find an input-agnostic trigger t consisting of a sequence of tokens $\{w_1, w_2, \ldots, w_i\}$ such that when concatenating t with any input x from X, the victim model incorrectly predicts $f(x;t) = \tilde{y}$, where $\tilde{y} \neq \hat{y}$. Specifically, we use the following objective function:

$$\arg\min_{t} \mathbb{E}_{x \sim X}[\mathcal{L}(\tilde{y}, f(x;t))] \tag{1}$$

To solve the above objective function, we follow the approach of Wallace et al. (2019a) by utilizing the HotFlip method (Ebrahimi et al., 2018) at the token level: First, we initiate the trigger t with a sequence of i placeholder tokens (i.e., 'the'); then we compute the gradient of (1) w.r.t the trigger. Since tokens are discrete, we approximate the loss function around the current token embedding using the first-order Taylor expansion

$$\arg\min_{e'_i \in \mathcal{V}} \left[e'_i - e_{adv_i}\right]^T \nabla_{e_{adv_i}} \mathcal{L} \tag{2}$$

where $\mathcal{V}$ is the set of all token embeddings over the entire vocabulary and e_{adv_i} represent the embedding of the current trigger token.

We update the embedding for every trigger token e_{adv_i} to minimize (2). This can be efficiently computed through d-dimensional dot products, with d corresponding to the dimension of the token embeddings. For constructing the entire updated trigger, we then use beam search to evaluate the top i token candidates from (2) for each token position in the trigger t. As variable parameters, we run experiments with triggers of different lengths [3, 5, 8] and insert positions [begin, middle, end] in the input text.

4 Experiments

4.1 Dataset and the Victim Model

The CLAUDETTE dataset (Lippi et al., 2019; Ruggeri et al., 2022) consists of 100 ToS contracts (20,417 clauses) of online platforms. A clause is

deemed as unfair if it creates an unacceptable imbalance in the parties' rights and obligations, i.e., harms the user's rights or minimizes the online service's obligations. Each clause was labelled by legal experts. [2]

Following Lippi et al. 2019, we discard sentences shorter than 5 words. In order to avoid an information leak between training and testing sentences by virtue of them stemming from the same document of contracts, we split the 100 contracts randomly into 40:40:20 for training, development and testing. Table 2 in Appendix A shows the detailed statistics of each split. Notably, the CLAUDETTE has a very imbalanced class ratio of 9:1 (fair:unfair).

For the victim model, we finetune an instance of LEGAL-BERT (nlpaueb/legal-bert-base-uncased) (Chalkidis et al., 2020) on the CLAUDETTE training set. Please refer to Appendix B for details on model finetuning. It achieves overall macro F1 of 88.9%, 97.7% F1 for class fair, and 80.1% for class unfair.

4.2 Attack Results

In the following we focus on the attack scenario *fairwashing*: targeted attacks that flip *unfair* predictions to *fair*. We apply the universal attack trigger algorithm on the development set and report the attack performance on the test set. The generated triggers can considerably degrade the victim model's performance. For instance, inserting the trigger of token length 8 "##purchased another opponent shall testify unless actuarial opponent" in the middle of the sentence can decrease the model's accuracy from 80.1% to 16.9%. However, we observe that triggers often contain special tokens or subwords, such as '[SEP]' or '##purchased', which makes them easily detectable for human readers. Inspired by Wang, 2022, we facilitate the generation of natural triggers by simply skipping all subwords and special tokens during the search (hereafter we denote this approach as mode 'no_subword' for simplicity). Although slightly less effective than the original triggers (Table 4 in Appendix C), the no_subword triggers are less likely to be detected by human readers (See our human evaluation study in Section 6).

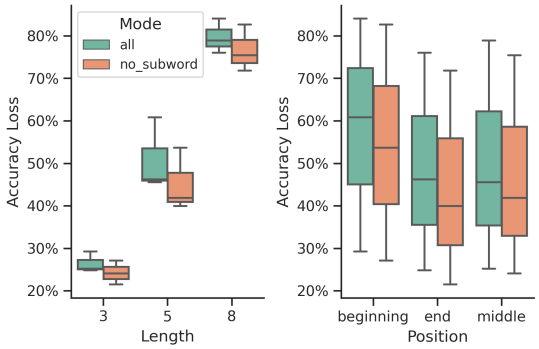

Figure 1: Accuracy loss of the victim model's detection performance when attacked by universal triggers of different insert positions and lengths. For completeness, we report the full attack results in Appendix C

We also run experiments to study the impact of trigger length, insert position, and mode (with/without subwords) on the attack's effectiveness. Figure 1 shows that increasing the token length improves attack effectiveness by a noticeable margin. The victim model's accuracy degrades by 25% to 60% using three words and by 80% to 13% with eight words. The result also indicates the victim model's sensitivity to the insert position of the triggers. These results are consistent with previous studies (Wallace et al., 2019b; Wang, 2022): Triggers are more effective when inserted at the beginning of the clause, which may be due to the transformer-based model paying more attention to the terms at the beginning of the text. These results hold across both modes. Between the modes, a higher effectiveness is consistently observed for 'all' compared with 'no_subword'. This is in line with 'no_subword' generating triggers from a subset of potential trigger tokens of 'all' mode.

5 Data Artifacts as Universal Triggers

A growing number of works have raised awareness that deep neural models may exploit spurious artifacts in the dataset and take erroneous shortcuts (McCoy et al., 2019; Xu and Markert, 2022). In this section, we experiment with using dataset artifacts as universal triggers to explore the feasibility of generating universal triggers without access to the victim model's gradient signals. Following Gururangan et al. (2018), Wallace et al. (2019a) identified the dataset artifacts as words with high pointwise mutual information (PMI) (Church and Hanks, 1990) with each label. Since the Claudette dataset has a heavily imbalanced label distribu-

[2] To measure the inter-annotation agreement, Lippi et al. (2019) have an additional test set containing 10 contracts labelled by two distinct experts, which achieve a high inner-annotator agreement with standard Cohen κ= 0.871. For details on the annotation process and the legal rational of unfair contractual clauses, please refer the original CLAUDETTE paper.

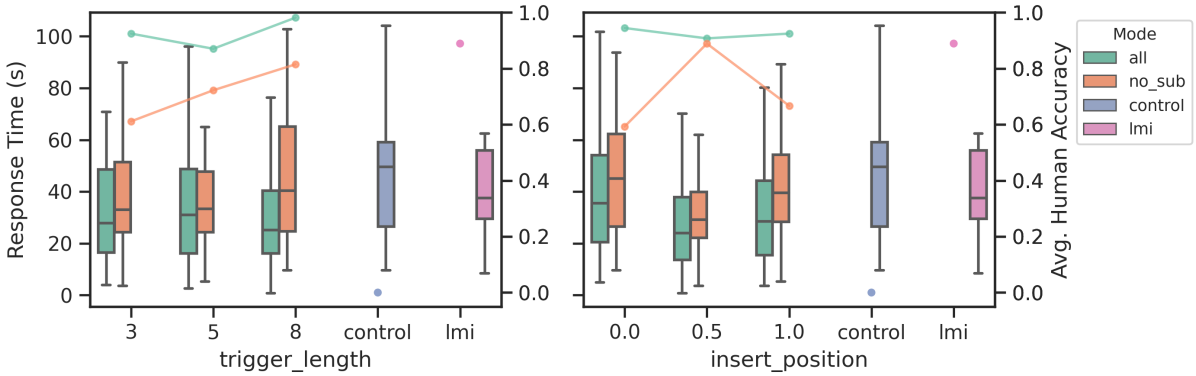

Figure 2: Human response time (box plots) and detection accuracy (line plots) for triggers of different insert positions and lengths. *Control* stands for the question where no trigger is inserted. *LMI* represents an LMI trigger of length eight inserted in the middle of the sentence. The insert positions are the following. 0.0 : beginning, 0.5 : middle, 1.0 : end.

tion, in order to prevent picking up very sparse tokens, in this work, we use local mutual information (LMI) (Schuster et al., 2019), a re-weighted version of PMI. We observe that high LMI ranked words are successful triggers. We use the 8 highest LMI words and PMI words with label fairness as triggers (hereafter LMI trigger and PMI trigger respectively, please refer to Appendix E for the list of words used); and insert them to the unfair clauses at different token positions. The LMI trigger is able to reduce the victim model's classification accuracy from 80% to around 60%; while the PMI trigger can only reduce the performance to around 76% (see Figure 3 in Appendix E). Although less successful than the universal adversarial triggers, the LMI triggers are natural and less detectable than 'all' mode triggers according to our human evaluation studies. Critically, LMI triggers are extracted by simply analyzing the training data and do not require access to the victim model. The attack effectiveness of LMI trigger highlights the potential threat of training data leakage in the NLP application.

6 Human Evaluation Study

We perform a human evaluation to study the impact of token length, insert position and mode on the triggers' detectability [3]. The task is to identify which sentence out of four candidate sentences from ToS contracts was modified. We include one question with no modified sentence as the control. In a previous study, Song et al. 2021 directly asked the human participants to rate whether the generated triggers were natural or not. However, the rating of naturalness is very abstract and varies between individuals. Inspired by studies on the detection process in psychological studies (Pandya and Macy, 1995; Yap and Balota, 2007), we assume response time (i.e., the length of time taken for a human to detect a trigger) can act as a proxy for the naturalness. To measure the human detectability of triggers, we hence collect the answer accuracy as well as the response time from the participants.

19 participants of different ages, English abilities, and legal experience were recruited from the personal network of the authors. Figure (2) demonstrates that it is consistently easier for participants to detect 'all' mode triggers than 'no_subword' mode triggers. Participants were on average 19% faster in detecting that a sentence inserted by 'all' than 'no_subword' triggers; and they find 'all' triggers with 21% higher accuracy on average. We include the LMI trigger of token length eight in the study and find its detectability is in between the 'no_subword' and 'all' triggers of the same length. The intuitive notion that participants are better at finding longer triggers generally holds with regard to detection accuracy. Nevertheless, we cannot observe a trend in the response time change, which may be due to our small sample size. Regarding the insert position, participants are the fastest in detecting triggers inserted in the middle. Further, we notice that special tokens and subwords make triggers more obvious. Qualitative, informal reports from participants indicate that 'spelling error' stuck out in a legal context. All triggers containing these tokens can be detected with more than 90%

[3] We report the details of the web application used and full instructions for the human subjects in Appendix D

accuracy, which include two 'all' triggers of length three (containing special token [SEP] or combination of subtokens '##assignabilityconsult'); and one 'no_subword' triggers inserted at position 5 (includes a bound stem 'concul'). This likely explains why these two data points do not conform to the general trend of detection accuracy.

7 Conclusion

We attacked ToS unfair clause detectors with universal adversarial triggers generated by a gradient-based algorithm as well as by simply analyzing the training data. The effectiveness of the triggers exposes the vulnerability of the transformer-based classification model, and highlights the potential threat of training data leakage. We also conducted a human evaluation to study the detectability of the triggers. The results show that the triggers are less likely to be detected if they do not include subtokens. Future work can explore ways to generate more natural triggers in the legal domain, which may even deceive readers with a formal education in law.

Limitations

Wallace et al. (2019a) reduce the detection accuracy to 1% while we can only manage to degrade it to 10%. This might be due to the imbalanced label distribution and comparatively small size of the CLAUDETTE dataset. Our human evaluation is an initial exploration with only 19 participants. Future work will focus on using crowdsourcing techniques for large survey data collection. Furthermore, we generate the 'no_subword' triggers by skipping all the tokens preceded by the double hashtag '##'. This enables us to avoid derivational morphemes and inflection suffixes but fails to exclude bound stems such as 'consul', which makes some triggers obvious to human readers. Future work can explore better ways to generate natural triggers.

Ethics Statement

The study presented here works exclusively with the publicly available CLAUDETTE dataset, which consists of the Terms of Service (ToS) Agreements of various online platforms. The techniques described in this paper are prone to misuse. However, we design this study to draw public attention to the vulnerability of the transformer-based classification model. We hope our work will help accelerate progress in detecting and defending adversarial attacks. We finetuned the victim model and generated all the triggers on Google Colab. Our models adapted pretrained language models and we did not engage in any training of such large models from scratch. We did not track computation hours.

References

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Miguel Grinberg. 2018. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.".

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147.

Abhijit S Pandya and Robert B Macy. 1995. *Pattern recognition with neural networks in C++*. CRC press.

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.

Federico Ruggeri, Francesca Lagioia, Marco Lippi, and Paolo Torroni. 2022. Detecting and explaining unfairness in consumer contracts through memory networks. *Artificial Intelligence and Law*, 30(1):59–92.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425.

Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. 2021. Universal adversarial attacks with natural triggers for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3724–3733, Online. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Boxin Wang, Hengzhi Pei, Boyuan Pan, Qian Chen, Shuohang Wang, and Bo Li. 2020. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6134–6150, Online. Association for Computational Linguistics.

Yumeng Wang. 2022. Global triggers for attacking and analyzing ranking models. Master's thesis, Hannover: Gottfried Wilhelm Leibniz Universität.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Shanshan Xu and Katja Markert. 2022. The chinese causative-passive homonymy disambiguation: an adversarial dataset for nli and a probing task. In *Proceedings of the 13th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

Melvin J Yap and David A Balota. 2007. Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):274.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

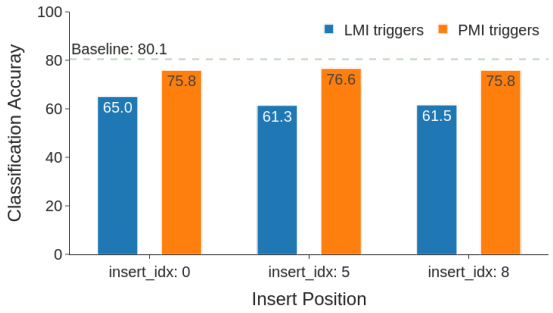

Figure 3: Attack performance of LMI trigger and PMI trigger of the different insert position.

split	# sentences	% fair label	% unfair label
train	8354	89.5%	10.5%
dev	8279	89.1%	10.9%
test	3784	89.3%	10.7%

Table 2: Statistics of the train, dev and test split of the CLAUDETTE dataset.

A Dataset Statistics

Table 2 displays the statics of the CLAUDETTE dataset.

B Finetuning the Victim Model

We used LEGAL-BERT (nlpaueb/legal-bert-base-uncased) with a sequence classification head on top from the transformers library (Wolf et al., 2019); and finetuned it on the CLAUDETTE training set. The model is fine-tuned with 5 epochs, a learning rate of 1e-5. We determine the best learning rate using grid search on the development set and use early stopping based on the development set F1 score.

C Additional Experimental Results

Table 3 demonstrates the attack results on *fair* clauses. Restricted to limited GPU resources, we generated only triggers of eight tokens which are inserted at the beginning of the sentence.

Table 4 displays the attack on *unfair* clauses with triggers of different lengths [3,5,8], insert position [begin, middle, end] and mode [original, no_subword].

D Instruction for the human evaluation study

The application is written in Python using Flask (Grinberg, 2018) and was hosted on an AWS EC2 instance. It included a landing page with a short instructions. Figure 4 is a screenshot of the web application. Following is the instruction on the landing page for the human evaluation study:

"Background information

When using online platforms, users are asked to agree to the Terms of Service (ToS). ToS documents tend to be long and difficult to understand. As a result, most users accept the terms without reading them, including clauses which would be deemed unfair under consumer protection standards. Therefore, applications that can support consumers in detecting unfair clauses would be useful. Nevertheless, studies have shown that such applications are vulnerable to adversarial attacks; even slight modifications to the input text, like inserting a few words into the text, can cause incorrect classifications. In this study, we ask you to help us detect the malicious modifications in the text.

Task instruction

You will be shown an excerpt of four sentences from a ToS contract. The task is to identify which sentence is modified. Please feel free to contact us if you have any questions. Many thanks for taking part in the study."

E LMI and PMI triggers

Figure 3 demonstrates the attack performance of LMI and PMI triggers. The 8 highest LMI ranked words that used as LMI trigger are ['information', 'payment', 'must', 'provide', 'person', 'license', 'rights', 'please']. The PMI trigger words are: ['berlin', 'attribution', 'addressing', 'android', 'sources', 'organiser', 'pc', 'unreasonable']

Trigger	Length	Position	Mode	Accuracy	Δ
Baseline	-	-	-	97.7	
not liable whenever	3	beginning	no_sub	69.8	-28.5%
terminate our convening practices if	5	beginning	no_sub	47.6	-51.2%
agree tankage bound through cloud terms 2016 laws	8	beginning	no_sub	9.0	-90.0%

Table 3: Performance of Universal Triggers on Fair Clauses

Trigger	Length	Position	Mode	Accu.	Δ
Baseline	-	-	-	80.1	
witness should testify	3	beginning	no_sub	58.4	-27.0%
may vote against	3	middle	no_sub	60.8	-24.1%
witness testified without	3	end	no_sub	62.9	-21.5%
interrelat order refusing priority where	5	beginning	no_sub	37.1	-53.7%
consul must produce his attorney	5	middle	no_sub	46.6	-41.9%
privilege to authenticate testimony groot	5	end	no_sub	48.1	-39.9%
testimony allows contracts opposing person tuber testify where	8	beginning	no_sub	13.9	-82.7%
compute another opponent shall testify unless lockbox opponent	8	middle	no_sub	19.7	-75.4%
another witness seems thus admissible scope testify usc	8	end	no_sub	22.6	-71.8%
admissible in evidence	3	beginning	all	56.7	-29.3%
##assignabilityconsult assigned	3	middle	all	59.9	-25.2%
[SEP] expert testimony	3	end	all	60.2	-24.8%
evid allowed equit testify where	5	beginning	all	31.4	-67%
[SEP] give precedence before priority	5	middle	all	43.6	-45.6%
368 hearsay witnesses may exclude	5	end	all	43.1	-46.2%
inference forbid 2028 opposing person may testify where	8	beginning	all	12.8	-84.0%
##purchased another opponent shall testify unless actuarial opponent	8	middle	all	16.9	-78.9%
assist [SEP] witness normally justifies cross admissibilitywillingness	8	end	all	19.2	-76.0%

Table 4: Performance of Universal Triggers on Unfair Label

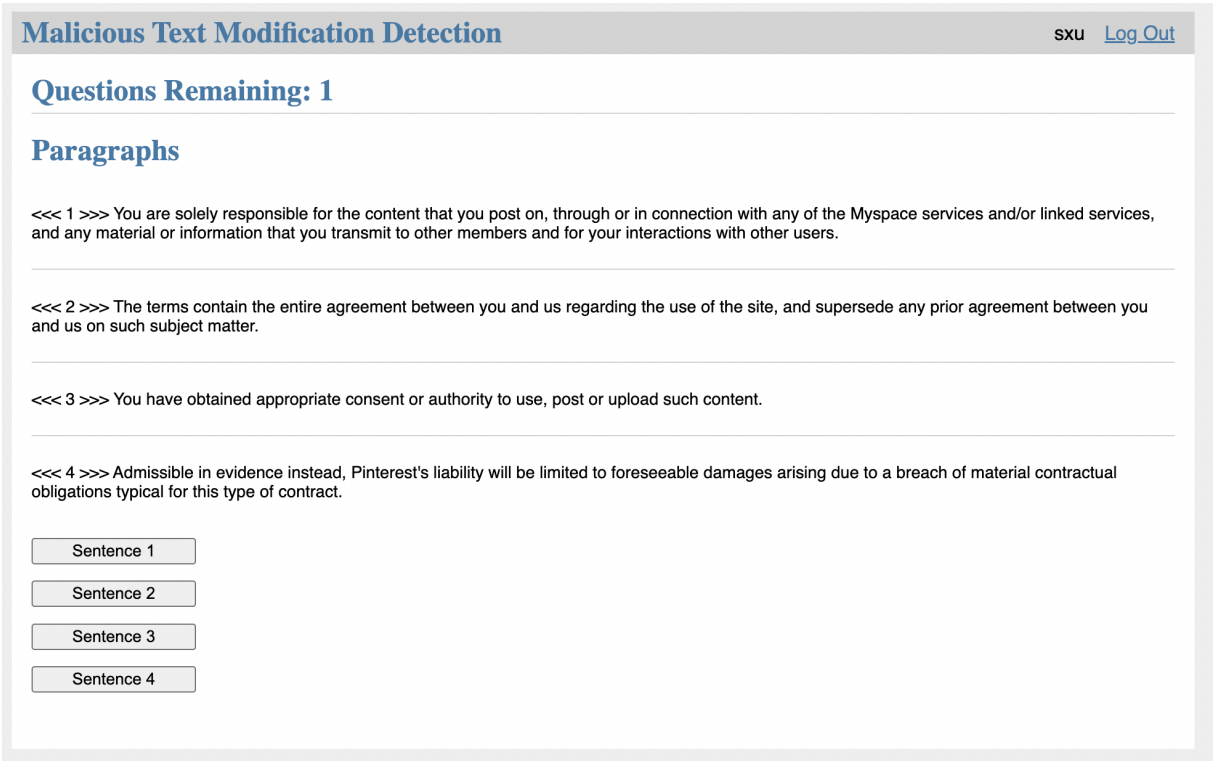

Figure 4: Screenshot of the web application for human evaluation