Enhancing Adversarial Attack Detection in NLP with Weighted Word Differential Reaction

Chenchen Ye, Yijie Lu, Yiyun Chen

Abstract

Adversarial attacks in natural language processing (NLP) present serious challenges to the development of reliable and ethical AI systems. Current techniques, like the Word-level Differential Reaction (WDR) metric, have shown potential in identifying adversarial inputs by analyzing how logits fluctuate. However, these methods typically do not consider the relative importance of the word in the context of the model's focus. This project introduces a refined detection method that uses part-of-speech (POS) tagging and attention scores to prioritize meaningful tokens and consider semantic importance of words and adjust the detector score sensitivity based on contextual importance. By focusing on these critical words with attention scores , the proposed approach aims to enhance the accuracy of detecting adversarial inputs. The improved method will be benchmarked against the original WDR technique and the WDR technique only using the part-of-speech (POS) tagging filter on the IMDb and AG News datasets.

1 Introduction

Adversarial machine learningGoodfellow et al. [2014] has become a crucial area of research in natural language processing (NLP), as AI-powered applications increasingly handle sensitive tasks where safety and ethics are of great concern. Despite impressive progress recently, these systems remain vulnerable to adversarial attacks—small, often invisible changes to input data that can cause modelsGarg and Ramakrishnan [2020] to make incorrect predictions. This vulnerability underscores the urgent need for reliable methods to detect such attacks and ensure trust in NLP systems.

One promising detection method gaining attention is the Word-level Differential Reaction (WDR) metricMosca et al. [2022]. WDR works by monitoring changes in a model's output logits (pre-activation scores) when specific words in a sentence are altered. However, WDR only measures how the removal of a word impacts the model's prediction, but it does not inherently account for the relative importance of the word in the context of the model's focus.

Our project addresses these efficiency and context sensitivity challenges by enhancing WDR through part-of-speech (POS) taggingKumawat and Jain [2015] and attention scoresVaswani et al. [2017]. By focusing on key word categories—nouns, verbs, adjectives, and adverbs—that most impact sentence meaning, we reduce the weights of unimportant words analyzed, significantly cutting down the focus of the detector on these words while maintaining high detection accuracy.

To test our approach, we conducted a series of experiments using the DistilBERT transformer modelSanh et al. [2020], chosen for its strong track record in NLP tasks and resilience to adversarial attacksMorris et al. [2020]. As the detection model, we selected XGBoostChen and Guestrin [2016], building on prior successful implementations. We evaluated our method on well-established NLP benchmarks, including the IMDbMaas et al. [2011], AGNewsPang and Lee [2005], and Rotten Tomatoes datasetsZhang et al. [2015].

Preliminary results are promising: our method not only outperforms the original WDR in terms of F1-scores and adversarial recall rates but also achieves faster processing times. These improvements suggest that our approach strikes a practical balance between detection accuracy and efficiency, making it a viable solution for real-world NLP applications.

This report is organized as follows: Section 2 reviews relevant research on adversarial ML and text-based attack detection. Section 3 explains our proposed method, including how POS tagging and attention scores enhance WDR. Section 4 covers our experimental setup, results, and key findings. We conclude in Section 5 with reflections on our work and future research directions.

2 Related Work

Adversarial text attacks involve altering a text input in a way that causes a model to make incorrect predictions while keeping the changes subtle enough to go unnoticed by humans. Common methods include changing, adding, or removing words or characters. Examples of

such attacks are DeepWordBugGao et al. [2018] and ViperEger et al. [2019]. More advanced techniques like Adversarial Training and Synonym Encoding MethodsGoodfellow et al. [2014] generate sentences that still make sense while being grammatically correct. These attacks often involve removing or inserting words, or replacing words with their synonymsRen et al. [2019b].

Adversarial detection aims to identify such attacks, alerting both models and developers. This approach was initially applied in the image domain by analyzing patterns in corresponding Shapley valuesFidel et al. [2020]. In text-based detection, basic methods like spell and grammar checks can spot some attacks, but they struggle when attacks maintain correct spelling and syntaxWang et al. [2021]. A more advanced method, FGWSMozes et al. [2020], focuses on how often words typically get replaced to detect suspicious changes.

Another detection strategy involves examining a model's output logits to spot adversarial samples. While this approach is common in image processing, Mosca et al. [2022] adapted it for NLP tasks by using logit-based metrics to detect manipulated text inputs.

3 Methodology

The proposed methodology consists of three components: (1) The basis of Word-Level Differential Reaction(WDR); (2) the application of POS tagging [Kumawat and Jain, 2015] to preliminarily filter and select words for further examination based on their grammatical roles for WDR; (3) the application of attention scoresVaswani et al. [2017] as weights to help detector model focus on significant words on WDR with POS tagging

3.1 Word-Level Differential Reaction (WDR)

The foundation of our proposed improvement is built on the Word-Level Differential Reaction (WDR) metric, which provides an innovative method for detecting adversarial attacks in natural language processing. The WDR metric is based on the observation that the logits (output scores) of a model show distinct patterns when responding to adversarial versus original inputs, even when the changes to the input are small but meaningful. This method allows us to systematically measure how substituting individual words impacts the model's prediction confidence across different classes, providing a clear indication of potential adversarial manipulations.

To use the WDR metric, we evaluate how the target model reacts to each word in an input text. This involves assessing the effect of removing or altering each word on the model's output logits. For a specific word x_i in an input x, its WDR score is defined as follows:

$$WDR(x_i, f) = f(x \backslash x_i)_{y^*} - \max_{y \neq y^*} f(x \backslash x_i)_y$$

In this equation, $f(x \backslash x_i)_y$ refers to the logit value for class y when x_i is removed or replaced with a placeholder. The term y^* represents the true class of the input, as determined by the model without any modification to the sentence. A negative WDR score indicates that removing x_i causes the model's prediction to shift away from the original class y^*, identifying x_i as a word that might be critical for adversarial manipulation.

The adversarial detector is then trained to classify inputs as either adversarial or genuine by using the aggregated WDR scores. This involves generating adversarial examples, calculating WDR scores for both original and manipulated inputs, and using these scores as input features to train a classifier. Importantly, this detector is model-agnostic, meaning it can be applied to a variety of target classifiers without being constrained by the specific characteristics of the NLP task or the architecture of the target model.

3.2 Part-of-Speech (POS) Tagging on WDR

Part-of-Speech (POS) tagging is a fundamental task in natural language processing, where each word in a sentence is labeled with its grammatical category, or "part of speech." These categories include nouns, verbs, adjectives, adverbs, and other grammatical components that make up the structure of a language. POS tagging can be accomplished using various methods, ranging from rule-based systems that follow predefined grammatical rules to statistical approaches and, more recently, neural network-based techniques. Hidden Markov Models (HMMs) have historically been widely used for POS tagging, treating the sequence of words as observable events and the corresponding tags as hidden states to be predicted. With advancements in deep learning, however, modern approaches like RNNs, LSTMs, and transformer-based models such as BERT have significantly improved performance. These methods excel at understanding long-range dependencies and capturing the subtle context of words within a sentence, making them highly effective for POS tagging tasks.

The initial step involves selecting words based on their Part-of-Speech (POS) tags. POS tagging is applied to the input text to determine the grammatical roles of each word. By analyzing these roles, we prioritize specific words for further WDR analysis, thereby optimizing the detection process and reducing computational demands. In particular, we focus on *nouns*, *verbs*, *adjectives*, and *adverbs* as the subset of words to be used for WDR score computation. This enhancement takes advantage of linguistic principles, operating under the assumption that certain parts of speech, such as nouns and verbs, are more likely to be targets of effective adversarial manipulations.

After identifying these key words using POS tagging, we calculate WDR scores exclusively for this subset. This targeted approach not only makes the detection process more efficient but also aims to improve precision by concentrating on words that are more likely to impact the semantic meaning of the input under adversarial conditions.

To formalize this, we define an indicator function $I_T(x_i)$, where T is the set of POS tags deemed relevant for analysis:

$$I_T(x_i) = \begin{cases} 1 & \text{if POS}(x_i) \in T \\ 0 & \text{otherwise} \end{cases}$$

Using this indicator function, the updated Word-Level Differential Reaction (WDR) score for a word x_i in the input x, restricted to the selected POS tags in T, is defined as:

$$\text{WDR}_T(x_i, f) = I_T(x_i) \cdot \left(f(x \setminus x_i)_{y^*} - \max_{y \neq y^*} f(x \setminus x_i)_y \right)$$

In this equation, $f(x \setminus x_i)_y$ denotes the logit value for class y when x_i is removed or replaced with a placeholder, and y^* represents the true class of the input as determined by the model in the absence of any modifications. The indicator function $I_T(x_i)$ ensures that WDR scores are calculated only for words x_i whose POS tags belong to the selected set T.

3.3 Attention on WDR with POS Tagging

Attention scores are a fundamental concept in deep learning, particularly in attention mechanisms like those used in transformer-based models such as BERT, GPT, and others. They measure the importance of one element in a sequence relative to others in determining the output of the model. We introduce the attention scores on WDR with POS Tagging because attention scores provide insight into where the model "looks" when making predictions, which can complement WDR by highlighting the words the model deems more important. Typically, attention scores are the average scores of the last layer of attention on multiple heads.

$$A_{ij} = \text{softmax} \left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_j^T}{\sqrt{d_k}} \right)$$

Then we normalize them as follows:

$$A(x_i) = \frac{\text{Attention}(x_i)}{\sum_j \text{Attention}(x_j)}$$

In this case, we used 0.2 for other words not in T

$$I_T(x_i) = \begin{cases} 1 & \text{if POS}(x_i) \in T \\ 0.2 & \text{otherwise} \end{cases}$$

$$T = [Noun(NN), Verb(VB), Adjective(JJ), Adverb(RB)]$$

Why we do so is because The value 0.2 is chosen as a proportional weight to significantly reduce the influence of less important words while still retaining some impact. It ensures that these words do not dominate the computation (e.g., in weighted differential reactions or saliency scores), while still being present to avoid losing grammatical structure entirely. Finally, our refined WDR is listed below:

$$\text{WDR}_{T,A}(x_i, f) = I_T(x_i) \cdot A(x_i) \cdot \left(f(x \setminus x_i)_{y^*} - \max_{y \neq y^*} f(x \setminus x_i)_y \right)$$

4 Experimental Setups and Results

4.1 Datasets

To comprehensively investigate the problem, we perform experiments on three benchmark datasets. The first two datasets, IMDbMaas et al. [2011] and RTMRZhang et al. [2015], involve binary sentiment classification, where each review is labeled as either positive or negative. In contrast, the AG News dataset is a four-class classification task, requiring the categorization of news articles into one of the following topics: World, Sports, Business, or Sci/Tech.

In line with previous research, we generate adversarial text samples using four established word-substitution-based attack techniques: Probability Weighted Word Saliency (PWWS)Ren et al. [2019a], Improved Genetic Algorithm (IGA) Jia et al. [2019], TextFooler Jin et al. [2020], and BERT-based Adversarial Examples (BAE) Garg and Ramakrishnan [2020]. We ensure that our dataset is balanced, containing equal proportions of adversarial and non-adversarial examples for each task.

Briefly, PWWS employs a greedy algorithm that prioritizes word replacements based on their saliency and the prediction probability Ren et al. [2019b]. IGA generates adversarial samples by iteratively mutating sentences and selecting those that most effectively alter the model's predictions Jia et al. [2019]. TextFooler identifies critical words and replaces them with suitable alternatives to maximize misclassification rates Jin et al. [2020]. Finally, BAE leverages a BERT language model to propose contextually appropriate token substitutions Garg and Ramakrishnan [2020]. All adversarial examples are produced using the TextAttack library.

4.2 Evaluation Metrics

The performance of the proposed model will be compared to the original WDR method using the average F1-score as the main metric for detection. However, as in adversarial detection false negatives can have major consequences, we also report the recall of adversarial sentences.

4.3 Model Backbone

Following [Mosca et al., 2022], we use the transformer-based model DistilBERT [Sanh et al., 2020] as the main attacked model for the experiment, where DistilBERT is a transformer architectures are widely used in NLP applications, and shows its resilient to adversarial attacks in previous works. Also, since our method does not restrict the choice of detector architecture, we choose the best-performed detector architecture, XGBoost [Chen and Guestrin, 2016], among XGBoost, AdaBoost [Schapire], LightGBM [Ke et al., 2017], SVM [Hearst et al., 1998], Random Forest [Breiman, 2001], and Perceotron NN [Singh and Banerjee, 2019]. The detailed performance is reported in Table 1 from [Maas et al., 2011]. All models are compared on adversarial attacks generated with PWWS from IMDb samples and targeting a DistilBERT model.

Table 1: Performance comparison of different detector architectures on IMDb adversarial attacks generated with PWWS and targeting a DistilBERT transformer.

Model	F1-Score	Adv. Recall
XGBoost	**92.4**	95.2
AdaBoost	91.8	**96.0**
LightGBM	92.0	93.7
SVM	92.0	94.8
Random Forest	91.5	93.7
Perceptron NN	90.4	88.1

4.4 Main Experiments

The results, as shown in Table 2, demonstrate that our proposed method, which combines attention scores with selective POS weighting, consistently outperforms the original WDR and FGWS baselines across all datasets. For instance, in the AGNews Alzantot configuration, our method achieves an F1-score of 0.865 and an adversarial recall of 0.872, compared to WDR's 0.831 F1-score and 0.830 adversarial recall. Similarly, in the AGNews TextFooler and IMDB BAE configurations, our method achieves F1-scores of 0.945 and 0.910, respectively, surpassing WDR's 0.930 and 0.882. Even in the Rotten-Tomatoes Alzantot dataset, where the performance margin is narrower, our method still outperforms WDR with an F1-score of 0.825 compared to WDR's 0.819.

The consistent improvement highlights the effectiveness of integrating attention scores and selective POS weighting. Attention scores enhance contextual sensitivity by prioritizing tokens that are critical to model predictions, making it easier to detect adversarial manipulations targeting these high-attention tokens. Meanwhile, selective POS weighting reduces noise by focusing on semantically significant words, such as nouns, verbs, adjectives, and adverbs, which are more likely to be perturbed in adversarial attacks. Together, these components form a robust detection mechanism that improves both F1-scores and adversarial recall across diverse datasets and attack methods.

This consistent performance improvement suggests that our method is not only more effective at detecting adversarial perturbations but also generalizes well across different datasets and attack strategies, providing a significant advancement in adversarial detection for natural language processing.

Table 2: Performance comparison of adversarial detection methods with different configurations

Configuration	Weighted WDR (POS + Attention)		Selective POS Weight		WDR		FGWS	
	F1-score	Adv Recall	F1-score	Adv Recall	F1-score	Adv Recall	F1-score	Adv Recall
AGNews Alzantot	**0.865**	**0.872**	0.848	0.852	0.831	0.830	0.686	0.583
AGNews TextFooler	**0.945**	**0.920**	0.935	0.906	0.930	0.898	0.870	0.794
IMDB BAE	**0.910**	**0.918**	0.894	0.906	0.882	0.861	0.656	0.552
Rotten-Tomatoes Alzantot	**0.825**	**0.850**	0.811	0.838	0.819	0.842	0.681	0.552

4.5 Model Ablation

We further conduct an ablation study. The ablation model Selective POS model is discussed in Section 3.2, where other POS-tagged words are given a weight of 0. The results are also shown in the main table.

The selective POS weighting mechanism plays a pivotal role in reducing noise by focusing on linguistically significant tokens such as nouns, verbs, adjectives, and adverbs. By deprioritizing function words (e.g., prepositions, articles), which are less critical to semantic meaning, this approach allows the model to detect perturbations in content-rich tokens more effectively. This is particularly evident in datasets like AGNews and IMDB, where adversarial manipulations often target meaningful tokens, leading to notable improvements in adversarial recall and F1-scores.

However, there is a potential risk of overly aggressive filtering. By assigning negligible importance to function words and structural tokens, the model might overlook important adversarial signals in datasets where these tokens contribute significantly to context and meaning. For instance, in Rotten-Tomatoes Alzantot, adversarial manipulations that exploit function words (e.g., negations like "not" or intensifiers like "very") could be underdetected due to their low weighting. Despite this, our method still outperforms WDR in this configuration, suggesting that the combined effect of attention and selective POS weighting mitigates this limitation to some extent.

Table 3: Statistics of POS-tagged Words (Nouns, Verbs, Adjectives, Adverbs)

Dataset	# Total Tokens	% Selective POS-tagged (N, V, Adj, Adv)
AGNews Alzantot	17052	68.25%
AGNews TextFooler	48952	69.38%
IMDB BAE	181840	61.16%
Rotten-Tomatoes Alzantot	21774	55.26%

4.6 Further Analysis

Figure 1a illustrates the number of unique keywords used in attacks across four datasets under various experimental setups. A comparison of keyword diversity before and after replacement shows an increase in diversity for AG News following the attacks. To refine the model's keyword selection process, we examined the part-of-speech (POS) distribution of the attacking keywords, as shown in Figure 1b. The most common POS categories identified were nouns, verbs, adjectives, and adverbs, which were subsequently used in our experiments.

5 Conclusion

This project has demonstrated the effectiveness of an enhanced adversarial detection approach in NLP by incorporating part-of-speech (POS) tagging and attention to prioritize meaningful tokens and consider semantic importance of words and adjust the detector score sensitivity based on contextual importance. This refinement capture a more nuanced understanding of word-level impacts and adversarial signals. By integrating POS tagging and attention scores with the Word-level Differential Reaction (WDR) metric, we significantly increased the accuracy of identification of adversarial texts. Our experiments, conducted using the DistilBERT transformer model on the IMDb, AG News, and Rotten Tomatoes datasets, showed notable improvements in F1-scores and adversarial recall rates compared to the standard WDR method. These results highlight the potential of our approach to strengthen the reliability and responsiveness of NLP systems against adversarial attacks. Future research could explore further optimization of the detection mechanism and its applicability to a wider range of NLP tasks and adversarial scenarios.

References

Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, San Francisco California USA, August 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL https://dl.acm.org/doi/10.1145/2939672.2939785.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. Text processing like humans do: Visually attacking and shielding nlp systems. *arXiv preprint arXiv:1903.11508*, 2019.

Gil Fidel, Ron Bitton, and Asaf Shabtai. When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based Adversarial Examples for Text Classification. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.498. URL https://aclanthology.org/2020.emnlp-main.498.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL https://api.semanticscholar.org/CorpusID:6706414.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, July 1998. ISSN 1094-7167. doi: 10.1109/5254.708428. URL http://ieeexplore.ieee.org/document/708428/.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. Certified robustness to adversarial word substitutions. 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 4129–4142. Association for Computational Linguistics, 2019. URL https://arxiv.org/abs/1909.00986.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Deepika Kumawat and Vinesh Jain. Pos tagging approaches: A comparison. *International Journal of Computer Applications*, 118(6), 2015.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.16. URL https://aclanthology.org/2020.emnlp-demos.16.

Edoardo Mosca, Shreyash Agarwal, Javier Rando Ramírez, and Georg Groh. "That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7806–7816, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.538. URL https://aclanthology.org/2022.acl-long.538.

Maximilian Mozes, Pontus Stenetorp, Bennett Kleinberg, and Lewis D Griffin. Frequency-guided word substitutions for detecting textual adversarial examples. *arXiv preprint arXiv:2004.05887*, 2020.

Bo Pang and Lillian Lee. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 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, 2019a.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085–1097, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1103. URL https://aclanthology.org/P19-1103.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, February 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108 [cs].

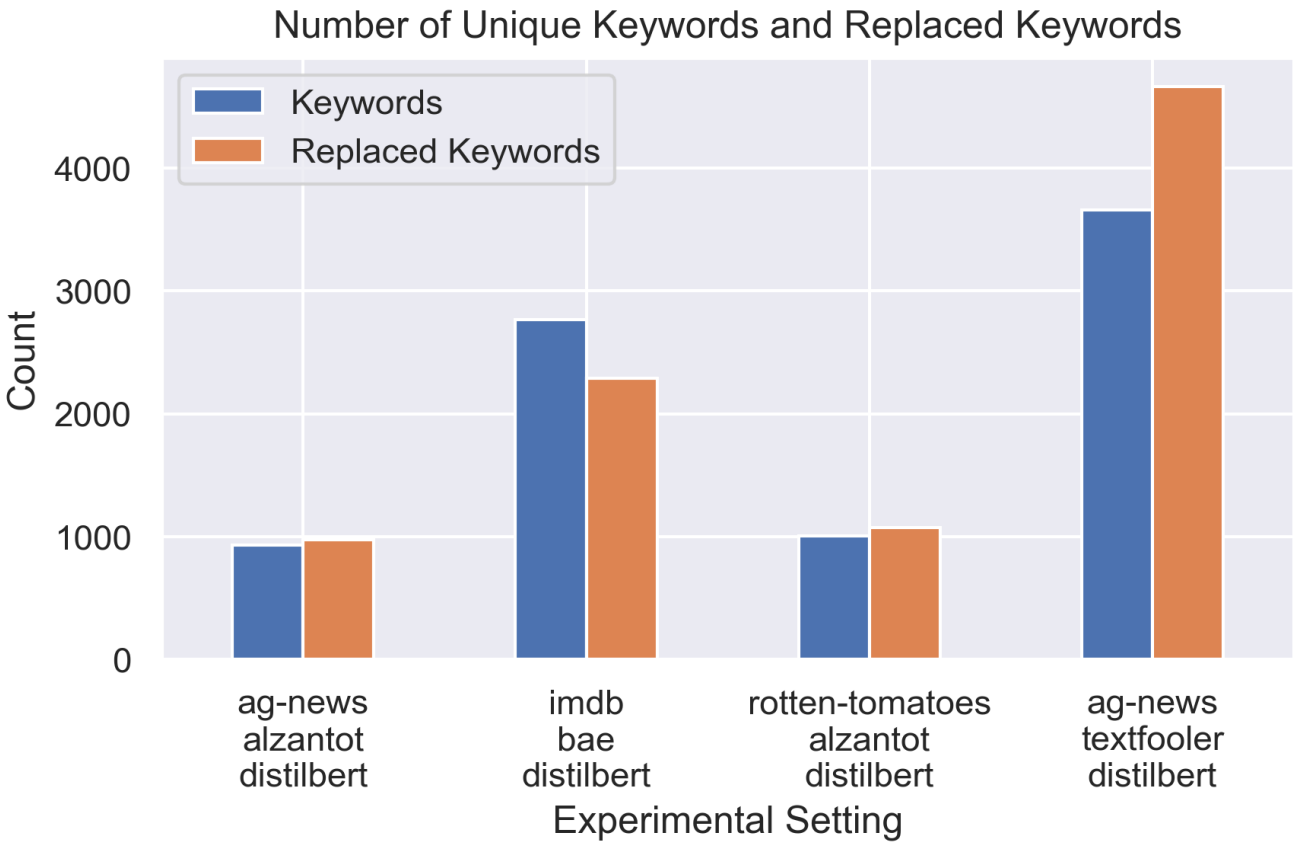
Robert E Schapire. A Brief Introduction to Boosting.

Jaswinder Singh and Rajdeep Banerjee. A Study on Single and Multi-layer Perceptron Neural Network. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 35–40, Erode, India, March 2019. IEEE. ISBN 978-1-5386-7808-4. doi: 10.1109/ICCMC.2019.8819775. URL https://ieeexplore.ieee.org/document/8819775/.

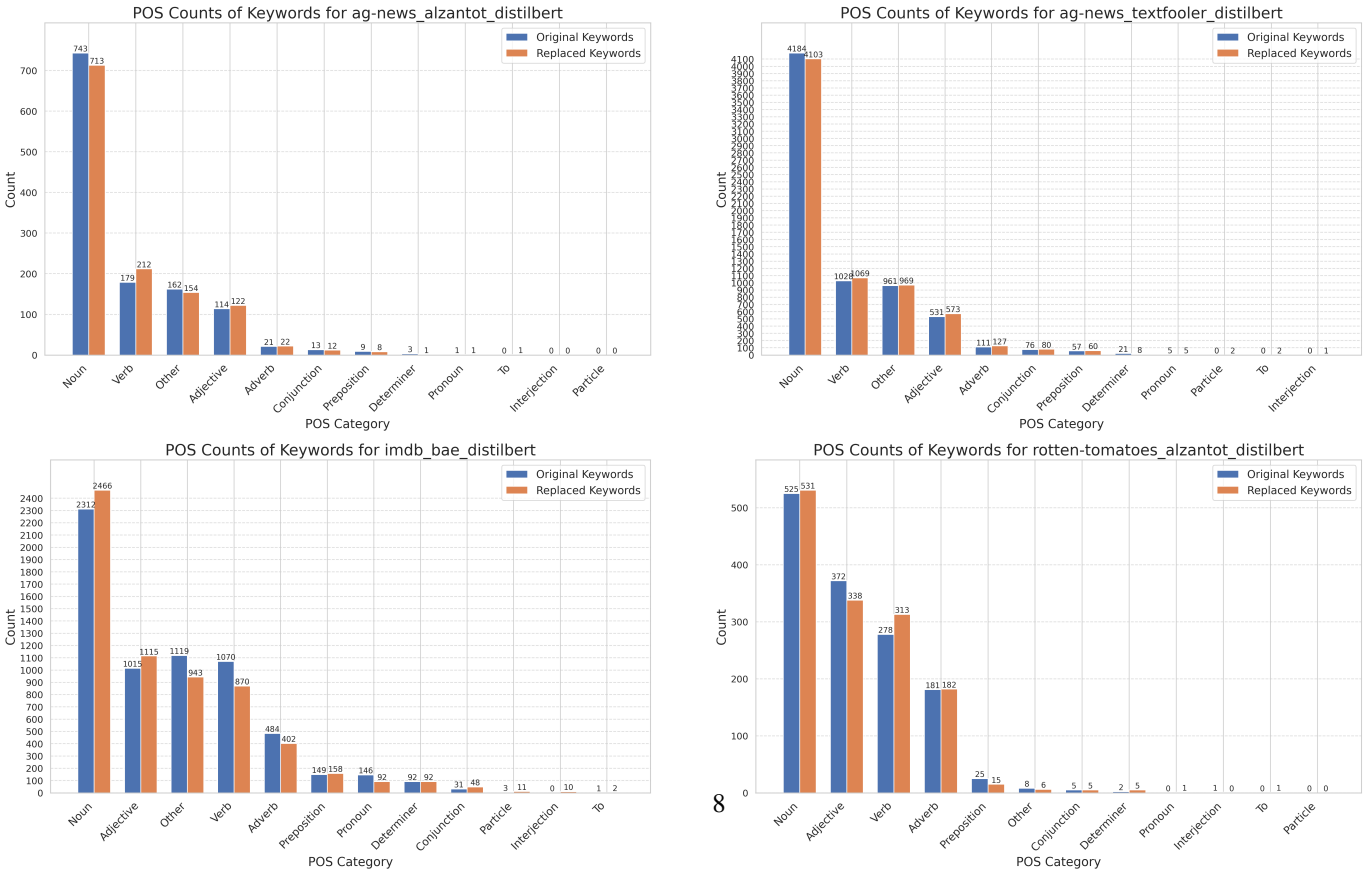
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Xiaosen Wang, Jin Hao, Yichen Yang, and Kun He. Natural language adversarial defense through synonym encoding. In *Uncertainty in Artificial Intelligence*, pages 823–833. PMLR, 2021.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.



(a) Number of unique keywords to perform attacking across 4 different experimental settings.



(b) Part-of-speech distribution for keywords