

SMAB: MAB based word Sensitivity Estimation Framework and its Applications in Adversarial Text Generation

Saurabh Kumar Pandey^{1*}, Sachin Vashista^{2*}, Debrup Das³,
Somak Aditya², Monojit Choudhury¹

saurabh2000.iitkgp@gmail.com, sachinvashista.phdcse@kgpian.iitkgp.ac.in,
saditya@cse.iitkgp.ac.in, monojit.choudhury@mbzuai.ac.ae

¹MBZUAI,

²Indian Institute of Technology, Kharagpur

³University of Massachusetts Amherst

Abstract

To understand the complexity of sequence classification tasks, Hahn et al. (2021) proposed sensitivity as the number of disjoint subsets of the input sequence that can each be individually changed to change the output. Though effective, calculating sensitivity at scale using this framework is costly because of exponential time complexity. Therefore, we introduce a Sensitivity-based Multi Armed Bandit framework (SMAB), which provides a scalable approach for calculating word-level local (sentence-level) and global (aggregated) sensitivities concerning an underlying text classifier for any dataset. We establish the effectiveness of our approach through various applications. We perform a case study on CHECKLIST generated sentiment analysis dataset where we show that our algorithm indeed captures intuitively high and low-sensitive words. Through experiments on multiple tasks and languages, we show that sensitivity can serve as a proxy for accuracy in the absence of gold data. Lastly, we show that guiding perturbation prompts using sensitivity values in adversarial example generation improves attack success rate by 15.58%, whereas using sensitivity as an additional reward in adversarial paraphrase generation gives a 12.00% improvement over SOTA approaches. **Warning: Contains potentially offensive content.**

1 Introduction

Classifiers leveraging pre-trained Language Models (PLMs) while being empirically successful, are often opaque. Identifying input subspaces where the models classify correctly (or incorrectly) or the input patterns the model is sensitive towards, is not straightforward. In the presence of model weights (white box), it may be feasible to explore such spaces but computationally intractable. On the other hand, methods attempting to explain black-box models resort to robust diagnostic tests (such as

CHECKLIST; Ribeiro et al. (2020)) at scale, which requires human input and may not be sufficient. Visualization techniques such as LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) attempt to explain local causes towards a prediction but do not provide a global view of these models. Therefore, a systematic framework is required to understand the model’s weaknesses and strengths related to the input space without assuming access to model weights.

Hahn et al. (2021) proposed a theoretical framework for understanding the complexity of sequence classification tasks using sensitivity, where sensitivity is the number of disjoint subsets of the input sequence that can each be individually changed to change the output. High-sensitivity functions are complex because a single change in the input can alter the output, whereas low-sensitivity functions are simple. Sensitivity helps predict the complexity of the tasks for various machine learning methods. While the proposed method effectively captures the task complexity, it requires iterating exhaustively over all possible independent subsequences in the input, resulting in a time complexity that is exponential to the number of input tokens. *In this work, we extend this definition of sensitivity and propose a scalable framework based on multi-armed bandits to calculate word-level sensitivities on sequence classification tasks, without assuming access to model weights and gold labels.*

Specifically, our framework provides an effective tool for local attributions of the model output to individual segments of the input (sentence level local sensitivities) as well as dataset-level sensitivities to specific words/phrases, which is independent of any context (global sensitivities). We introduce SMAB, a sensitivity-based multi-armed bandit framework, which utilizes masked language modeling (MLM) to compute word-level sensitivities in the dataset through effective exploration-exploitation strategies in a multi-armed bandit

* indicates equal contribution, Order chosen at random

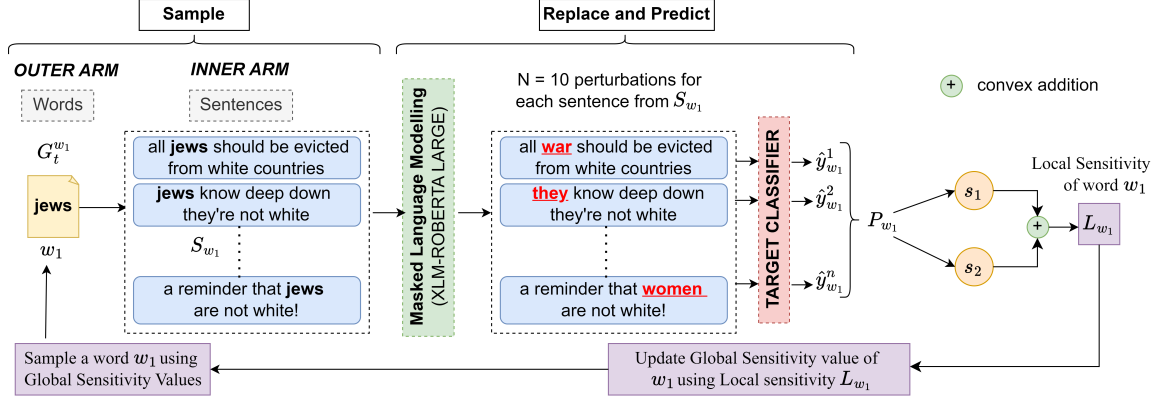


Figure 1: Overview of our **SMAB** framework. The outer arm consists of all words in the corpus, each linked to a set of sentences in the inner arm. w_1 is a word in the outer arm, and S_{w_1} is the set of sentences in the inner arm that contains w_1 . $G_t^{w_1}$ is the Global Sensitivity of word w_1 at step t . We utilize a **sample-replace-predict** strategy to estimate local sensitivity values L_{w_1} for a word w_1 . Here, P_{w_1} is the set of predicted labels obtained after perturbing S_{w_1} sentences and using the target classifier. s_1 is chosen randomly from P_{w_1} while s_2 is chosen such that it has the highest reward. The local sensitivity values of a word help to update its Global Sensitivity values, which helps in better outer arm selection in the next time step.

setup. Using these word-level sensitivities, we prove the effectiveness of the SMAB framework on three different tasks for three different application scenarios: (1) **Case Study on CHECKLIST**, where we show that for template generated datasets, the global sensitivity values obtained from SMAB can help us identify high and low sensitive words across different test types. (2) **Sensitivity as a Proxy for Accuracy**, we show that sensitivity can serve as an unsupervised proxy for accuracy for a given classifier when the gold labels are unavailable. (3) **Adversarial Example Generation**, where we propose perturbation instructions that use global sensitivities obtained from the SMAB framework along with perturbation instructions proposed in Xu et al. (2023) to attack LLMs like GPT-3.5 (Brown et al., 2020) and Llama-2-7B (Touvron et al., 2023). We also demonstrate that using local sensitivities as an additional reward helps design highly accurate paraphrase attacks for LMs using the adversarial attack generation method proposed in Roth et al. (2024). The two primary contributions¹ of this work are summarized as follows.

- We propose an efficient, scalable algorithm, SMAB, to estimate the *local* (sentence-level) and *global* (dataset-level) sensitivities of words concerning an underlying classifier without *access to model weights and gold labels*.
- We propose and empirically establish the use-

fulness of SMAB by (i) a case study on a templated dataset (using CHECKLIST) to identify high and low-sensitive words, (ii) proposing sensitivity as an unsupervised proxy for accuracy (drops), and (iii) adversarial example generation using *local* (sentence-level) and *global* (dataset-level) word sensitivities.

2 Methodology

We formalize the definitions of the global sensitivity of a word for a given text classifier. Subsequently, we explain our proposed sensitivity estimation framework with examples.

2.1 Problem Formulation

Given an input space X containing the input sentences and output space of possible labels Y , we have a pre-trained classifier $f_\theta : X \rightarrow Y$ that maps the input text $x = [w_1 \cdot w_2 \cdot w_3 \cdots w_n] \in X$ to a class $\hat{y} \in Y$. We are interested in finding the minimal subset of words to replace in x (by contextually relevant words), such that for the new sentence x' , $f_\theta(x') \neq f_\theta(x)$.

2.2 Definitions

Local Sensitivity: We define the local sensitivity of a word for a specific input text (x). For an underlying classifier (f_θ), local sensitivity estimates the relative importance of a word towards the predicted label ($f_\theta(x)$). We estimate

¹Code: <https://github.com/skp1999/SMAB>

singleton sensitivity (Hahn et al., 2021)², which is proportional to the number of flips of the predicted label when we replace a target word, say w_i , with contextually relevant words.

Global Sensitivity: Consider a text containing m words $W_x = \{w_1, w_2, \dots, w_m\}$. For an underlying classifier (f_θ), we assume there exists a minimal subset of words $W_k \subseteq W_x$ that can be replaced to change the predicted label. The global sensitivity of a word provides a greedy heuristic to discover such a minimal subset. The higher the global sensitivity, the higher the chance that the word belongs to the minimal subset. We estimate the global sensitivity of a word by aggregating the local sensitivity of words per sentence.

2.3 SMAB Framework

Multi-armed bandits offer a simple yet powerful framework for algorithms to optimize decision-making over a given period of events. Our proposed framework SMAB can interpret the importance of all words (*global sensitivity*) present in a given dataset for a particular task from a language modeling perspective. The framework is described below.

Multi-armed Bandit. Our use of multi-armed bandits has two levels – with words at the outer arms and sentences in the inner arm. The **outer arm** consists of all the words present in a dataset obtained after applying preprocessing techniques like removal of stopwords, removal of random URLs, and lemmatization. The outer arm is associated with a reward value termed as *Global Sensitivity* of a word w , which provides a greedy heuristic to discover a minimal subset of words in a sentence that needs to be changed to flip the predicted label. The **inner arm** comprises all the sentences from the dataset in which a particular word w from the outer arm is present. We denote S^w as the set of all the sentences in which a particular outer arm or word (say w) is present. Each inner arm also has a reward value, defined as *Local Sensitivity* for the inner arm.

Calculating Local Sensitivity: Local Sensitivity L_w for a word w at each step t is calculated by perturbing sentences using *sample-replace-*

²For subset sensitivity, Hahn et al. (2021) captures the variance among predicted labels, we capture mean over flips of the predicted label from the original label.

predict strategy. First, we sample a word from all the words present in the outer arm using either **Upper Confidence Bound1 (UCB)** (Auer et al., 2002) or **Thompson Sampling (TS)** (Thompson, 1933). We use **TS** to draw a sample with maximum value from the Beta distribution of sensitivity of all words using:

$$w_{t+1}^* = \operatorname{argmax}_{w \in W} (\text{Beta}(\alpha, \beta)), \quad (1)$$

where $\alpha \in (0,1)$ and $\beta = 1 - \alpha$. Then, we replace the word w in all the sentences of the inner arm using predictions of a masked language model, ensuring that the resulting sentence with the new word w' remains coherent and semantically sound. This process is repeated N times (here $N = 10$). If, across the N replacements, the new word w' matches the original word w , we discard that instance. Let P_w denote the set of all valid instances i.e., instances that have not been discarded. Lastly, we use the target model to predict the labels of the newly constructed P_w sentences. By utilizing the target model predictions of the P_w sentences, we select a randomly sampled sentence $s_1 \in P_w$ with reward r_1 (sentence-level local sensitivity) and a sentence $s_2 \in P_w$ with the highest reward r_2 is selected. The local sensitivity for a word L_w is calculated as the convex combination of rewards from s_1 and s_2 , $\epsilon \in (0, 1)$

$$L_w = \epsilon * r_1 + (1 - \epsilon) * r_2, \quad (2)$$

Calculating Global Sensitivity: Global sensitivity value (represented as G_t^w) for a particular word w at a particular step/iteration t is calculated as follows:

$$G_t^w = \frac{(N^w * G_{t-1}^w + L_w)}{1 + N^w}, \quad (3)$$

where N^w represents the number of times the word w has been picked up till now. $G_t^w \in (0,1)$. We assign L_w to 1 if $L_w > 0$. We minimize the total regret R_t over the total number of iterations. The estimation of total regret and the pseudocode of the complete algorithm is presented in Algorithm 1.

2.4 SMAB Training Details

We initialize the global sensitivity values of all the words (arms) present in a dataset with $\text{Beta}(\alpha, \beta)$, where $\alpha = \text{Random}(0, 0.5)$ and $\beta = (1 - \alpha)$. We use ϵ in the convex addition function as 0.9. We iterate with the total number of steps, $N = 2,00,000$. We report the *#datapoints* used for training, inner and outer arm details in Table 9.

3 SMAB: A Case Study on CHECKLIST

Dataset. Evaluating the global sensitivity values of words is challenging as they depend on a specific task and classifier combination. There is no straightforward way to determine the ground truth global sensitivity of different words. Therefore, we start with a template-generated dataset such as CHECKLIST (Ribeiro et al., 2020), where for a template, we know that changing specific keywords may cause a label flip while changing others should not have any effect on the label. We note that this still provides only an approximation of sensitivity values, as the global sensitivity value of a word (as defined in §2.2) also depends on the target classifier.

Method. The CHECKLIST framework creates targeted test cases inspired by standard software engineering practices. Each test belongs to one of the categories – MFT (*Minimum Functionality Test*), INV (*Invariance Test*), and DIR (*Directional Expectation Test*). INV applies perturbations that preserve the original label, whereas DIR tests if the confidence of a label changes in a specific direction. These tests (INV & DIR) consist of templates that vary according to the test type in consideration. For example, *change names* test suite of INV test type has sentences where we only vary a single word (name) in the complete sentence. In the texts shown below, only the name (Alicia) changes in the newly created example.

Example:

@JetBlue Thank you **Alicia**!Exceptional Service
 @JetBlue Thank you **Haley**!Exceptional Service

We utilize test types from INV and DIR to identify words with low and high sensitivity using our SMAB framework. We experiment with two outer arm sampling strategies, UCB and TS. We use the perturbed Twitter US airline sentiment dataset³ obtained from CHECKLIST. We sample $\sim 35k$ sentences from all the test types covering **38 test types**. We extract **8498** arms from the above sentences after some initial preprocessing (stopwords removal, lemmatization). Finally, we run our SMAB algorithm (§2.3) on the above set of arms and sentences to obtain the global sensitivity of all the words.

Observations. We observe that the perturbation of words (arms) present in INV templates does not

³<https://github.com/marcotcr/checklist>

TYPE	EXAMPLE	G_s^w
INV	@united happens every time in and out of <Newark>	0.0397
INV	@JetBlue and of course that was supposed to say <Jeremy> , not login.	0.0883
DIR	Thanks @JetBlue. Next up we will see how the slog from JFK to the city goes. You are <exceptional> .	0.6185
DIR	@USAirways Delays due to faulty engine light. Great work guys. Coming up on 2 hrs sitting on the plane. WorstAirlineInAmerica. You are <creepy> .	0.7574
INV	@SouthwestAir I didit's just been such a disheartening experience for me and my family ...and a lot of taxi money wasted. <@MZ0q19>	0.9311

Table 1: A few examples from the CHECKLIST test suite showing the highlighted words in the template and their respective estimated global sensitivity (G_s^w) using SMAB. The templated words are enclosed in <word>. The low and high sensitivity words are highlighted in blue and red respectively.

tend to change the label of the original sentence. In contrast, the words in DIR templates are more prone to flipping the label since they contribute to confidence score manipulation. Further analysis (details in Figure 5 in the Appendix) shows that the words from DIR templates have higher estimated global sensitivity and have a much wider spread of values ranging from 0 to 1. In contrast, as expected, the words from INV are concentrated in the low-sensitivity range of (0-0.2). Further, in Table 1, we present some qualitative examples from various test types, words, and their estimated global sensitivities.

Evaluation. Sensitivity threshold is the value above which a word present in a sentence if perturbed, is highly likely to change the predicted label. To evaluate the performance of our framework, we propose a metric, **Sensitivity Attack Success Rate (SASR)**, calculated as follows. Given a test dataset, a word, w , from the set of all the words present in the dataset, S_w , the set of sentences in which the word is present, and G_s^w , its estimated global sensitivity from our SMAB framework, if the word is above the sensitivity threshold and replacing the word with the predictions of a masked language model flips the predicted label in any one of the sentences from S_w , it is called a success. SASR is the fraction of all the words in a dataset above the sensitivity threshold that can flip the predicted label.

We calculate SASR for a test set from CHECKLIST templated dataset of 1800 sampled data points and plot SASR for different sensitivity thresholds

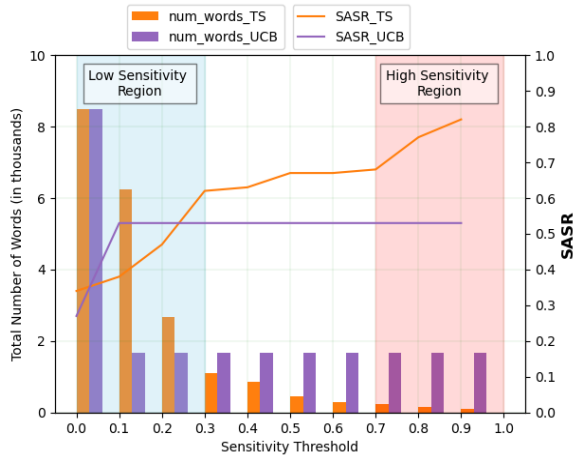


Figure 2: Variation of SASR with sensitivity threshold on CheckList test dataset for UCB and TS. For UCB, words are only present in bins (0-0.1) and (0.9-1.0), hence SASR becomes constant after 0.1. It shows that Thompson Sampling proves to be a better sampling strategy for this task as compared to UCB.

for both the algorithms, UCB and TS, as shown in Figure 2. We observe that the SASR_TS increases as we increase the sensitivity threshold, which signifies that the words in the high sensitivity region (0.7-1.0) are responsible for the change in the predicted label. SASR_UCB remains constant after threshold of 0.1 since the estimated sensitivity values from UCB lie only in two bins, (0.0-0.1) and (0.9-1.0). Additionally, we plot the number of words obtained at different sensitivity thresholds to ensure sufficient words are present to calculate SASR. For TS, we get 104 words even above the sensitivity threshold of 0.9, with around 83 of these words able to produce a flip. Hence, it shows that the estimated global sensitivities from our SMAB framework, in a true sense, capture the impact of various words in a sentence for a particular task in a given dataset.

4 Sensitivity as a Proxy for Accuracy

Using our multi-armed bandit framework (in §2.3), we calculate the word sensitivities for each word in a given dataset based on the model predictions (in an unsupervised way). Here, we aim to quantify the correlation between the accuracy of various models and the difference between their corresponding sensitivity distributions obtained from the SMAB framework. We experiment with two different settings - (i) Correlation across languages (same model) and (ii) Correlation within language (different models). We compare the KL divergence

of the sensitivity distributions from two different models (/languages) with the relative drop in accuracy between models (languages). We compute $D_{KL}(P \parallel Q) = \sum_{i=1}^N P(i) \log \frac{P(i)}{Q(i)}$, where P and Q represent sensitivity distributions obtained from two different runs of SMAB. N represents the number of sensitivity bins (here 10). We calculate accuracy on the same dataset using ground truth labels. We hypothesize that KL divergence is negatively correlated with accuracy drop for both the settings, which signifies that the sensitivity distributions may serve as an *unsupervised proxy for accuracy* for a given target classifier when gold labels are absent.

4.1 Tasks & Datasets

Hate Speech Classification Task. Hate Classification is a challenging task that contains many words spanning different sensitivity bins (contains highly-sensitive target words). We selected hate speech classification datasets from various sources, covering nine languages - English, Bengali, French, German, Greek, Hindi, Italian, Spanish, and Turkish. Details for these datasets is provided in the appendix B.1.1. Hereafter, we refer to this dataset as the mHate dataset.

Natural Language Inference Task. We use the XNLI (Conneau et al., 2018) dataset, a cross-lingual NLI dataset for this task, which expands upon the English-based MultiNLI dataset (Williams et al., 2018) by translation into 14 languages. We select five languages - English, French, Greek, Hindi, and Spanish to evaluate our hypothesis.

4.2 Correlation Across Languages

Robust evaluation and benchmarking of low-resource languages have always been challenging because of the lack of sufficient and reliable evaluation datasets (Ahuja et al., 2022a), (Ahuja et al., 2022b). SMAB may be highly effective in the evaluation and benchmarking of low-resource languages. We experiment with various languages of mHate and XNLI datasets. We quantify the relative zero-shot drop in accuracy and attempt to correlate it with the KLD. We utilize the mBERT (Devlin et al., 2019) classifier for mHate and mDeBERTa (He et al., 2021) for XNLI. Given a classifier, we get the predictions on the mentioned language split of the dataset. Then, we compare the KLD between sensitivity distributions of different languages from a base language (English, in our case) and the model’s accuracy in various lan-

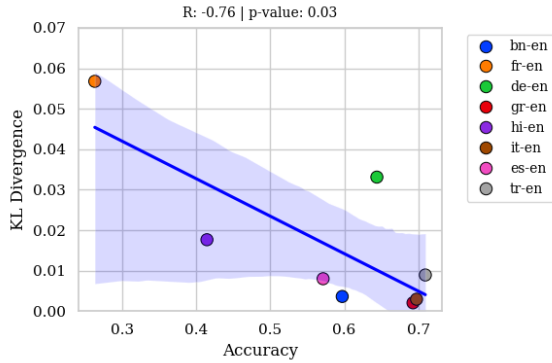


Figure 3: KL Divergence v/s accuracy across languages of mHate dataset using mBERT.

guages. We plot KLD v/s accuracy to study nine languages for a particular target classifier under study.

4.3 Results

From Figures 3 and 4, we observe a negative correlation between KL Divergence and Accuracy on the test set. For mHate, we calculate KLD between sensitivity distributions obtained from our SMAB framework across eight languages and sensitivity distribution for English. We observe a negative correlation with Pearson Correlation Coefficient (R) of -0.75 (statistically significant with p -value of 0.03). Similarly, for XNLI, we perform it for 5 different languages against English and obtain a Correlation Coefficient of -0.91 (statistically significant with p -value of 0.03). We also carry out experiments for **within the language** setting using various pre-trained classifiers and obtain a similar correlation (Appendix C). The results show that KLD between sensitivity distributions is *negatively correlated* with the accuracy of a target classifier. Hence, the sensitivity values obtained from SMAB act as an *unsupervised proxy for accuracy* for a given target classifier on a dataset when the gold labels are absent.

5 Adversarial example generation

5.1 Datasets

We evaluate the utility of word-level sensitivities in *adversarial example generation* using the Sentiment Analysis task. We conduct experiments on the Cornell Movie Review dataset (RT dataset) (Pang and Lee, 2005) and SST-2 dataset (Socher et al., 2013). We extend two perturbation-based (or paraphrase) attack baselines using sensitivities

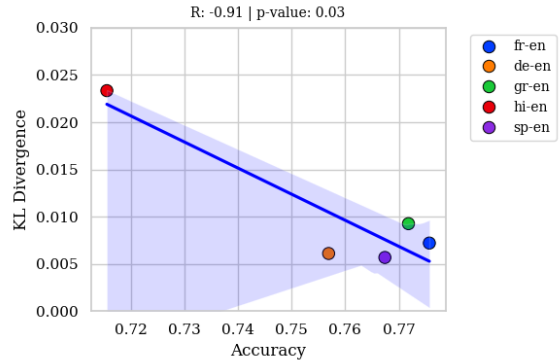


Figure 4: KL Divergence v/s accuracy across languages of XNLI dataset using mDeBERTa.

predicted from our SMAB framework.

5.2 Evaluation Metrics

Attack Success Rate: Following Wang et al. (2018), we evaluate our attack methods using the Attack Success Rate (ASR). Given a dataset $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^N$ consisting of N pairs of samples $x^{(i)}$ and ground truth labels $y^{(i)}$, for an adversarial attack method A that generates an adversarial example $A(x)$ given an input x to attack a surrogate model f , ASR is calculated as follows:

$$\sum_{(x,y) \in \mathcal{D}} \frac{1[(f(A(x)) \neq y) \wedge (f(x) = y)]}{1[f(x) = y]} \quad (4)$$

After Attack Accuracy: It measures the robustness of a model to an adversarial attack, A low After-attack accuracy represents a highly vulnerable model to adversarial attacks. It is calculated as follows:

$$\frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} 1[(f(A(x)) = f(x) = y)] \quad (5)$$

5.3 PromptAttack using Global Sensitivities

Our first baseline, PromptAttack (Xu et al., 2023), is a prompt-based adversarial attack that can effectively audit the adversarial robustness of a target LLM. For a target LLM, PromptAttack prompts the same LLM with a perturbation instruction such that the LLM generates an adversarial sample for a given input text, effectively fooling itself. Authors propose character-level, word-level, and sentence-level perturbations. As we capture word-level sensitivities, we define three word-level perturbation instructions utilizing the

Perturb Type	Perturbation instructions
$W1$	Replace at most two words in the sentence with synonyms.
$W2$	Choose at most two words in the sentence that do not contribute to the meaning of the sentence and delete them.
$W3$	Add at most two semantically neutral words to the sentence.
$W4$ (Ours)	For a given sentence, there always exists a minimal subset of words that need to be replaced to flip the label of the sentence while preserving its semantic meaning. Global Sensitivity of a word provides a greedy heuristic to discover such a minimal subset. The higher the global sensitivity, the higher the chance that the word belongs to the minimal subset. Given the minimal subset of the words "[Word1]", "[Word2]" and their global sensitivity values in the decreasing order [GS1, GS2], replace these words in the original sentence with semantically close words.
$W5$ (Ours)	The words "[Word1]" and "[Word2]" are highly sensitive in the given sentence, and perturbing either "[Word1]", "[Word2]", or both can change the label of the sentence while preserving the semantic meaning of the new sentence as that of the original.
$W6$ (Ours)	Add at most two semantically close words to the sentence, replacing the words "[Word1]" or "[Word2]", or both.

Table 2: Perturbation instructions used in Prompt Guidance: In $W4$, $W5$ and $W6$, [Word1] is replaced with the most sensitive word in the sentence, while [Word2] is replaced with the second most sensitive word. In $W4$, GS1 and GS2 represent the global sensitivity values of the most sensitive word and the second most sensitive word.

global sensitivity values obtained from the SMAB framework on the SST-2 dataset. Prompts for the sentiment classification task on SST-2 can be found in Table 11 and all perturbation instructions are outlined in Table 2. We experiment with GPT-3.5 (gpt-35-turbo), Llama-2-7B and Qwen-2.5-7B as the target LLMs. Our instructions primarily rely on perturbing the high-sensitive words in the input text. We apply the *fidelity filter* used in Xu et al. (2023) with the following constraints: BERTScore ≥ 0.92 (Zhang et al., 2019) and Word Modification Ratio ≤ 1.0 (Wang et al., 2018).

Results. As shown in Table 3, using GPT-3.5 as the target LLM, we compute sensitivity values from our SMAB framework with six different LLMs serving as classifiers, referred to as **SMAB LLMs**: BERT, GPT-3.5, Llama-2-7B, Llama-2-13B, Llama-3.1-8B and Qwen-2.5-7B. For all four SMAB LLMs, our word level perturbation instructions outperform the baselines $W1$, $W2$, and $W3$. We obtain the best ASR of **52.06%** and the lowest After Attack Accuracy of **44.23%** when we use Qwen-2.5-7B as the SMAB LLM. It demonstrates improvement over the baselines by a margin of **15.58%**. Likewise, with Llama-2-7B as both the target LLM and the SMAB LLM, perturbation type $W5$ outperforms the top baseline $W3$. With Qwen-2.5-7B as both the target LLM and SMAB LLM, our perturbation type $W6$ outperforms the top baseline $W3$. The results show that our instructions utilizing high-sensitive words consistently outperform the baselines and can generate high-quality

adversarial examples as shown in Table 12.

5.4 ParaphraseAttack using Local Sensitivities

Roth et al. (2024) studies adversarial attacks on text classifiers using an encoder-decoder paraphrase model (T5), trained to generate adversarial examples using a reinforcement learning algorithm (Williams, 2004) with a *constraint-enforcing reward* that helps generation of semantically close, label invariant and grammatically correct adversarial examples. We modify the original reward function by incorporating an additional **Sensitivity reward**, defined as the difference between the sensitivity of the input text and the sensitivity of the generated text. Details of the modified reward function and the keyphrase sensitivity calculation can be found in Appendix D. During inference, we generate 8 paraphrases for each input text using various decoding mechanisms under constraints used in Roth et al. (2024) and use the target classifier to predict the label. We used the RT dataset to conduct experiments using the same hyperparameter and design choices used in the referenced work. The details of the different models used, including the paraphrase model are provided in Table 6.

Human Evaluation. We also perform a human evaluation of the generated adversarial examples. Two co-authors (graduate CS students trained in NLP) quantify the quality of generated paraphrases on three aspects. We use the human evaluation metrics proposed in Das et al. (2024). Specificity (SPE) measures how specific the aspects obtained in the

SMAB LLM	Perturb Type	ASR \uparrow	After Attack Accuracy \downarrow
Target LLM \rightarrow GPT-3.5 (Before Attack Accuracy - 92.16%)			
\times	W1	22.20	72.09
	W2	24.63	69.73
	W3	36.48	58.68
BERT	W4	37.32	58.20
	W5	38.34	56.96
	W6	48.23	47.84
Llama-2-7B	W4	37.49	57.85
	W5	38.20	57.04
	W6	48.26	47.80
Llama-2-13B	W4	38.61	56.95
	W5	38.56	56.79
	W6	50.09	46.23
Llama-3.1-8B	W4	48.16	47.80
	W5	39.96	55.40
	W6	46.98	48.90
Qwen-2.5-7B	W4	51.62	44.61
	W5	38.87	56.37
	W6	52.06	44.23
GPT-3.5	W4	37.94	57.55
	W5	39.40	56.09
	W6	46.06	49.95
Target LLM \rightarrow Llama-2-7B (Before Attack Accuracy - 92.32%)			
\times	W1	56.19	36.13
	W2	43.97	48.35
	W3	60.20	32.11
Llama-2-7B	W4	30.55	61.76
	W5	61.69	30.62
	W6	44.85	47.46
Target LLM \rightarrow Qwen-2.5-7B (Before Attack Accuracy - 76.38%)			
\times	W1	36.70	39.68
	W2	14.79	61.58
	W3	28.78	47.71
Qwen-2.5-7B	W4	28.44	47.94
	W5	36.01	40.25
	W6	37.16	39.22

Table 3: PromptAttack results on SST-2 dataset. **SMAB LLM** is the LLM used for calculating sensitivity values from the SMAB framework. ASR and After Attack Accuracy are in (%). All **perturb types** that outperformed the highest baseline score is highlighted in green and the best-performing perturb type is highlighted in green.

generated text are in response to the input. Grammaticality (**GRM**) measures how grammatically correct are the paraphrased sentences. Choose-or-Not (**CHO**) represents whether a human considers

the generated paraphrase adversarial. Humans evaluate SPE and GRM on a Likert scale of 1-5 (least to most). CHO is 1 if a human considers the generated paraphrase adversarial, else 0. We compute the average score across all instances for each metric. For human evaluation, we selected up to 100 unique paraphrases generated by each model variation that flipped the label of a text. The number of unique paraphrases varies across strategies, as detailed in Table 7.

Results. Table 4 shows that incorporating *sensitivity reward* consistently results in a higher **ASR** and **CHO** for a particular strategy and temperature combination. Each strategy uses a combination of decoding mechanisms and whether the sensitivity reward is used or not. We achieved the highest ASR of **93.03%** with DistilBERT (Sanh, 2019) as the SMAB Model, *beam search* decoding strategy and a temperature of 1.00 that exceeds the corresponding baseline variant by **74.09%** and the best baseline variant by **62.39%**. Similarly, we achieved the best CHO using BERT as the SMAB model that exceeds the corresponding baseline variant by **20.14%** and the best baseline variant by **12.00%**. We present the qualitative examples in the Table 10.

Strategy	Temp	ASR \uparrow	SPE \uparrow	GRM \uparrow	CHO \uparrow
(Human Evaluation)					
SMAB Model \rightarrow \times					
DBS + (\times)	0.85	30.64	3.43	3.65	27.00
BS + (\times)	0.85	17.27	3.50	3.85	22.58
BS + (\times)	1.00	18.94	3.72	3.87	19.11
BS + (\times)	1.15	14.76	3.64	3.86	18.86
SMAB Model \rightarrow BERT					
DBS + (\checkmark)	0.85	50.41	3.28	3.35	21.00
BS + (\checkmark)	0.85	19.77	3.38	3.88	25.35
BS + (\checkmark)	1.00	22.00	3.46	3.65	24.05
BS + (\checkmark)	1.15	50.97	3.52	3.45	39.00
SMAB Model \rightarrow DistilBERT					
DBS + (\checkmark)	0.85	88.85	3.88	3.90	22.00
BS + (\checkmark)	0.85	38.99	3.44	3.57	37.00
BS + (\checkmark)	1.00	93.03	3.86	3.88	26.00
BS + (\checkmark)	1.15	20.89	3.40	3.74	22.66

Table 4: ParaphraseAttack Results on RT dataset. The strategy represents the combination of decoding mechanism (**DBS**: Diverse Beam Search, **BS**: Beam Search) and usage of sensitivity reward (\times : No sensitivity, \checkmark : Sensitivity using BERT/DistilBERT). Temp: decoding temperature. ASR/CHO in (%). All variants outperforming the highest baseline score are highlighted in green and the best-performing variant is highlighted in green.

6 Related Work

Sensitivity has been commonly used as a measure for the *complexity* of sequence classification tasks (Hahn et al., 2021). Sensitivity has also been used for the understanding and optimization of prompts in related paradigms such as in-context learning. Lu et al. (2024) performs an analysis of accuracy and sensitivity, for different prompts in an *ICL* setting and observes a negative correlation between them. *FormatSpread* (Sclar et al., 2024) also presents a Multi-armed bandit framework for a model-agnostic evaluation of performance spread across different prompt formats. Multiple research work have focused on the paradigm of adversarial text generation by perturbing safe input examples through gradient-based approaches (Ebrahimi et al., 2018; Cheng et al., 2020). Wallace et al. (2019) deployed a gradient guided search over all tokens to extract *universal adversarial triggers*, which are input-agnostic tokens to trigger a model. Ribeiro et al. (2018) followed a similar approach by presenting simple and universal *semantically equivalent adversarial rules* (SEARs) that create adversaries on safe inputs. Other approaches such as Iyyer et al. (2018) and Roth et al. (2024), have delved into the training of paraphrase networks for controlled generation of attack examples, whereas Xu et al. (2023) proposed prompt-based adversarial attack to audit LLMs.

7 Conclusion

We introduce the notion of *local* (sentence-level) and *global* (word-level) sensitivities to capture the intricacies of a text classifier for a given dataset. We introduce a novel, cost-effective sensitivity estimation framework, SMAB. Through experiments on CHECKLIST-generated dataset, we show that our SMAB framework captures high-sensitive and low-sensitive words effectively. We observe that the comparative accuracy between two models (for the same language or for across language on the same task) has strong negative correlations with KL divergence between (global) sensitivity distributions of the models – showing sensitivity can be used as an unsupervised proxy for accuracy (drops). Further, we define three word-level perturbation instructions utilizing the global sensitivity values obtained from the SMAB framework to attack LLMs such as GPT-3.5 with a high success rate. We also show that sensitivity can be used as an additional reward in paraphrase-based attacks to improve the

attack success rate of adversarial models. Hence, word-level sensitivities provide a closer look at how opaque language models work.

Limitations

The work explores the proposed framework for sequence classification tasks. Further exploration is needed to extend to other tasks, such as generation and translation. The hypothesis of sensitivity acting as an unsupervised proxy is valid under the specific conditions we tested. A more detailed study of various families of models and tasks might provide deeper insights into the correlation, which will be highly useful for evaluating and benchmarking low-resource languages. It is also important to note that we performed experiments concerning adversarial example generation in English, and a full-fledged multilingual study needs to be performed.

Ethics Statement

Although our framework helps identify words with different sensitivity levels, there can be a few repercussions. It is important to note that the method does not guarantee that the examples generated will always be adversarial. The framework, and hence the sensitivity values, may be misused by people to develop better jailbreak techniques.

Acknowledgements

This research is partially supported by SERB SRG/2022/000648. We acknowledge the OpenAI and Azure credits from the Microsoft Accelerate Foundation Models Research (AFMR) Grant. Sachin Vashistha is partially supported by the Prime Minister’s Research Fellowship (PMRF) grant.

References

- Kabir Ahuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2022a. Beyond static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages. *arXiv preprint arXiv:2205.06356*.
- Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022b. Multi task learning for zero shot performance prediction of multilingual models. *arXiv preprint arXiv:2205.06130*.
- Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256.

- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. [Hostility detection dataset in hindi](#). *Preprint*, arXiv:2011.03588.
- Florian Boudin. 2016. [pke: an open source python-based keyphrase extraction toolkit](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 69–73, Osaka, Japan.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. [Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3601–3608.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- Çağrı Çöltekin. 2020. [A corpus of Turkish offensive language on social media](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). *Preprint*, arXiv:1911.02116.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Mithun Das, Saurabh Pandey, Shivansh Sethi, Punyajoy Saha, and Animesh Mukherjee. 2024. [Low-resource counterspeech generation for Indic languages: The case of Bengali and Hindi](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1601–1614, St. Julian’s, Malta. 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](#). *Preprint*, arXiv:1810.04805.
- 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.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. [Latent hatred: A benchmark for understanding implicit hate speech](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Michael Hahn, Dan Jurafsky, and Richard Futrell. 2021. [Sensitivity as a complexity measure for sequence classification tasks](#). *Transactions of the Association for Computational Linguistics*, 9:891–908.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [Deberta: Decoding-enhanced bert with disentangled attention](#). *Preprint*, arXiv:2006.03654.
- 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.
- Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. [How are prompts different in terms of sensitivity?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856, Mexico

- City, Mexico. Association for Computational Linguistics.
- Scott Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). *Preprint*, arXiv:1705.07874.
- Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. [Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages](#). In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '19*, page 14–17, New York, NY, USA. Association for Computing Machinery.
- Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. [Multilingual and multi-aspect hate speech analysis](#). *Preprint*, arXiv:1908.11049.
- Bo Pang and Lillian Lee. 2005. [Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales](#). In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.
- Juan Carlos Pereira-Kohatsu, Lara Quijano Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. [Detecting and monitoring hate speech in twitter](#). *Sensors (Basel, Switzerland)*, 19.
- Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. [Offensive language identification in Greek](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ["why should i trust you?": Explaining the predictions of any classifier](#). *Preprint*, arXiv:1602.04938.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. [Semantically equivalent adversarial rules for debugging NLP models](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.
- Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2020. [Hate speech detection in the bengali language: A dataset and its baseline evaluation](#). *Preprint*, arXiv:2012.09686.
- Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. [Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis](#). In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.
- Tom Roth, Inigo Jauregi Unanue, Alsharif Abuadba, and Massimo Piccardi. 2024. [A constraint-enforcing reward for adversarial attacks on text classifiers](#). *Preprint*, arXiv:2405.11904.
- V Sanh. 2019. [Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter](#). *arXiv preprint arXiv:1910.01108*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- William R Thompson. 1933. [On the likelihood that one unknown probability exceeds another in view of the evidence of two samples](#). *Biometrika*, 25(3-4):285–294.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. [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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ronald J. Williams. 2004. [Simple statistical gradient-following algorithms for connectionist reinforcement learning](#). *Machine Learning*, 8:229–256.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. 2023. [An llm can fool itself: A prompt-based adversarial attack](#). *Preprint*, arXiv:2310.13345.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#). *arXiv preprint arXiv:1904.09675*.

A MAB Framework: Additional Details

We summarize the details of the SMAB framework in an Algorithm format in Algorithm 1.

Outer Arm Selection. We employ the UCB sampling strategy to choose a specific word w at each step t using:

$$w_{t+1}^* = \operatorname{argmax}_{w \in W} (\operatorname{Beta}(\alpha, \beta))$$

where W is the set of all the words or outer arms. N^w is the no. of times a word w has been picked so far.

Estimation of Total Regret. In MAB, Total Regret R_t is defined as the total loss we get by not selecting the optimal action up to the step or iteration t . Let the outer arm or word w be picked up at the step or iteration t . Now, in turn, we will pick up sentences s_1 and $s_2 \in S_w$. Let L_w be the local sensitivity. Hence, the Total Regret R_t up to the iteration t is defined as:

$$R_t = R_{t-1} + ([L_{w^*} - L_w] * G_t^w) \quad (6)$$

where G_t^w is the Global sensitivity value of (the outer arm) the word w that was picked and L_{W^*} is the highest value of local sensitivity that can be obtained out of the set S^w .

Comparison of Time Complexity. Let $|D|$ be the size of the dataset i.e. the number of input sentences in the dataset, $|\Sigma|$ be the total number of words in the dataset, $|V|$ be the vocabulary size of the Language Model, and $\operatorname{cost}(f)$ be the cost to use a Language Model for various purposes like classification, Masked Language Modeling.

Time Complexity for Subset-sensitivity: Let $|P|$ be the size of the subset. For every input sentence in the dataset, we use an MLM to generate $|V|^{|P|}$ perturbed strings and then classify them using an LLM to see if the label flipped. Hence, the Time Complexity is:

$$\mathcal{O}(|D| \cdot |V|^{|P|} \cdot \operatorname{cost}(f)) \quad (7)$$

For calculating the block sensitivity, the subset sensitivity is calculated K times, corresponding to the K partitions, which can go exponential.

Time Complexity of SMAB: Local sensitivity in our algorithm is closely related to the subset sensitivity defined in [Hahn et al. \(2021\)](#). Our local sensitivity is obtained by taking $|P| = 1$ (singleton sensitivity) and global sensitivity is calculated using equation 3 in our paper, which is $\mathcal{O}(1)$. The time complexity of our SMAB algorithm given T as the total number of iterations is:

$$\mathcal{O}(T \cdot (|\Sigma| + (|D| \cdot |V| \cdot \operatorname{cost}(f)))) \quad (8)$$

It is calculated as

- **Initialization (Lines 1-3):** Initializing \mathbf{G} and \mathbf{N} , both of size $|\Sigma|$, has a complexity of $\mathcal{O}(|\Sigma|)$ and setting $t \leftarrow 0$ has $\mathcal{O}(1)$ complexity. Hence, the total complexity for the Initialization step is $\mathcal{O}(|\Sigma|)$.
- **Outer Loop (Lines 4–10):** The loop runs \mathbf{T} times. We analyze the per-iteration complexity below.
- **Selecting the Word (Line 5):** Finding w^* requires evaluating the expression
$$w_{t+1}^* = \operatorname{argmax}_{w \in W} (\operatorname{Beta}(\alpha, \beta))$$
that requires $\mathcal{O}(|\Sigma|)$ operations.
- **Fetching Sentences (Line 6):** Accessing S^w , the set of sentences for the selected word w^* , has $\mathcal{O}(1)$ complexity.

- **Selecting Two Sentences and Calculating Local Sensitivity (Line 7):** Here, first the set S^w is created by using xlm-roberta-large 10 number of times for each sentence. It takes $\mathcal{O}(|D| \cdot |V|)$ complexity as our setup replaces a single word every time. Then, for every sentence, a target classifier is used to predict the label that requires $cost(f)$ for each sentence. After that, selecting s_1 randomly from S^w has $\mathcal{O}(1)$ complexity. Selecting s_2 with the highest reward requires evaluating the reward for all sentences in S^w , which takes $\mathcal{O}(|D|)$ and finally, Computing the local sensitivity L_w is a constant-time operation: $\mathcal{O}(1)$. Total complexity for this step is $\mathcal{O}(|D| \cdot |V| \cdot cost(f))$
- **Updating Global Sensitivity (Line 8):** It is a constant-time operation: $\mathcal{O}(1)$.
- **Final Output Computation (Lines 9–10):** Calculating the final global sensitivities $\mathbf{G}[w]$ for all $w \in \mathbf{W}$ and the total regret R_t involves $\mathcal{O}(|\Sigma|)$.

The per-iteration complexity is:

$$\mathcal{O}(|\Sigma| + (|D| \cdot |V| \cdot cost(f))).$$

For \mathbf{T} iterations, the overall complexity becomes:

$$\mathcal{O}(T \cdot (|\Sigma| + (|D| \cdot |V| \cdot cost(f))))$$

This shows that our proposed algorithm is computationally more efficient than the existing algorithms. In practice, the top few replacements (perturbations) from the MLM probability distribution are used instead of iterating over all vocabulary symbols.

B CHECKLIST: Additional Results

In Figure 5, we show a scatter plot of the word sensitivities estimated using SMAB (with Thomson Sampling). As mentioned in the main paper, we observe that words from DIR templates have a larger distribution, present in the 0.2 – 1 range, whereas, the words from the invariant test cases (INV template) have lower sensitivities (mostly between 0 and 0.2).

B.1 Dataset details

B.1.1 Hate Speech Dataset

In this section, we describe the sources of our compiled dataset for the hate classification task for different languages. The complete statistics for all the languages can be found in Table 5.

Algorithm 1 Multi-Armed Bandit Algorithm

Input: A set of words/outer-arms \mathbf{W} , Dictionary \mathbf{D} containing the set \mathbf{S}^w of sentences as a *value* for every *key* i.e. word $w \in \mathbf{W}$ and total number of iterations $\mathbf{T} \leftarrow 200000$.

Output: The set \mathbf{G} containing final global-sensitivity values for every word $w \in \mathbf{W}$

- 1: Initialize the set \mathbf{G} as the initial values of the global sensitivities of the words. Here, $|\mathbf{G}| = |\mathbf{W}|$
- 2: Initialize the set \mathbf{N} ($|\mathbf{N}| = |\mathbf{W}|$) to zero. \mathbf{N} represent the count of every word $w \in \mathbf{W}$.
- 3: $t \leftarrow 0$
- 4: **Repeat steps 5 to 9 until $t \neq \mathbf{T}$:**
- 5: Select a word $w \in \mathbf{W}$ such that

$$w_{t+1}^* = \operatorname{argmax}_{w \in \mathbf{W}} (Beta(\alpha, \beta))$$

- 6: $\mathbf{S}^w \leftarrow \mathbf{D}[w^*]$, $N^w \leftarrow N^w + 1$
- 7: Select two sentences $s_1, s_2 \in \mathbf{S}^w$ and calculate Local sensitivity as:

$$L_w = \epsilon \cdot r_1 + (1 - \epsilon) \cdot r_2$$

- 8: Update Global Sensitivity G_t^w as

$$G_t^w = \frac{(N^w * G_{t-1}^w + L_w)}{1 + N^w}$$

- 9: Final Global Sensitivity Values: $\mathbf{G}[w] \leftarrow G_t^w$
 - 10: Total Regret $R_t = R_{t-1} + (L_{w^*} - L_w) * G_t^w$
-

English: We used the **Stage 1: High Level Categorization** of the *Implicit Hate* dataset provided by (ElSherief et al., 2021). It contains three labels namely implicit_hate, explicit_hate, and not_hate. We converted this multi-classification task into a binary classification task where the labels implicit_hate, and explicit_hate are treated as the hate label.

Bengali: We used a subset of the Bengali Hate speech dataset provided by (Romim et al., 2020) which includes the categories *crime*, *religion*, and *politics* for hate label and all the categories for non-hate label.

Hindi: For the Hindi language, we combined datasets collected from two different sources: 1) (Mandl et al., 2019) provides a binary (hate/no-hate) version of the Hindi dataset. 2) We used a subset of the Hindi dataset (excluding the *fake cat-*

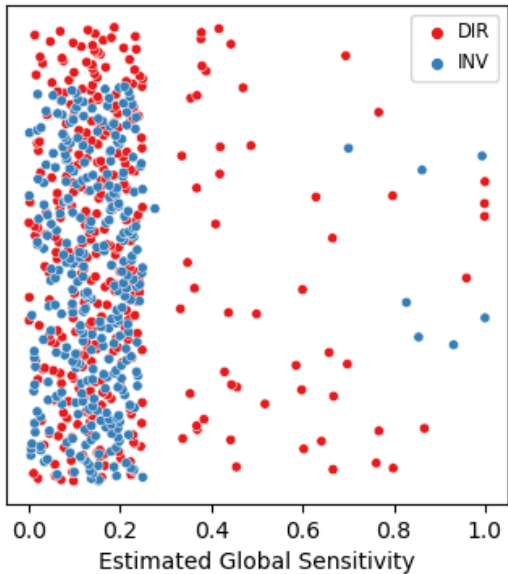


Figure 5: Scatter plot of estimated global sensitivities of arms of INV and DIR templates using TS. Words from DIR templates have higher estimated global sensitivity and are spread in the whole space as opposed to words from INV templates.

egory) provided by (Bhardwaj et al., 2020) which includes the category *non-hostile* for non-hate label and all remaining categories for hate label.

Spanish: We used two datasets for the Spanish language: 1) HatEval dataset is provided by (Basile et al., 2019), and 2) (Pereira-Kohatsu et al., 2019) has provided the hate speech dataset in the Spanish language.

Turkish: We used a corpus of offensive content for the Turkish language collected from Twitter posts (Çöltekin, 2020).

French: We used the *french* version of the Multilingual and Multi-Aspect Hate Speech Analysis dataset (Ousidhoum et al., 2019) and converted it into a binary (hate or non-hate) classification task.

German: We combined the dataset from two different sources: 1) The german hate speech dataset provided for the HASOC track at FIRE 2019 (Mandl et al., 2019), and 2) the german hate speech dataset related to the Refugee crises (Ross et al., 2016).

Greek: We used the Offensive Greek Tweet Dataset (OGTD) provided by (Pitenis et al., 2020).

Language	Hate	No Hate	Total
English	1026	1658	2684
Bengali	806	1194	2000
French	1593	421	2014
German	904	1607	2511
Greek	501	1100	1601
Hindi	783	549	1332
Spanish	1010	1290	2300
Turkish	400	1290	1690

Table 5: Dataset Statistics for val split of mHate dataset

C KLD v/s Accuracy Experiments

Correlation Within Language. We experiment with various models on mHate and XNLI dataset. We use 5 different target classifiers for a language to estimate global sensitivities. We use XLM-R (Conneau et al., 2020), mBERT, mDeBERTa (He et al., 2021), FlanT5-L (Chung et al., 2022) and GPT-3.5 for predictions on mHate and XNLI dataset. For a given language, we then compare the difference in sensitivity distributions of different models with respect to a base model (XLM-R), measured as KL Divergence with the performance of different models on that language (accuracy). We plot KLD v/s Accuracy plots for different models under study.

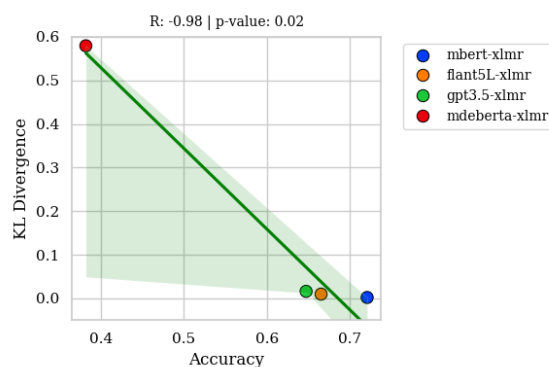


Figure 6: KLD v/s accuracy within language on mHate-English dataset.

D Improving Paraphrase Attack with Local Sensitivity

D.1 Sensitivity Reward Calculation

We adjust the original reward function $R(x, x')$ by incorporating an additional **Sensitivity reward** $S(x, x')$ weighted by the scaling constant $\alpha \in$

(0, 1) as shown:

$$R(x, x') = R(x, x') + \alpha S(x, x'). \quad (9)$$

Here, $S(x, x')$ is the difference between the sensitivity of the input text x and the sensitivity of the generated text x' : $S(x, x') = s(x) - s(x')$. The sensitivity of a text is calculated as

$$s(x) = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} L_s^{ij}}{\sum_{i=1}^m n_i} \quad (10)$$

where m represents the total number of keyphrases, n_i represents the number of words in the i -th keyphrase and L_s^{ij} represents the local sensitivity of j^{th} word in the i^{th} keyphrase. For our experiments, we used Scaling constant $\alpha = 0.25$. We extract the keyphrases from a text using a TopicRank keyphrase extraction model using an open source toolkit pke (Boudin, 2016).

D.2 Key Phrase sensitivity Calculation

We present the key-phrase sensitivity estimation process in a pseudo-code form in Algorithm 2. This algorithm takes as input the original text x and calculates its sensitivity. The algorithm has three main steps: In the **first step**, it extracts the key phrases from the input text x using the Topic Rank Key phrase extraction Model M and stores them in a list K . This first step is shown in line 6 of the pseudo-algorithm. In the **second step**, it calculates the sensitivities of each key phrase. The algorithm iterates through all the key phrases stored in the list K and does the following: ① It extracts all the words that are present in the current key phrase and stores it in the list $words$. Similarly, it maintains a global variable $Total_words$ that will store the total number of words extracted from each Keyphrase. ② It creates a masked sentence x_{masked} from the input text x by replacing all the words that are present in the current key phrase with [MASK]. ③ It then uses bert-large-uncased to generate 10 perturbed sentences by predicting new words at the locations where [MASK] is present in x_{masked} . These 10 perturbations are stored in the list $Masked_output$. ④ The algorithm then predicts the label of the input text x and 10 perturbed sentence present in the list $Masked_output$ and stores the predictions in the variable $Prediction_original$ and in the list $Prediction_List$ respectively. ⑤ Then, it iterates through all the labels in the $Prediction_List$ and increments the variable $flips$ if the $Prediction_original$ doesn't match the current label. ⑥ Finally, it calculates the local sensitivity of the current keyphrase as the pro-

portion of the total flips that we got out of the total labels i.e. $flips/\text{len}(Prediction_List)$ and stores it in a list $Keyphrase_sensitivity$ at the index corresponding to the current key phrase. This second step is shown in lines 7 to 25 in the pseudo-algorithm. In the **third step**, the algorithm adds up all the values in the list $Keyphrase_sensitivity$ and divides it by $Total_words$ to get the sensitivity of the input text x . This third step is shown in lines 27 to 33 in the pseudo-algorithm.

D.3 Qualitative analysis of the type of Attacks

Table 10 shows the three different types of attacks carried out by different variants when using **distilbert** as the SMAB Model. ① **Type 1** attack involves generating adversarial examples by replacing one or more words in a sentence with new words and rephrasing the original sentence. Model variant with decoding strategy **BS + (✓)** and a temperature of 0.85 has learned to carry out **Type 1** attacks. ② **Type 2** attack involves generating adversarial examples by removing one or more words from the original sentence while preserving its semantic meaning. Model variant with decoding strategy **BS + (✓)** and a temperature of 1.15 has learned to carry out **Type 2** attacks. ③ **Type 3** attack involves appending various suffixes to sentences, such as adding *but it's true*, or words like *but why* followed by a ?. Model variant with decoding strategy **BS + (✓)** and a temperature of 1 has learned to carry out **Type 3** attacks.

Purpose	Model	Threshold
Paraphraser	prithivida/parrot_paraphraser_on_T5	-
Target (RT)	textattack/distilbert-base-uncased-rotten-tomatoes	-
Sensitivity (RT)	textattack/bert-base-uncased-rotten-tomatoes	-
Linguistic Acceptability	textattack/albert-base-v2-CoLA	0.5
Semantic Consistency	sentence-transformers/paraphrase-MiniLM-L12-v2	0.8
Label Invariance	howey/electra-small-mnli	0.2

Table 6: Various models as target model, in sensitivity reward functions and constraints used in Paraphrase Attack using Local Sensitivity based adversarial example generation experiment. **RT**: Rotten Tomatoes dataset.

Strategy	Temp	Total generated paraphrases	Total used paraphrases
SMAB Model → \mathcal{X}			
DBS + (\mathcal{X})	0.85	62	62
BS + (\mathcal{X})	0.85	53	53
BS + (\mathcal{X})	1.00	110	100
BS + (\mathcal{X})	1.15	68	68
SMAB Model → BERT			
DBS + (\checkmark)	0.85	181	100
BS + (\checkmark)	0.85	71	71
BS + (\checkmark)	1.00	79	79
BS + (\checkmark)	1.15	183	100
SMAB Model → DistilBERT			
DBS + (\checkmark)	0.85	319	100
BS + (\checkmark)	0.85	140	100
BS + (\checkmark)	1.00	334	100
BS + (\checkmark)	1.15	75	75

Table 7: Total unique paraphrases generated by each model variation from a total of 359 test instances that resulted in a label flip and the total paraphrases selected for human evaluation for the **Rotten Tomatoes** dataset.

Strategy	Temp	FSR \uparrow	After Attack Accuracy \downarrow
SMAB Model → \mathcal{X}			
DBS + (\mathcal{X})	0.85	24.79	69.35
BS + (\mathcal{X})	0.85	15.87	82.72
BS + (\mathcal{X})	1.00	17.54	81.05
BS + (\mathcal{X})	1.15	13.64	85.23
SMAB Model → BERT			
DBS + (\checkmark)	0.85	44.87	49.58
BS + (\checkmark)	0.85	18.10	80.22
BS + (\checkmark)	1.00	20.33	77.99
BS + (\checkmark)	1.15	47.62	49.02
SMAB Model → DistilBERT			
DBS + (\checkmark)	0.85	86.35	11.14
BS + (\checkmark)	0.85	35.93	61.00
BS + (\checkmark)	1.00	92.20	6.96
BS + (\checkmark)	1.15	20.33	79.10

Table 8: ParaphraseAttack Results on Rotten Tomatoes (**RT**) dataset. The strategy represents the combination of decoding mechanism (**DBS**: Diverse Beam Search, **BS**: Beam Search) and usage of sensitivity reward (\mathcal{X} : No sensitivity, \checkmark : Sensitivity using bert/distilbert). Temp: decoding temperature. FSR/After Attack Accuracy in (%). **Before Attack Accuracy** is 100%. All variants outperforming the highest baseline score are highlighted in green and the best performing variant is highlighted in green.

Task	Dataset	Target Classifier	Languages	#datapoints	#arms	#edges
Sentiment Analysis	CheckList	cardiffnlp/twitter-roberta-base-sentiment-latest	english	34486	8498	52359
	RT	textattack/distilbert-base-uncased-rotten-tomatoes	english	1066	5104	18835
	SST-2	distilbert/distilbert-base-uncased-finetuned-sst-2-english gpt-3.5-turbo meta-llama/Llama-2-7b	english	872	4088	8041
			english	872	4088	8041
			english	872	4088	8041
Hate Classification	Multilingual hate	google-bert/bert-base-multilingual-uncased google/flan-t5-large MoritzLaurer/deberta-v3-base-zeroshot-v1 gpt-3.5-turbo	bengali	2000	7880	18929
			english	2684	7061	23149
			french	2014	5680	14184
			german	2511	11767	24226
			greek	1601	9799	22031
			hindi	1332	7181	20140
			italian	1846	7551	17876
			spanish	2150	8780	21982
			turkish	1690	12308	20631
Natural Language Inference	XNLI	MoritzLaurer/mDeBERTa-v3-base-mnli-xnli	english	5010	9543	68156
			french	5010	12431	75322
			german	5010	13099	61780
			greek	5010	15114	90149
			hindi	5010	9484	70938
spanish	5010	11994	69166			

Table 9: Training details of SMAB for each task and target classifier and the languages. #datapoints represent the number of sentences in the training set, #arms represents the unique words present in the dataset after preprocessing.

Attack	Examples	Flip Result
Type 1	Original: smarter than its commercials make it seem. Perturbed: it's smarter than the commercials make it appear.	pos → neg
	Original: it has become apparent that the franchise's best years are long past. Perturbed: it's clear that the best years of the franchise are long gone.	neg → pos
Type 2	Original: crush is so warm and fuzzy you might be able to forgive its mean-spirited second half. Perturbed: crush is so warm and fuzzy you might forgive its mean-spirited second half	pos → neg
	Original: you can practically hear george orwell turning over. Perturbed: You can hear george orwell turning over.	neg → pos
Type 3	Original: provides a porthole into that noble, trembling incoherence that defines us all. Perturbed: It provides a porthole into that noble trembling incoherence that defines us all. But why?	pos → neg
	Original: this feature is about as necessary as a hole in the head. Perturbed: This feature is about as necessary as a hole in the head. But it's true.	neg → pos

Table 10: Examples of successful adversarial attacks across the three different attack types when using DistilBERT as the SMAB Model. The **Flip Result** column shows label changes from Original Label → New Label, where pos and neg denote *positive* and *negative* classes respectively in the **Rotten Tomatoes** dataset.

LLM	Prompt
GPT-3.5	Given a sentence that is a movie review, your task is to assign a label based on its sentiment. Label 1 if the sentence is a positive review and Label 0 if the sentence is a negative review. Remember to only provide the label. Sentence: [input_sentence] Label:
Llama-2-7B	Please label the sentiment of the given movie review text. The sentiment label should be "positive" or "negative". Answer only a single word for the sentiment label. Do not leave answer as empty. Do not generate any extra text. Text: [input_sentence] Answer:
Qwen-2.5-7B	Please label the sentiment of the given movie review text. The sentiment label should be "positive" or "negative". Answer only a single word for the sentiment label. Do not leave answer as empty. Do not generate any extra text. Text: [input_sentence] Answer:

Table 11: Prompts used for sentiment classification task on SST-2 dataset

Perturb Type	Qualitative Example	Flip Result
W4	Original Sentence: the title not only describes its main characters, but the lazy people behind the camera as well. New Sentence: The title not only portrays its main protagonists, but also the laid-back crew behind the camera as well.	neg → pos
W5	Original Sentence: an important movie, a reminder of the power of film to move us and to make us examine our values. New Sentence: an important movie, a reminder of the influence of cinema to move us and to force us to question our beliefs.	pos → neg
W6	Original Sentence: It's a charming and often affecting journey. New Sentence: It's a charming but sometimes disconcerting journey.	pos → neg

Table 12: Qualitative examples for perturbation types W_4 , W_5 , and W_6 when using Llama-2-13B both as the as the SMAB Model and the target Model. The **Flip Result** column shows label changes from Original Label → New Label, where pos and neg denote *positive* and *negative* classes respectively in the **SST-2** dataset. In W_4 , the placeholders "[Word1]", "[Word2]", GS1, and GS2 are substituted with *characters*, *people*, 0.9865988772394045, and 0.968535617081986 respectively. In W_5 , the placeholders "[Word1]" and "[Word2]" are substituted with *make* and *film* respectively. In W_6 , the placeholders "[Word1]" and "[Word2]" are substituted with *affecting* and *often* respectively.

Algorithm 2 Algorithm to calculate the sensitivity of Input Text

Input: $x \leftarrow$ Input Text**Require:** $M \leftarrow$ TopicRank keyphrase extraction model, $C \leftarrow$ Classifier Model, $m \leftarrow$ bert-large-uncased Model**Representation:** $G(x) \leftarrow$ Gold label of the Input Text x , $C(x) \leftarrow$ Predicted label of the Input Text x using the Classifier Model C , $m(x) \leftarrow$ generates $K = 10$ perturbations of the input text x using the model m **Output:** $s \leftarrow$ Sensitivity of the Input Text

```
1: // Initialization
2:  $s \leftarrow 0$  ▷ This variable will store the sensitivity of the input text  $x$ .
3:  $Total\_words \leftarrow 0$  ▷ This variable represents the total number of words across all the keyphrases.
4:  $Keyphrase\_sensitivity \leftarrow \{\}$  ▷ A dictionary to store the sensitivity values of each keyphrase.
5: // Extract keyphrases using the TopicRank Model  $M$  and store them in the list  $K$ .
6:  $K \leftarrow M(x)$ 
7: // Main algorithm
8: for  $k$  in  $K$  do
9:    $words \leftarrow k.split()$  ▷ Get a list of all the words in the current keyphrase.
10:   $Total\_words += len(words)$ 
11:  for  $w$  in  $words$  do
12:     $x_{masked} \leftarrow$  mask the word  $w$  in the text  $x$  using "[MASK]"
13:  end for
14:   $Masked\_output \leftarrow m(x_{masked})$  ▷ Store all the perturbed samples in this list.
15:   $Prediction\_List \leftarrow C(Masked\_output)$  ▷ Store the prediction labels of all the perturbed samples in this list.
16:   $Prediction\_original \leftarrow C(x)$  ▷ Store the prediction label of the input text  $x$ .
17:   $flips \leftarrow 0$  ▷ A temporary variable to count the flips.
18:  for  $labels$  in  $Prediction\_List$  do
19:    if  $labels \neq Prediction\_original$  then
20:       $flips += 1$ 
21:    end if
22:  end for
23:   $local\_sensitivity \leftarrow flips/len(Prediction\_List)$  ▷ Total flips normalized by the total number of labels.
24:   $Keyphrase\_sensitivity[k] \leftarrow local\_sensitivity$ 
25: end for
26: // Calculate the sensitivity of the input text  $x$ .
27: if  $Total\_words \neq 0$  then
28:    $t \leftarrow 0$  ▷ A temporary variable.
29:   for  $value$  in  $Keyphrase\_sensitivity.values()$  do
30:      $t += value$ 
31:   end for
32:    $s \leftarrow t/Total\_words$ 
33: end if
return  $s$ 
```
