Attacking text is achievable and can have a significant impact in certain fields. Below, I will describe a scenario of text attacks from several aspects, including a concrete scenario, the aftermath of it, the reason it's dangerous, and in what way it will turn out to be harmful. The details are as follows:

（1）A concrete scenario

In the current social environment, with the popularization of computer technology for civilian use, more and more applications tend to recommend to corresponding targets through big data based on public evaluations, such as movie recommendations, takeout reviews, restaurant and scenic area reviews, and even public opinion effects on YouTube and Twitter. These reviews are crucial for newly registered users, This is the key to judging whether a goal is worth choosing or trusting. This type of comment screening is a classic sentiment analysis system that is susceptible to adversarial sample attacks. In the information I searched online, there is an implementation of ParallelDots' adversarial attack sentiment analysis API, which only needs to change specific words or generate adversarial samples by adding noise to the text (DNN is a typical example) to change the final sentiment judgment result (such as changing from negative to positive or vice versa), This will create a huge difference in the final submitted results. Over time, it will have a serious impact on a store, a landscape, and even the evaluation of a certain person or region, resulting in a black and white effect.

（2）The aftermath of it, the reason it’s dangerous, and in what way it will turn out to be harmful.

Among the scenarios I have listed above, I would like to elaborate on the consequences, reasons, and ways in which they will be proven harmful from both the public and social perspectives. People pay more attention to civil apps, such as uber and other take out taxi software, Netflix and other video software. This attack on comment text will seriously reverse the black and white effect, making many unscrupulous businesses profit from it, while harming the income and ratings of many conscientious businesses or conscientious movies. This not only damages the reputation of the platform, but also causes profound damage and impact to individual income or reputation. This result will receive feedback from other platforms or channels after many people visit, ultimately making the truth known to the public. However, the trauma guided by public opinion is difficult to repair. At the social level, there are many official channels for national governments to convey information on YouTube and Twitter, and many people learn about the trends of the world's international landscape through this way. However, organized or serious text attacks can change the attitude of the country or government towards publishing, or modify the attitude in the comment section to guide public opinion, This approach may reverse attitudes among different governments that were originally in neutral or partnership relationships, especially in triggering public opinion among the public and generating hatred towards an unrelated ethnic group. Even if network platform administrators make adjustments and explanations in the future, it may be difficult to recover the negative effects caused by flooding, and it may take a long time to alleviate them before they can return to their original state. In the 21st century with the development of the internet, the internet has become a place with relatively shallow legal jurisdiction, which has led many criminals to guide public opinion through text attacks similar to nlp, like the forest fire that erupts in a short period of time, ultimately leaving wounds that are difficult to heal even if eliminated.

The input of NLP is discrete, so when attacking it, it is not as simple as CV. CV attacks only require adding a certain noise (any noise) to the image, while NLP attacks are more complex. Assuming there is a sentence, we cannot simply add noise to words to change the prediction. This idea is like adding noise to word embedding, which is unreasonable behavior, Because the model internally converts tokens into continuous embeddings. And what we can operate on are only discrete tokens of input, which cannot change anything else. This is extremely unique compared to other models (CVs), with specific methods such as Estimation Attacks, Impetation Attacks, and Backdoor Attacks. Specific details such as text obfuscation attacks can have an impact on the model's production, which also indicates that the NLP model is relatively fragile.

1. Goal: What the attack aims to achieve. 2. Transformations: How to construct perturbations for possibleadversaries. 3. Constrains: What a valid adversarial example should satisfy. 4. Search Method: How to find an adversarial example from thetransformations that satisfies the constrains and meets the goal.

What I really like and would like to introduce is the TextFooter method. According to the combination of four ingredients of evaluation attachments, its Goal is Untargeted Classification; Constraints are Word embedding distance, USE sensitivity similarity, and POS consistency; Transformation lies in the Word substitution by counter filled GloVe embedding space; The final Search Method is Greedy word search with importance ranking. These important components are reflected in the algorithm of textfooler, and their specific working principles are as follows: (1) The USE sense similarity in Constraints can limit the similarity between words and sentences, and POS consistency can restrict grammar. Therefore, in the algorithm, its priority use before the search process will filter out some commonly occurring words such as semi nouns and articles, Replacing these words will affect grammar to achieve the final effect; (2) The use of GloVe in Transformation is because the model places synonyms in similar spaces when implementing counter rated, thus achieving the effect of synonym replacement; (3) The search process now requires transformation substitution, but the substitution also requires the participation of constraints (such as the minimum similarity between the final result sentence and the original sentence). Simply put, the function of the search component is to combine different transformations and combine some constraints during the transformation process to eliminate unreasonable ones, which is the core process of the algorithm; (4) In addition, in the preparation work of the algorithm, the importance of each word will be calculated through the WIR process to pave the way for subsequent processing and search. Overall, TextFooter is an effective attack method that can generate adversarial examples for various NLP tasks. Its components work together to modify input text in difficult to classify situations and can be used to evaluate the robustness of NLP models against attacks