OVERVIEW OF THE ATTACK ALGORITHM

The algorithm consists of three functions:

1) Evaluate_Word_Saliency
2) Find_Similar_Words
3) Generate_Adversary

- Evaluate_Word_Saliency:
  - Determine importance of each word in a review. This is done by removing one word at a time from a review and measuring the difference in confidence of the model before and after the word was replaced. Stop words are not considered.
  - Words are sorted in decreasing order based on their importance. For reviews with more than words, the top three words are kept, and rest are discarded. For reviews with less than 3 words, all the words are kept.
  - This function outputs a list of all salient words of a review for all reviews. Each review has at most 3 salient words.

- Find_Similar_Words:
  - This function takes the salient words of the reviews computed in the previous function as input.
  - For each of the salient words in a review, 50 most similar words are extracted from a pretrained fastText word embedding. The salient word in the review is replaced by each of the 50 words one at a time and fed to the model. A word is considered if:
    - Substituting the word inplace of the original word changes the prediction of the model and the adversarial review and the original review are 70% similar in context. The context similarity is determined by Universal Sentence Encoder.
    - Substituting the word inplace of the original word leads to a difference of 0.4 in the confidence of the model for the true label compared to the original confidence and is 70% similar in context as mentioned above.
    - Substituting the word inplace of the original word leads to the maximum difference in the confidence of the model for the true label compared to the original confidence and is 70% similar in context as mentioned above.

  The first similar word that satisfies any of the first two conditions are considered and the process is stopped for that salient word. If none of the similar words satisfy any of the first two conditions then, the word that satisfies the third condition is picked.

  - This function outputs the most similar word for a salient word of a review. From the last function, we know that each review has at most 3 salient words. And each salient word has similar words.

- Generate_Adversary:
  - This function takes in the output of the previous function as input. We have at most 3 salient words for a review. The goal is to replace only one word from the original review.
  - For a review, each of the salient words are replaced by their corresponding similar words and fed into the model. The salient word whose similar word changes the prediction of the model as well as have a 70% similarity in context to the original review is considered and the process is stopped. If no such word exists, then the salient word whose similar word leads to the maximum difference in the confidence of the model and have a 70% similarity in context to the original review is considered.

This algorithm is based on word saliency and FGSM and is inspired by https://arxiv.org/abs/1907.11932 and https://www.aclweb.org/anthology/P19-1103/ . While the underlying idea behind the algorithm is not novel, the implementation is.

The key differences between the algorithms mentioned in the papers and my algorithm lie in the similar word's selection and substitution strategy.

- These papers use a synonym-based word similarity. The problem with this approach is that not all words have synonyms. My algorithm does not choose similar words based on synonyms. I am using pre-trained fasttext embeddings. This allows to find similar words for words without synonyms, misspelled words, concatenation of words.
- My algorithm alters only one word of the original review. Altering only one word especially in a long review makes it hard for humans to recognize the adversary reviews.
- Multiple layers of checks are put in to maintain review context similarity of 70% between the original and the adversarial reviews.

RESULTS

NOTE: Adversarial reviews are generated from reviews that were correctly predicted by the target model.

CMU MOSI

Original Reviews	Adversarial Reviews	Original Prediction	Prediction After Attack
i mean the blood is like twilight with some scifi with some action and with **likable** characters and that that is important guys	i mean the blood is like twilight with some scifi with some action and with **lovable** characters and that that is important guys	Positive	Negative
it was not a **real** huge saw twist moment	it was not a **great** huge saw twist moment	Negative	Positive
at how much i really did **enjoy** it	at how much i really did **enjoys** it	Positive	Negative
but i feel that the movie **wasnt** really concerned with being exciting	but i feel that the movie **werent** really concerned with being exciting	Negative	Positive
i mean corpse bride **hes** really good in corpse bride	i mean corpse bride **isnt** really good in corpse bride	Positive	Negative
but i just **couldnt** find it funny	but i just **couldent** find it funny	Negative	Positive
the premiere itself was really **cool**	the premiere itself was really **coolest**	Positive	Negative
yeap a **horrible** protagonist	yeap a **horrid** protagonist	Negative	Positive
um like i said it **wasnt** big deal	um like i said it **havent** big deal	Negative	Positive
i **really** did	i **actually** did	Positive	Negative
that so what is still a really **great** movie to take your whole family to	that so what is still a really **tremendous** movie to take your whole family to	Positive	Negative
and and we kind of wanted it to be like but exept not that obvoious but that wouldve been **good** right like coz that it would give tham a reason coz they wanted to get married but they couldnc because she was gonna marry the other guy	and and we kind of wanted it to be like but exept not that obvoious but that wouldve been **goood** right like coz that it would give tham a reason coz they wanted to get married but they couldnc because she was gonna marry the other guy	Positive	Negative
i will even **pick** up this movie on bluray	i will even **picking** up this movie on bluray	Positive	Negative

TWITTER SENTIMENT DATASET

Original Tweets	Adversarial Tweets	Original Prediction	Prediction After Attack
is up and getting **ready** for work	is up and getting **finished** for work	Negative	Positive
i think that it cell phone had language options like on a computer that would just be **amazing** xmx	i think that it cell phone had language options like on a computer that would just be **jaw-dropping** xmx	Positive	Negative
awww **thanks** i am a regular to your site aswell	awww **sorry** i am a regular to your site aswell	Positive	Negative

dead battery right in the middle of an alexisonfire song	**doornail** battery right in the middle of an alexisonfire song	Negative	Positive
aww my sister needs tech support from college at 1 11am good thing big brother that s me is in the ha s sucks for anyone w o sibling	**cute.** my sister needs tech support from college at 1 11am good thing big brother that s me is in the ha s sucks for anyone w o sibling	Negative	Positive
yesss it was **awesome** meeting u yesterday	yesss it was **mind-blowing** meeting u yesterday	Positive	Negative
mb **wish** i was in town	mb **prefer** i was in town	Negative	Positive