Adversarial Examples for Text Classification

Anonymous Author(s)
Affiliation
Address
email

Abstract

Natural Language processing helps the computer to understand the way that the humans write and speak as it is a very complex task because of the involvement of large amount of unstructured data. Through Natural Language Processing, computer can effectively communicate with the humans in their own language, along with it also helps to scale the language related tasks as well. As Natural Language Processing makes it easy for the computer to read the text, hear the speech, measure the sentiment and also determines which parts are important. Now a days, with the evolution of Natural Language Processing, machines can now easily process large number of language based data than the humans in a more compatible and objective way. As a large amount of un structured data is being generated each day, from different social media sites, websites, through different contents, from office records to the medical records, so there is an need of an automation that can fully analyze the text data in an efficient way. The major motivation behind working over Natural Language Processing is that it makes it possible for the computers to read the text, hear the speech, measures the sentiment and also determines which part are important as well. Modern text classification models are basically vulnerable to the opposite example. Discomposed versions of the original text which are indistinguishable by the humans get misclassified by the models. In the recent research, rule based synonym replacement strategies have been considered to generate the adversarial examples. But the limitation of this approach is that it leads to out of the context and very complex tokens replacements. In this research, we formulate attacks against a trained model (LSTM Neural Networks, Naive Bayes Classifier, Random Forest Classifier) by testing it against inputs that are semantically similar to the original but have slight paraphrasing or synonym substitutions. Through result analysis, it can be seen that the model trained on the augmented data performed better and gives better accuracy, against perturbations and yielded nearly double the amount of failed attacks.

1 Introduction and Motivation

Before discussing adversarial text attacks in NLP we need to know what Adversarial Examples are, adversarial example is a fake data that mimics the training data but produces misclassified label when the machine learning algorithm encounters it. On the other hand adversarial attack is an pipeline which generates the adversarial examples Moustafa Alzantot et al. [1].

Deep Neural Network has been very popular since from the start and AlexNet performance increased the hope in deep learning. One of the major problems faced in deep neural network is to design such

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Positive (77%)**
Adversarial example [Visually similar]	**Aonnoisseurs** of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (52%)**
Adversarial example [Semantically similar]	Connoisseurs of Chinese **footage** will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (54%)**

Figure 1: Adversarial Examples in NLP

an deep neural network that is robust at detecting the changing and resisting against the adversarial examples Jacob Devlin et al. [2].

The concept of the black box attack was originally introduced in the computer vision domain which is easier to mimic the original data by infusing fuzzy data (noise) to the original data represented in the continuous pixel intensity. Unlike, the computer vision, in NLP we have an concept of tokens, NLP utilizes the discrete word as tokens and it is often very challenging to swap out tokens without changing the meaning of the sentences, sentiments, implications and syntax of a sentence Javid Ebrahimi et al. [3].

The motivation behind working on the adversarial text attacks in NLP is that, usually the data is imbalanced and to balance the data usually Up Sampling and Text Augmentation, SMOTE, ENN and KNN are performed to balance the data. In some of these approaches the data is balanced by the generating the copy of the data from the original data. In some cases the data is balanced by inserting, deleting words, replacing them with synonyms and so. But it must be noted that swapping an input's words or randomly deleting several letters can severely alter the performance of our Natural Language Processing models (NLP). We formulate attacks against a trained model (i.e LSTM Neural Networks, Random Forest Classifier, Naive Bayes Classifier)by testing it against inputs that are semantically similar to the original but have slight paraphrasing or synonym substitutions. In this research Text Attack library is used and an adversarial attack is run on each of the trained models (the one trained with up sampled data and the one with augmented data). The attack will run until 1000 attacks are successful at fooling each model.

The research contribution is as follows

• The data is balanced using the multiple approaches (i.e. using Up Sampling, Text Augmentation and using SMOTE). Along with this the text is converted into vectors using TF-IDF and the count vectorizer and the different Machine Learning and Deep Learning models (LSTM Neural Network, Naive Bayes Classifier, AdaBoost Classifier and the AdaBoost Classifier) are trained using the balanced datasets.

• An adversarial attack is formulated against the trained models by testing it against the inputs that are semantically similar to the original but have slight paraphrasing or synonym substitutions.

• A performance comparison of each of the model trained on the balanced data (i.e. the data balanced using Up Sampling, text augmentation and through SMOTE) is done considering all the scenarios, the original accuracy, accuracy under attack, the number of successful attacks, the number of failed attacks, the number of skipped attacks.

2 Related Work

The recent research have shown the possibility of the machine learning models being attacked or exposed to adversarial attacks i.e. it is an approach which tries to deceive/ fool the model with the deceptive data and it has emerged as an major challenge in the field of artificial intelligence and machine learning, small input perturbation in the system, i.e. small change in the system which may be as a result of third object interacting with the system, and it results in the misclassification by the machine learning model. Adversarial attacks or adversarial example is an major challenge in the field of Natural Language processing and it is quite a complex and major problem as compared to computer vision tasks.

In the initial work, the text models were attacked considering the introduction of the error at the character level or adding or deleting the words. These approaches result in an very un natural adversarial examples that lacks the grammatical correctness and thus can be easily identified by the humans.

The rule based synonym replacement approaches have resulted in generating more natural looking adversarial examples. Nicolas Papernot et al. combined all the previous approaches and proposed an TextFooler which is an strong black box attack baseline used in the text classification model. The adversarial examples being generated through the Textfooler only considers the token level similarity using the word embeddings and does not considers the overall sentence structure, which results in a out of context and an very un naturally complex replacement.

In extension of topics in biomedical decision making, vulnerabilities in biomedical NLP could be devastating for medical decision tasks. If wrong decision is proposed by AI, it will negatively impact the patients' health. Adversarial Examples for Biomedical NLP Tasks discusses adversarial examples generated by BERT-based model in BioNLP

Besides performing tasks in text classification in general NLP domains,it is also important to leverage existing techniques for improving detections of adversarial examples in other domains, ie. medical domain in support of fraudulent billing activities.

In this paper we have proposed a novel approach, by using the text attack library an adversarial attack is run on each of the trained models i.e (LSTM Neural Network, Random Forest Classifier, Naive Bayes Classifier), and the models are trained on the balanced data sets as the data is balanced using multiple approaches i.e using SMOTE, up sampling and the augmented text data approach, the performance comparison of each of the model is done considering all the scenarios considering the original accuracy, accuracy under attack, the number of successful attacks, the number of failed attacks, the number of skipped attacks are considered as well

3 Datasets and Methods

To analyze the performance of the proposed approach the dataset considered for the implementation is Women's E-Commerce Clothing Reviews. The data-set is available publically on Kaggle. The dataset contains 23486 rows and 10 feature variables. In the proposed approach only two feature variables are considered i.e "Review Text" and "Recommended IND", as this paper is focused on text classification.

In the "Recommended IND" column, the 1 represents the "Recommended" and 0 represents the "Not Recommended". If we take a closer look at the "Recommended IND" column, we can see that the data is imbalanced, to solve this issue, Up Sampling and Text Augmentation is performed to balanced the dataset.

Four different machine learning algorithms are used to train the model on the given dataset. The model used for the training include the LSTM Neural Network, Random Forest Classifier and Naive Bayes Classifier. Each of the model is trained considering bot the up sampled data as well as the augmneted data.

In the next step, using the TextAttack library, an adversarial attack is run on each of the trained models (the one trained with upsampled data and the one with augmented data). The attack will run until 1000 attacks are successful at fooling each model.

4 Experiment and Results

The first step, involves installing all the required libraries and importing them. After this the data-set is loaded and the data analysis and visualization is performed to find the hidden insights in the data and to analyze the data. The data-set contains 23486 rows and 10 feature variables. In the proposed approach only two feature variables are considered i.e "Review Text" and "Recommended IND", as this paper is focused on text classification. The "1" in the "Recommended IND" represents the recommended and the "0" represents the non recommended. After the data analysis and visualization the data-set is cleaned, all the missing values are removed and outliers are handled as well. The "Recommended IND" column contains the imbalanced data i.e the number of 1's which represents the recommended are more than the 0's which represents the non recommended. In this case if we do not balance the data-set. The probability of prediction accuracy of "0" will be very less as compared to the "1". So, it is needed to balance the dataset. There exists multiple approaches to balance the dataset which included ENN< KNN, Up Sampling and Text Augmentation technique. Each technique has its own pros and cons. In this research, Up sampling and text augmentation are considered to increase the number of 0 in the "Recommended IND" column from 4101 to 16104. Up sampling technique works by generating the copies of the original data and increasing the data size. While in the text augmentation, words are randomly swapped, deleted, as well as replaced or inserted with synonyms using pretrained word embeddings. In this research Easy Data Augmentation technique is used.

After this LSTM Neural Network, Naive Bayes Classifier and Random Forest Classifier model is trained on this dataset (with upsampled data and the one with augmented data). The evaluation of all three models is done considering the accuracy score, percision score, recall score and the f1-score.

The data set is spllited into the train and test set, and 80 percent of the data is used for the training purpose. After converting he text into the vectors and training the LSTM neural network on the training data set and evaluating the model on the test data 85 percent is obtained on the test, along with an precision of 80 percent for the 0 and 89 percent for the 1. In this case no adversarial attack is run on the trained model. The confusion matrix obtained is as follows as shown in Figure 3.

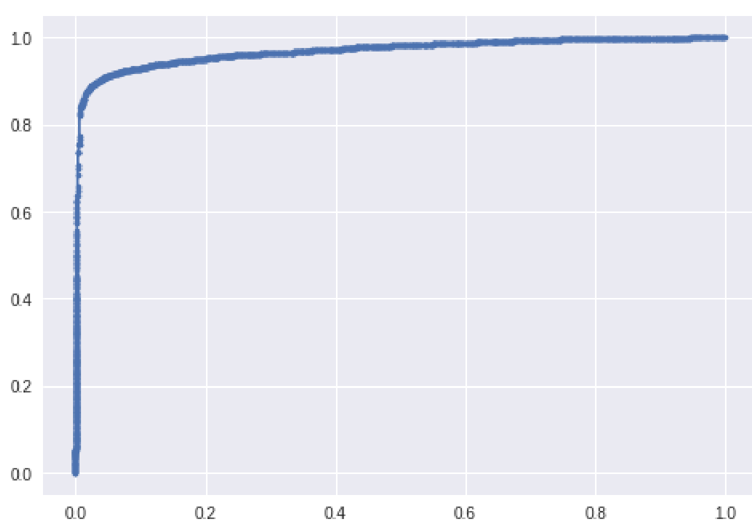

Figure 2: ROC Curve of LSTM Neural Network with out any Adversarial Attack

In the next step, after converting the text into vectors using the TF-IDF and training the model using Random Forest Classifier. From the results it can be seen that an accuracy of 97 percent is obtained on the test data set. Along with this a precision of 0.94 is obtained for the 0 and a precision of 0.99 for the one. Following is the confusion matrix obtained as shown in the Figure 4

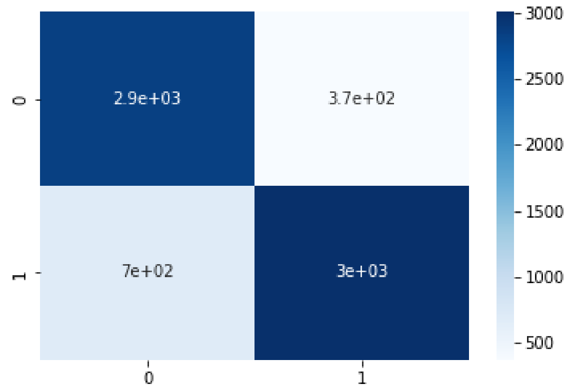

Figure 3: Confusion Matrix for the LSTM Neural Network with out any Adversarial Attack

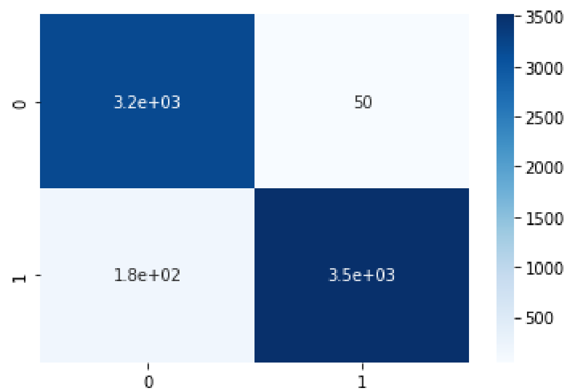

Figure 4: Confusion Matrix for the Random Forest Classifier with out any Adversarial Attack

Bag of Words can also be used to convert text into vectors, after converting the text into vectors using count vectorizer and implementing the Naive Bayes Classifier on the training dataset and evaluating the model on the test data set. Following is the confusion matrix obtained from the Naive Bayes Classifier model on the test data set as shown in Figure 5.

In the next step, using the Text Attack library, an adversarial attack is run on each of the trained models i.e. LSTM Neural Network, Random Forest Classifier and the Naive Bayes Classifier and

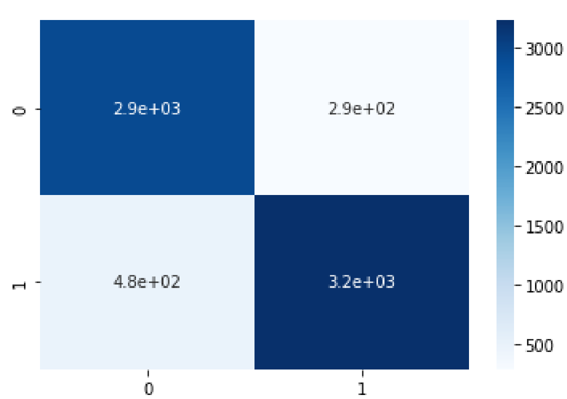

Figure 5: Confusion Matrix for the Naive Bayes Classifier with out any Adversarial Attack

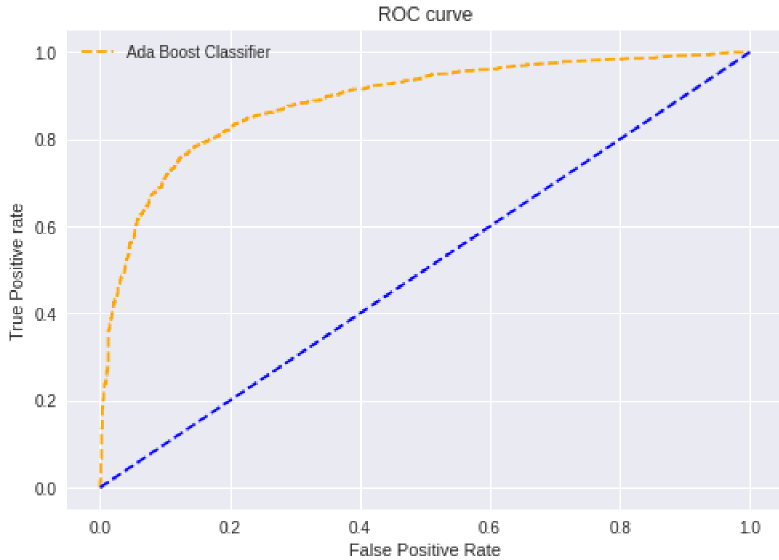

Figure 6: ROC Curve for the AdaBoost Classifier with out any Adversarial Attack

Attack Results	
Number of successful attacks:	1000
Number of failed attacks:	132
Number of skipped attacks:	35
Original accuracy:	97.0%
Accuracy under attack:	11.31%
Attack success rate:	88.34%
Average perturbed word %:	9.04%
Average num. words per input:	61.2
Avg num queries:	81.04

Figure 7: Attack Results using the LSTM Neural Network considering the Up Sampled Data set

AdaBoost classifier (the one trained with up sampled data, the one with augmented data and the one in which the dataset is balanced using SMOTE). The attack will as per the conditions defined by the user, in the case of LSTM Neural Network considering the Up Sample dataset's, the attack will run until 1000 attacks are successful at fooling the model while on the other hand in the case of LSTM Neural Network considering the augmented text data set, the attack will run until 10 attacks are successful at fooling the model and is shown in Figure 7 and Figure 8 as well. In the Figure 9, attack result using the LSTM neural network considering the data set balanced using SMOTE are given but it can be seen that the results obtained using SMOTE are not very satisfactory as the accuracy under attack is around 12 percent which is very less as compared to the accuracy under attack considering the LSTM Neural Network with the augmented text data set.

In the Figure 10, from the attack results using the AdaBoost Classifier considering the dataset balanced using SMOTE are given, from the results it can be analyzed that the original accuracy was 91.94 percent, while the accuracy under attack is 11.29 percent. Along with this, it can also be analyzed that the attack success rate is very high. In the Figure 11, from the attack results using the Random Forest Classifier considering the dataset balanced using Up Sampling are given, from the results it can be analyzed that the original accuracy was 99.94 percent, while the accuracy under attack is 0.0 percent. Along with this, it can also be analyzed that the attack success rate is 100 percent.

In the Figure 12, few examples of the original text vs the perturbed text results are shown considering the LSTM Neural Network with the Up Sampled Text Dataset. Considering the attack success rate

Attack Results	
Number of successful attacks:	10
Number of failed attacks:	14
Number of skipped attacks:	4
Original accuracy:	85.71%
Accuracy under attack:	50.0%
Attack success rate:	41.67%
Average perturbed word %:	9.99%
Average num. words per input:	63.43
Avg num queries:	96.0

Figure 8: Attack Results using the LSTM Neural Network considering the Augmented Text Data set

Attack Results	
Number of successful attacks:	100
Number of failed attacks:	12
Number of skipped attacks:	3
Original accuracy:	97.39%
Accuracy under attack:	10.43%
Attack success rate:	89.29%
Average perturbed word %:	8.95%
Average num. words per input:	58.64
Avg num queries:	75.59

Figure 9: Attack Results using the LSTM Neural Network considering the dataset balanced using SMOTE

i.e. 88.34 percent it can be analyzed that the model doesnot give satisfactory performance in the case of LSTM Neural Network with the up sampled text data set, however a satisfactory performance can be seen in the case of LSTM Neural Network consdering the Augmented Text Dataset.

In the Figure 13, few examples of the original text vs the perturbed text results are shown considering the LSTM Neural Network with the Dataset balanced using SMOTE. Considering the attack success rate i.e. 89.29 percent it can be analyzed that the model doesnot give satisfactory performance in the case of LSTM Neural Network with the data set balanced using SMOTE, however a satisfactory performance can be seen in the case of LSTM Neural Network considering the Augmented Text dataset.

In this research, adversarial attacks were launched on the trained model and a detailed result analysis is presented considering multiple scenarios i.e balancing the data using multiple approaches along with training the Machine Learning and Deep Learning model on different balancing datasets and launching attacks on the trained model and doing the accuracy comparison considering attack success rate, accuracy under attack, number of skipped attacks and the number of successful attacks. From

Attack Results	
Number of successful attacks:	100
Number of failed attacks:	14
Number of skipped attacks:	10
Original accuracy:	91.94%
Accuracy under attack:	11.29%
Attack success rate:	87.72%
Average perturbed word %:	9.79%
Average num. words per input:	63.17
Avg num queries:	85.84

Figure 10: Attack Results using the AdaBoost Classifier considering the dataset balanced using SMOTE

Attack Results	
Number of successful attacks:	100
Number of failed attacks:	0
Number of skipped attacks:	1
Original accuracy:	99.01%
Accuracy under attack:	0.0%
Attack success rate:	100.0%
Average perturbed word %:	6.78%
Average num. words per input:	58.01
Avg num queries:	114.06

Figure 11: Attack Results using the Random Forest Classifier considering the Up Sampled Dataset

	original_text	perturbed_text
0	the sleeves are a bit too voluminous and the solid color collar stands out a bit too much. but it's a cute pattern and style. very comfortable.	the sleeves are a bit too voluminous and the solid colored collar stands out a bit too much. but it's a charmer pattern and style. very cosy.
1	i purchased another top, same brand, in the same style from retailer and really love it. the fabric is lightweight, which makes it great for multiple seasons. i travel extensively and these tops hand wash and hang to dry very nicely. i wear a gray tank under this one and love the look. i also like the long length. as usual, i will probably tack the neckline up an inch or so, so that it's not so low - just my thing. overall, the pattern and the fabric are very nice. i usually wear an xs or s; i pur	i purchased another top, same brand, in the same style from retailer and really adores it. the fabric is lightweight, which makes it tremendous for multiple seasons. i trip extensively and these tops hand wash and hang to dry very politely. i wear a gray tank under this one and love the look. i also like the long length. as usual, i will probably tack the neckline up an inch or so, so that it's not so low - just my thing. overall, the pattern and the fabric are very nice. i usually worn an xs or s; i pur
2	the stone color was nice, but the material is heavy and doesn't breath. if you are partial to more natural fabrics i would go with a different item. upside would be that i imagine this shirt will last.	the stone coloured was nice, but the materials is hefty and doesn't breathing. if you are partial to more natural fabrics i would go with a different item. upside would be that i imagine this shirt will last.
3	i wanted to love this dress but it was too large... and once i washed it (which is recommended) it shrank so i had to return it.	i wanna to love this dress but it was too considerable... and once i washed it (which is recommending) it shrank so i had to comeback it.

Figure 12: Original Text vs Perturbed Text using the LSTM Neural Network with the Up Sampled Text Data set

the results it can be analyzed that the model trained on a dataset with augmented data outperformed all-around. In comparison, it had significantly better accuracy against perturbations and yielded nearly double the amount of failed attacks

	original_text	perturbed_text
0	the sleeves are a bit too voluminous and the solid color collar stands out a bit too much. but it's a cute pattern and style. very comfortable.	the sleeves are a bit too voluminous and the solid colored collar stands out a bit too much. but it's a charmer pattern and style. very cosy.
1	i purchased another top, same brand, in the same style from retailer and really love it. the fabric is lightweight, which makes it great for multiple seasons. i travel extensively and these tops hand wash and hang to dry very nicely. i wear a gray tank under this one and love the look. i also like the long length. as usual, i will probably tack the neckline up an inch or so, so that it's not so low - just my thing. overall, the pattern and the fabric are very nice. i usually wear an xs or s; i pur	i purchased another top, same brand, in the same style from retailer and really adores it. the fabric is lightweight, which makes it tremendous for multiple seasons. i trip extensively and these tops hand wash and hang to dry very politely. i wear a gray tank under this one and love the look. i also like the long length. as usual, i will probably tack the neckline up an inch or so, so that it's not so low - just my thing. overall, the pattern and the fabric are very nice. i usually worn an xs or s; i pur
2	the stone color was nice, but the material is heavy and doesn't breath. if you are partial to more natural fabrics i would go with a different item. upside would be that i imagine this shirt will last.	the stone coloured was nice, but the materials is hefty and doesn't breathing. if you are partial to more natural fabrics i would go with a different item. upside would be that i imagine this shirt will last.
3	i wanted to love this dress but it was too large... and once i washed it (which is recommended) it shrank so i had to return it.	i wanna to love this dress but it was too considerable... and once i washed it (which is recommending) it shrank so i had to comeback it.

Figure 13: Original Text vs Perturbed Text using the LSTM Neural Network with the Data set balanced using SMOTE

Discussion and Conclusion

In real life or in the case of many machine learning and deep learning problems usually the data is imbalanced and to balance the data usually Up Sampling and Text Augmentation, SMOTE, ENN and KNN are performed to balance the data. In some of these approaches the data is balanced by generating the copy of the data from the original data and in some cases the data is balanced by inserting, deleting words, replacing them with synonyms and so. But it must be noted that swapping an input's words or randomly deleting several letters can severely alter the performance of our Natural Language Processing models (NLP). In this project, we balanced our data using multiple approaches i.e. using SMOTE, Augmented Text Data approach and Up Sampling the Data and then after training the Machine Learning and Deep Learning models i.e. LSTM Neural Networks, Random Forest Classifier, Ada Boost Classifier and the Naive Bayes Classifier on each of the balanced data set obtained after implementing SMOTE, Up Sampling and the Text Augmentation, we formulate an adversarial attack using the text attack library on each of the trained model by testing it against inputs that are semantically similar to the original but have slight paraphrasing or synonym substitutions, and then performance comparison of each of the trained model is done considering the attack success rate, accuracy under attack and the average percentage of perturbed words

So after making the data balanced, we trained our model using multiple machine learning and deep learning algorithms (i.e LSTM Neural Networks, Random Forest Classifier, Naive Bayes Classifier) on each of balanced data i.e. (Using SMOTE, ENN, Up Sampling and Text Augmentation) and then formulate attacks against a trained model by testing it against inputs that are semantically similar to the original but have slight paraphrasing or synonym substitutions using the Text Attack Library. Through result analysis, it can be seen that the model trained on the augmented data performed better and gives better accuracy, against perturbations and yielded nearly double the amount of failed attacks.

References

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] 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, pages 31–36, Melbourne, Australia. Association for Computational Linguistics. **15**(7):5249-5262.

[4]Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of ACL.

[5]Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17, pages 506–519, New York, NY, USA. ACM.