Report Of Task - 1

Paper Title: Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers.

Paper Link: https://link.springer.com/article/10.1007/s13042-022-01553-3

Summery:

1. Introduction:

1.1 Motivation:

The study was motivated by the need to enhance text classification models, particularly in the domain of disaster events and sentiment analysis, through the utilization of data augmentation techniques. The objective was to investigate how well the GPT-2 language model could produce a variety of examples that were contextually relevant in order to enhance model training.

1.2 Contribution:

The primary contribution of this research lies in the exploration and application of data augmentation strategies, specifically leveraging the GPT-2 model. The study's objective was to find out how well these methods worked for enhancing text classification models' performance, with a particular emphasis on sentiment analysis and disaster-related tasks.

1.3 Methodology:

The methodology involved the analysis of generated data by selecting instances from the GPT-2 model and comparing them to their original counterparts. The assessment involved calculating the Levenshtein distance for similarity, looking at examples from particular datasets (SST-2 and West Texas Explosions), and determining the different modifications the model made. The paper also covered how the GPT-2 model might produce a variety of examples with novel linguistic elements.

1.4 Conclusion:

In conclusion, the research suggests that the proposed method, utilizing GPT-2 for data augmentation, is capable of performing various transformations in text instances. The model demonstrated how to alter the generated data by removing, truncating, interpolating, enlarging, and substituting words. The study acknowledged instances of repetition and a rise in duplicates with the amount of created data, even though it showed diversity.

2. Limitations:

2.1 First Limitation:

One of the limitations identified is the challenge of finding similar counterparts for all generated instances. This raises questions about the origin of some examples and suggests that the model may be learning from diverse sources, impacting the interpretability of the augmented data.

2.2 Second Limitation:

The model's propensity to produce duplicates has also been noted as a weakness, particularly as the number of created data rises. This is explained by the higher likelihood of repeat resulting from the inherent probabilities associated with specific token sequences.

3. Synthesis:

The synthesis of the findings indicates that although the GPT-2 model adds diversity and carries out different transformations, there are certain limitations that call for more research.Future research could explore the optimal amount of generated data for maximum improvement and assess the impact of language model size on augmentation effectiveness. Despite limitations, the study offers insightful information about GPT-2's potential for data augmentation in text classification tasks.