# Enhancing sentiment analysis corpus through paraphrasing

Pablo Calcina Ccori[1], Carlos E. Atencio-Torres[2], and Julio Vera-Sancho[2]

[1,2,3] *Universidad Nacional de San Agustin, Peru,*
[1] pcalcinacc@unsa.edu.pe
[2] catencio@unsa.edu.pe
[3] jveras@unsa.edu.pe

**Keywords:** Corpus, Bert, Paraphrasing, Parrot, Pegasus

Nowadays, artificial intelligence have presented many solutions to different problems with very accurate results. Specifically in text classification (TC), many works have been proposed using diverse techniques such as Machine Learning (ML) or Deep Learning (DL) [1].

The DL-based solutions for TC usually works on large datasets and obtains good results. Therefore, the quantity and quality of a dataset determines the success of a DL-based solution for TC.

In this presentation, we show a solution to enhance a dataset by paraphrasing sentences of the original corpus. We tested two automatic paraphrasing techniques: Parrot [2] and Pegasus [3] and we prove the augmented corpus on a classifier based on BERT [4] to measure its performance

We used the IMDB dataset containing $50k$ movies reviews categorized in positive and negative . The results showed that without paraphrasing, the Bert model obtained 90% of f-measure; and with paraphrasing we obtained 99.56% and 99.86%.

Table 1: Results of classification methods applying to different corpus

| | Original | **Augmented Parrot** | **Augmented Pegasus** |
|---|---|---|---|
| **Precision** | 0.90722 | 0.99566 | 0.99860 |
| **Recall** | 0.90691 | 0.99565 | 0.99860 |
| **F-measure** | 0.90697 | 0.99565 | 0.99860 |

Finally, we would like to encourage the use of paraphrasing techniques to improve the corpus because it provides new sentences that improve any classifier model. Furthermore, the use of paraphrasing increases the richness of any corpus by adding new vocabulary.

## Bibliography

[1] Arunachalam N., S. J. Sneka, G. MadhuMathi *A survey on text classification techniques for sentiment polarity detection* Innovations in Power and Advanced Computing Technologies (i-PACT), 2017.

[2] P. Pamodara Parrot: Paraphrase generation for NLU 2021.

[3] J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,* Proceedings of the 37th International Conference on Machine Learning , 2020.

[4] Sun, C., Qiu, X., Xu, Y., Huang, X. *How to Fine-Tune BERT for Text Classification?* Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science, vol 11856, 2019.