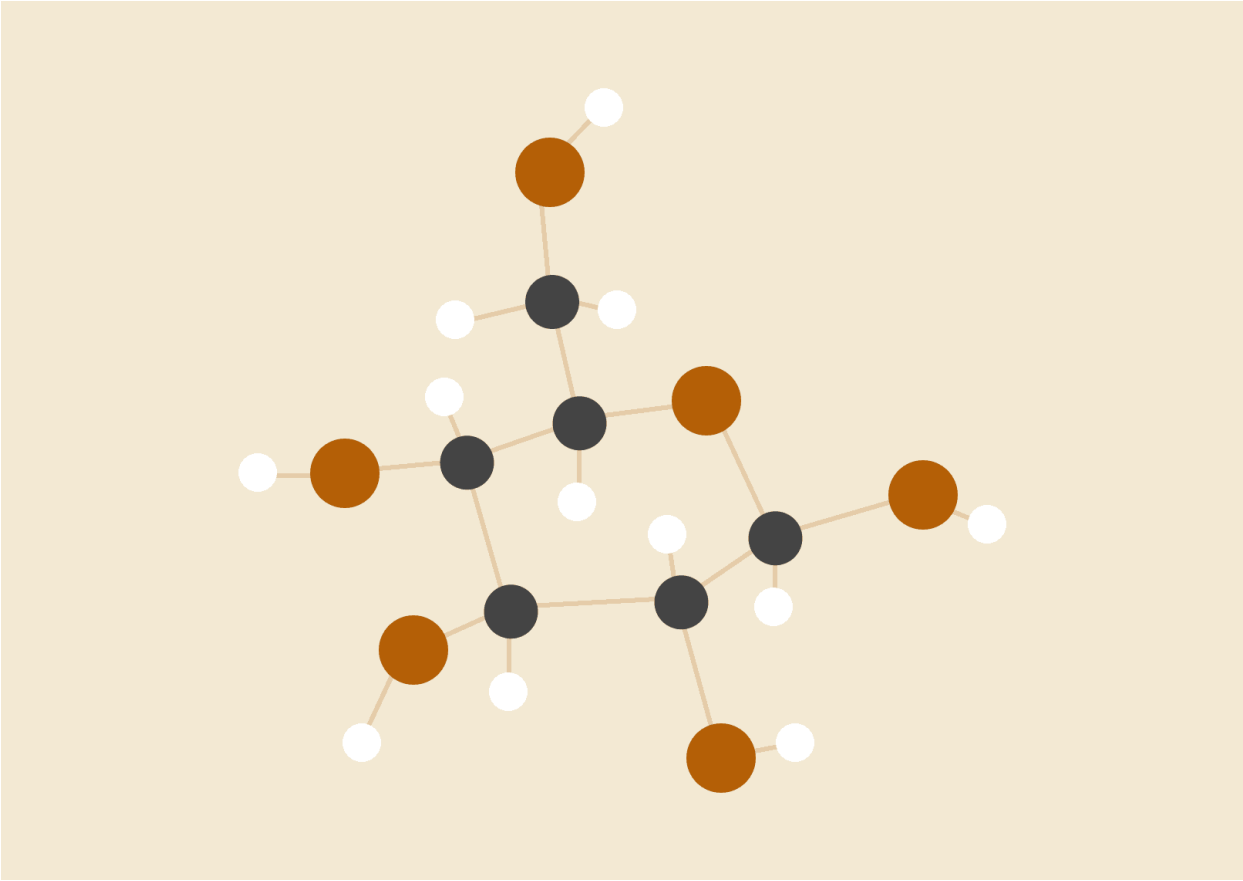
NLP REPORT



Dibyarup Pal (MT2023090)

&

B Rahul (MT2023038)

1. INTRODUCTION

Give a brief overview of the assignment's goals and the work for the NLP. Emphasise the significance of text preparation and the many approaches to model training that are employed in NLP.

2. DATA PREPROCESSING

This section details the techniques applied to transform the raw review data into a format suitable for model training. We can further break down the pre-processing into two stages: text cleaning and text transformation.

2.1. TEXT CLEANING

We performed text lowercasing, html tag and link removal, punctuation removal, etc.

2.2. TEXT TRANSFORMATION

This stage focuses on converting the raw text into a more informative representation of the model. We decontracted the phrases (like won't" to "will not), and expanded some phrases (like he'd to he would).
For the initial submissions, we used **Stemming** to reduce the word to its word stem and then passed on to the next stages before training. But for the final submission, we used the BERT classifier which doesn't require us to stem the review text explicitly.

2.3. DATA UPSAMPLING AND DOWNSAMPLING

The unbalanced data would lead to the learning of reviews with a greater number of examples (for instance, reviews with four and five stars).
We therefore downsampled the remaining reviews and upsampled the zero and one star reviews in order to lessen this phenomenon.

3. VECTORIZATION

We applied TF-IDF for feature extraction, on the preprocessed data, and then passed it onto the training models. We also tried Word2Vec but it took a huge amount of time, so we went with TF_IDF. But for final submission (using BERT) we didn't have to vectorize the review text.

4. MODEL TRAINING

We tried a lot of different models for this NLP assignment as it was a multi-class classification problem.

4.1. MODELS EXPLORED

- SVC
- SVC Balanced
- Naive Bayes
- RNN
- GRU
- BERT

4.2. MODEL EVALUATION

We various metrics to evaluate the model accuracy weighted f1-score and accuracy from the sklearn library. The final model selection was based on the accuracy obtained on submitting on Kaggle.

5. MODEL TRAINING

Following extensive experimentation with various models, BERT stood out as the optimal choice for our final submission due to its decent performance on the data.