Performance Report
Group-11
Project-T1

Task - 1

The data is vectorized using a standard Tf-Idf vectorizer, the spoiler types are labeled, vectorized, and then classified using an SVM classifier. Finally, the model is trained using a train.jsonl file.

If we check the accuracy of the classifier on the train.jsonl file, the accuracy is around 90%, and when we tested the model on validation.jsonl file, the accuracy was around 59%.

The SVM classifier with TF-IDF vectorizer is an effective method to detect clickbait spoiler types. We could have used a naive Bayes classifier, which would give similar accuracy because the data set isn't that big.

Task - 2

The data present in the training dataset and the validation dataset were stored in dictionaries.

For generating spoilers we use Question Answering models(DistilBERT) using transformers(hugging face transformers).

As we know that DistilBert is a pretrained model on a large dataset we need to fine-tune the model so that it can give accurate results.

After fine-tuning, After training the model with train.jsonl data, it evaluates the model which is used to make predictions on the validation data.
The predictions are based on the probabilities of all possible answers for each question in the validation data.

Now coming to the accuracy part, we cannot generate exact predictions of spoilers, it will generate similar types of spoilers sometimes.

While training the train.jsonl, the results are as below.
(729, {'global_step': [729],

'correct': [35],
'similar': [620],
'incorrect': [145],
'train_loss': [0.0022680554538965225],
'eval_loss': [-5.345833968298108]})

For the predictions made on validation dataset from the model trained based on training validation dataset is as follows:

{'correct': 35,
'similar': 620,
'incorrect': 145,
'eval_loss': -5.345833968298108}

The final predicted spoiler for each UUID is shown in the output dataframe.