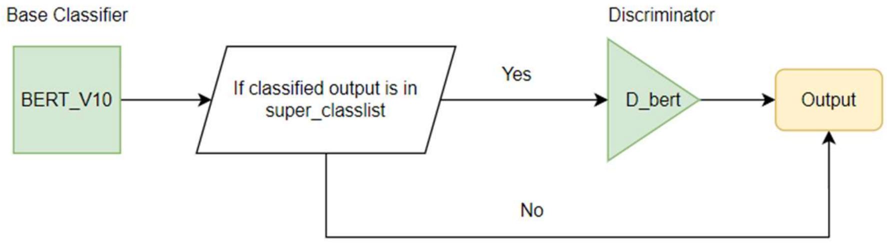
Fibe Hack the Vibe Season 2

Team – ml_dtypes

The final submission for the text classification problem consisted of the following approach:

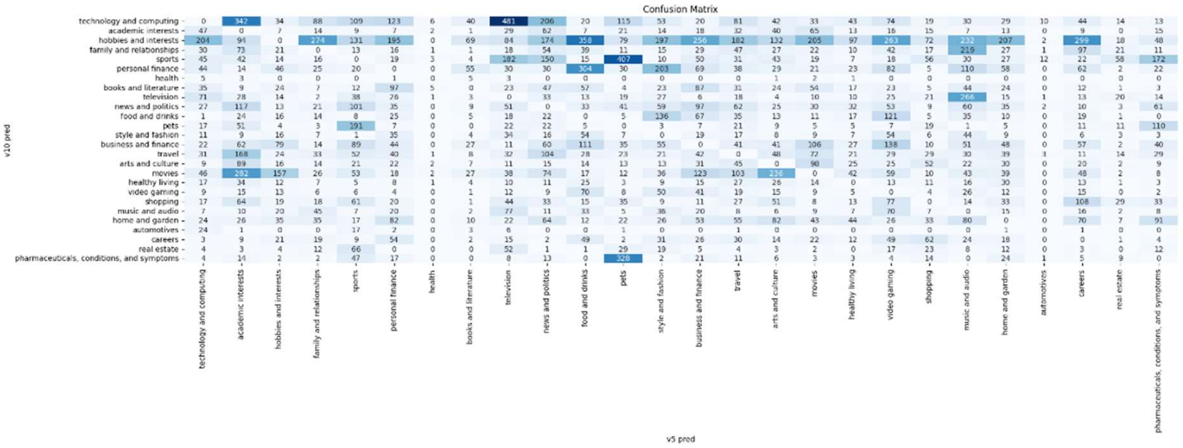
1. **Initial Pretrained Embeddings & MLP Classifier**: Started with pretrained BERT embeddings and a custom MLP classification head, achieving a 40% F1 score.
2. **KNN Voting Method**: Moved to a KNN-based voting approach using cosine similarity on vector embeddings, improving the F1 score to 65%.
3. **DistilBERT Model**: Due to resource constraints, switched to DistilBERT (an encoder-only model), reaching an F1 score of 84%.
4. **BERT Base Classifier & DistilBERT Discriminator**: Used BERT as the base classifier and DistilBERT as a discriminator to differentiate between semantically similar class pairs (e.g., business & finance, personal finance, real estate). This approach crossed the 85% F1 score mark.



5. **RoBERTa Model**: Switched to RoBERTa for successive training. The first 2 epochs were on a train subset, followed by 1.5 epochs on the full training set with a slow learning rate. This boosted performance to 86%.
6. **Text Cleaning & Augmentation**: Applied text cleaning and augmentation on the test set before inference, achieving the final F1 score of 87.64%. This approach delivered the best result for the competition.

Confusion matrix to see mismatch and between what labels are the models switching during training. This is done to actually understand which labels is the model finding hard to classify.

For BERT MODEL –



Final Model - [roberta-base-v1]