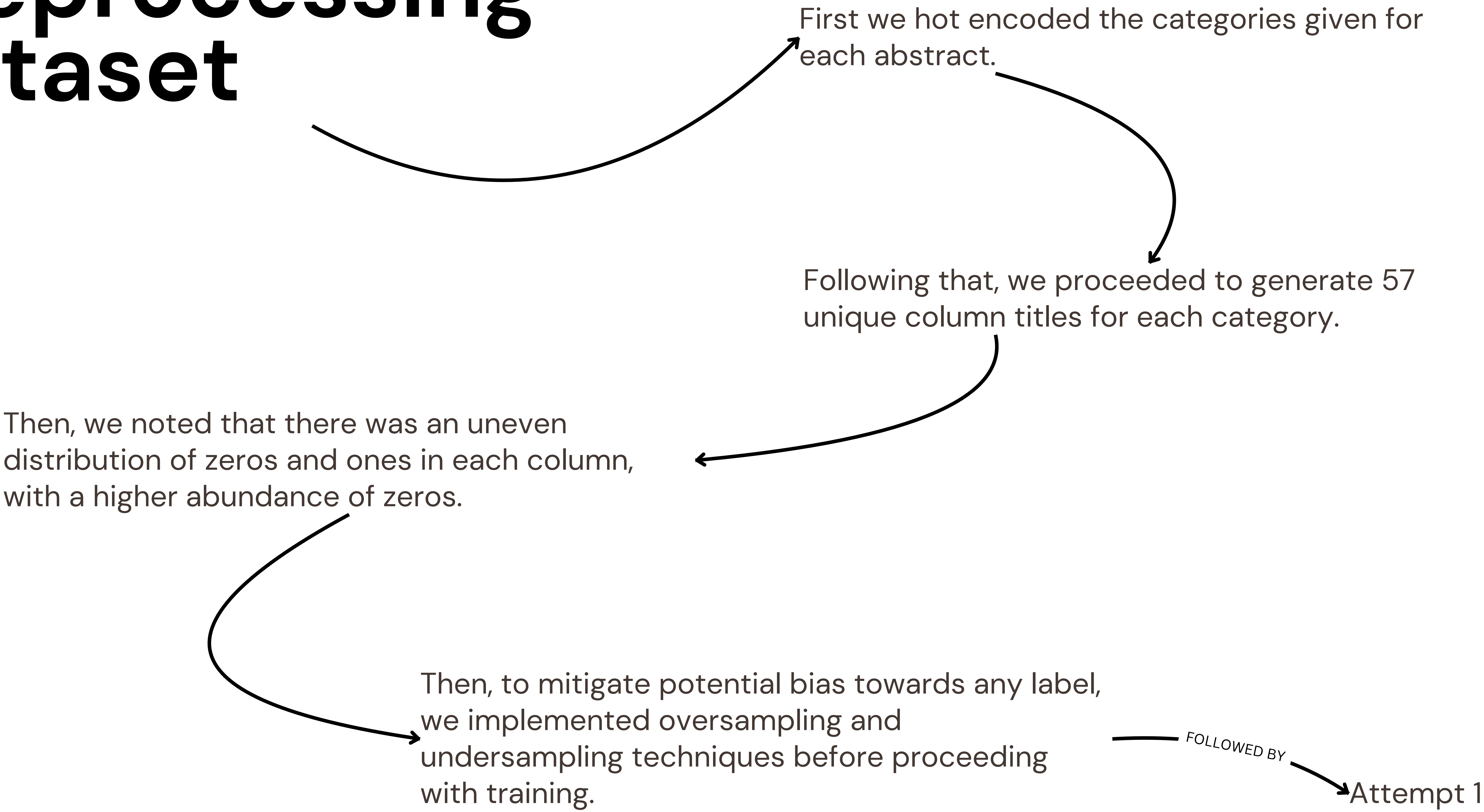
Brahmaputra
Hostel

Kriti ' 24

Automated Research Paper Categorization

Approach

Preprocessing Dataset



First we hot encoded the categories given for each abstract.

Following that, we proceeded to generate 57 unique column titles for each category.

Then, we noted that there was an uneven distribution of zeros and ones in each column, with a higher abundance of zeros.

Then, to mitigate potential bias towards any label, we implemented oversampling and undersampling techniques before proceeding with training.

FOLLOWED BY

Attempt 1

Word2Vec

Utilised Word2Vec embeddings, which are context-agnostic and utilise basic tokenizers.

Did it work???

Encountered issues with performance when using **logistic regression, random forest, SVM, and adaboost.**

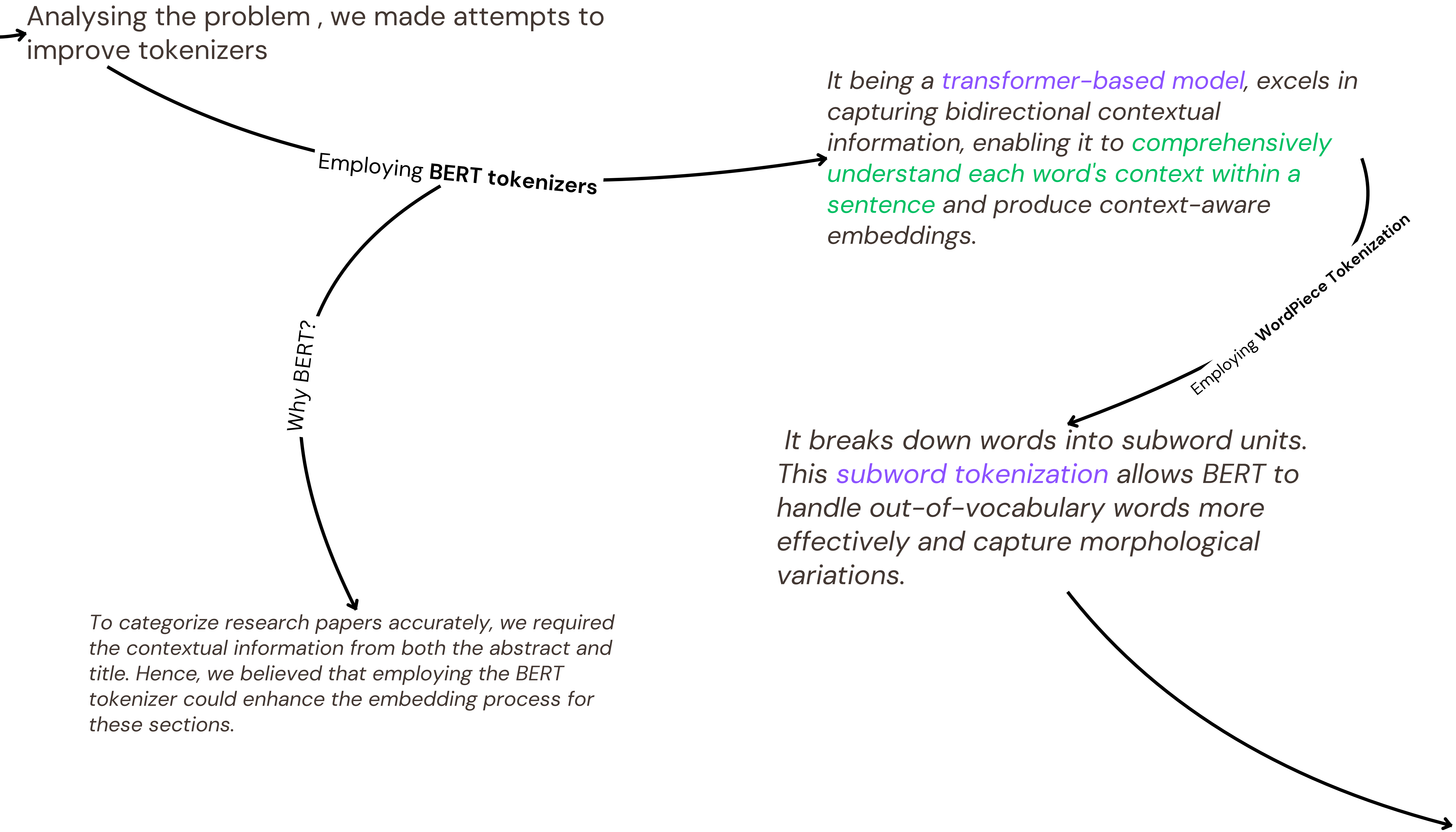
Random forest gave a very good f1 score on the train set but **performed poorly** on the test set **(score 0.01)** thus was probably overfitting the train dataset ;
Same issue was observed with **SVM**

why?

Word2Vec **uses a basic tokenizer** that separates words based on spaces and punctuation. It doesn't handle subword tokenization or more complex structures.

This might be why the word2vec embedding vector underperformed during training.

Further proceedings



Analysing the problem , we made attempts to improve tokenizers

Employing **BERT tokenizers**

It being a transformer-based model, excels in capturing bidirectional contextual information, enabling it to comprehensively understand each word's context within a sentence and produce context-aware embeddings.

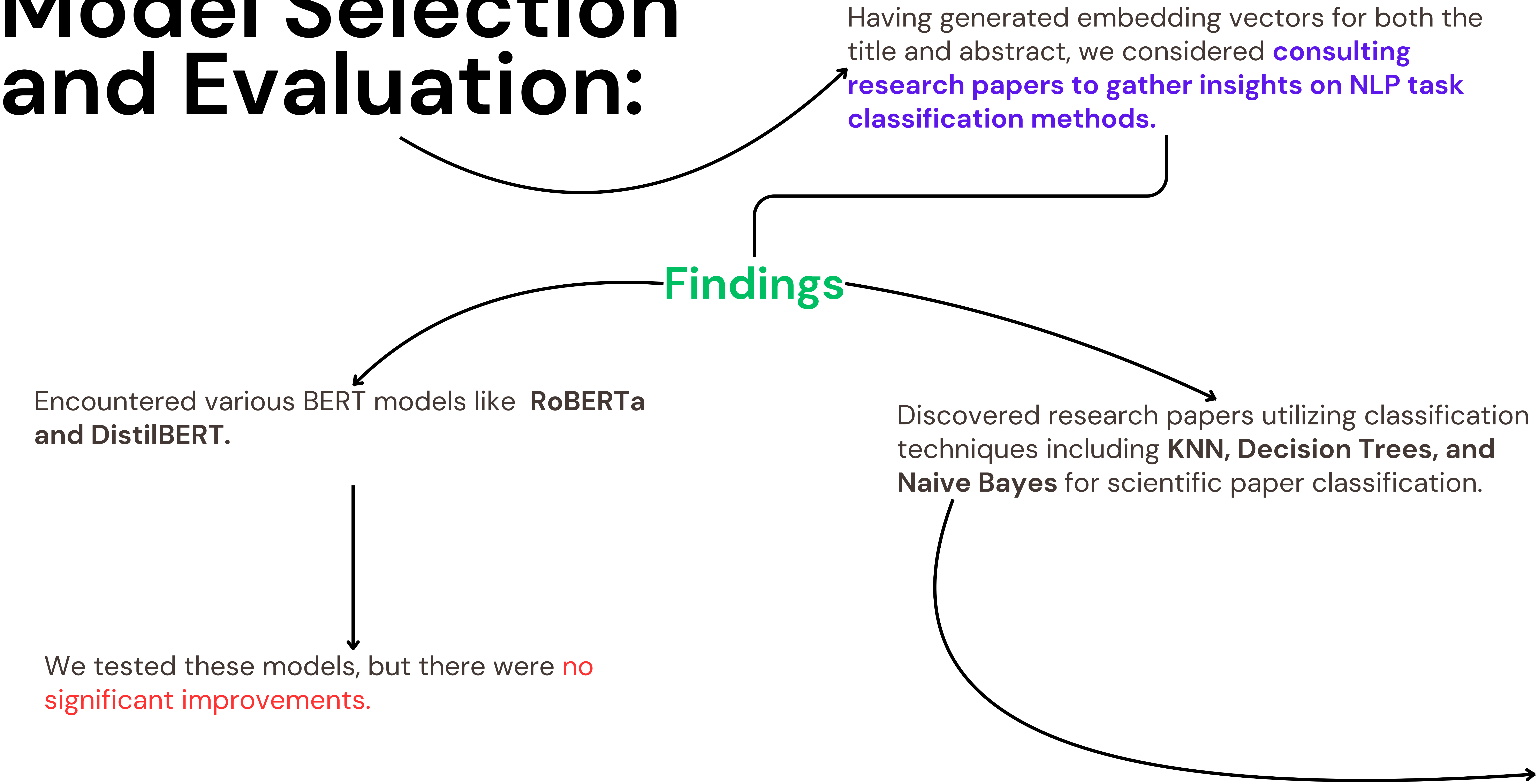
Employing **WordPiece Tokenization**

It breaks down words into subword units. This subword tokenization allows BERT to handle out-of-vocabulary words more effectively and capture morphological variations.

Why BERT?

To categorize research papers accurately, we required the contextual information from both the abstract and title. Hence, we believed that employing the BERT tokenizer could enhance the embedding process for these sections.

Model Selection and Evaluation:



Having generated embedding vectors for both the title and abstract, we considered **consulting research papers to gather insights on NLP task classification methods.**

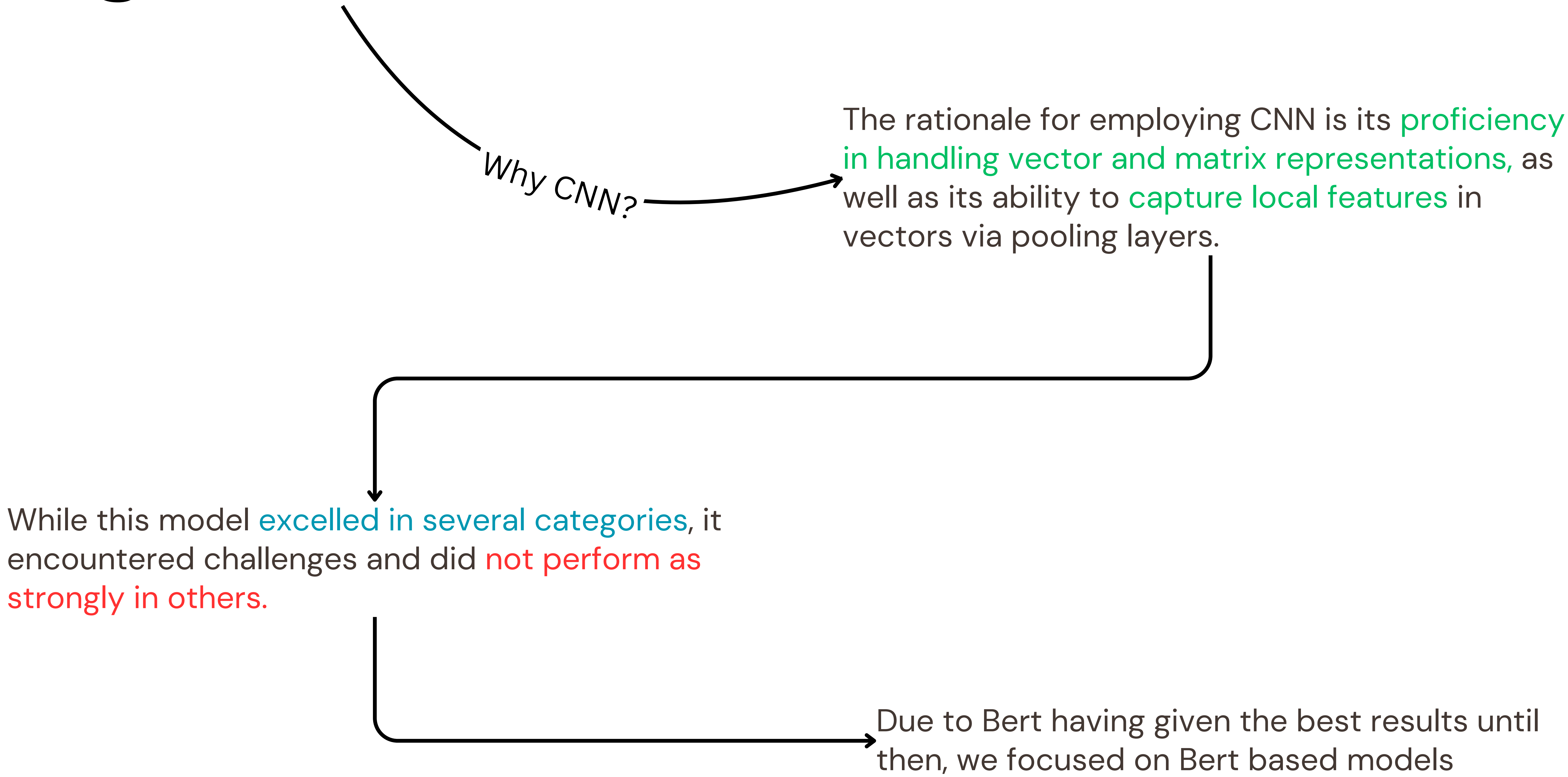
Findings

Encountered various BERT models like **RoBERTa and DistilBERT.**

We tested these models, but there were no significant improvements.

Discovered research papers utilizing classification techniques including **KNN, Decision Trees, and Naive Bayes** for scientific paper classification.

Using CNN



Why CNN?

The rationale for employing CNN is its proficiency in handling vector and matrix representations, as well as its ability to capture local features in vectors via pooling layers.

While this model excelled in several categories, it encountered challenges and did not perform as strongly in others.

Due to Bert having given the best results until then, we focused on Bert based models

SciBERT

As we directed our attention to models based on BERT and discovered **SciBERT** shortly after.

SciBERT is a variant of BERT (Bidirectional Encoder Representations from Transformers) specifically designed and pre-trained for scientific text.
It utilizes a large corpus of scientific publications to learn domain-specific language patterns and knowledge representations

A baseline model made with **SciBERT** brought us a **macro F1 score of 0.64!** With some tweaks and improvements, we **boosted that score to an impressive 0.67**, marking a significant step forward in our quest for optimum.

Challenges faced

Ensembling

Due to time constraints, the proposed strategy of ensembling SciBERT models could not be implemented. The plan was to create an ensemble of SciBERT models, a technique involving the combination of multiple instances of the same model to enhance predictive performance

THANK YOU