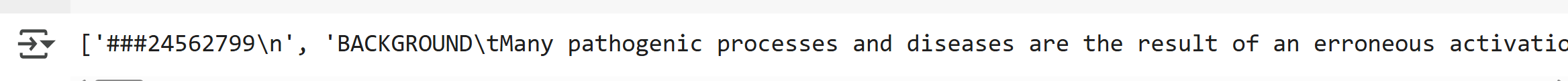
This project aims to develop a DistilBERT-based model for sequential sentence classification. After model development and training, performance will be evaluated using standard classification metrics. In a subsequent phase, dropout regularization will be introduced to assess its impact on accuracy and overall model performance.

The dataset consists of approximately 200,000 abstracts from randomized controlled trials (RCTs), containing around 2.3 million sentences. Each sentence is labeled according to its rhetorical role in the abstract, using one of five predefined categories: *background*, *objective*, *method*, *result*, or *conclusion*. Abd the numbers in the datasets are replaced by @ sign.

**The Dataset:**

After the data is uploaded in a notebook, the raw data needs to be preprocessed.

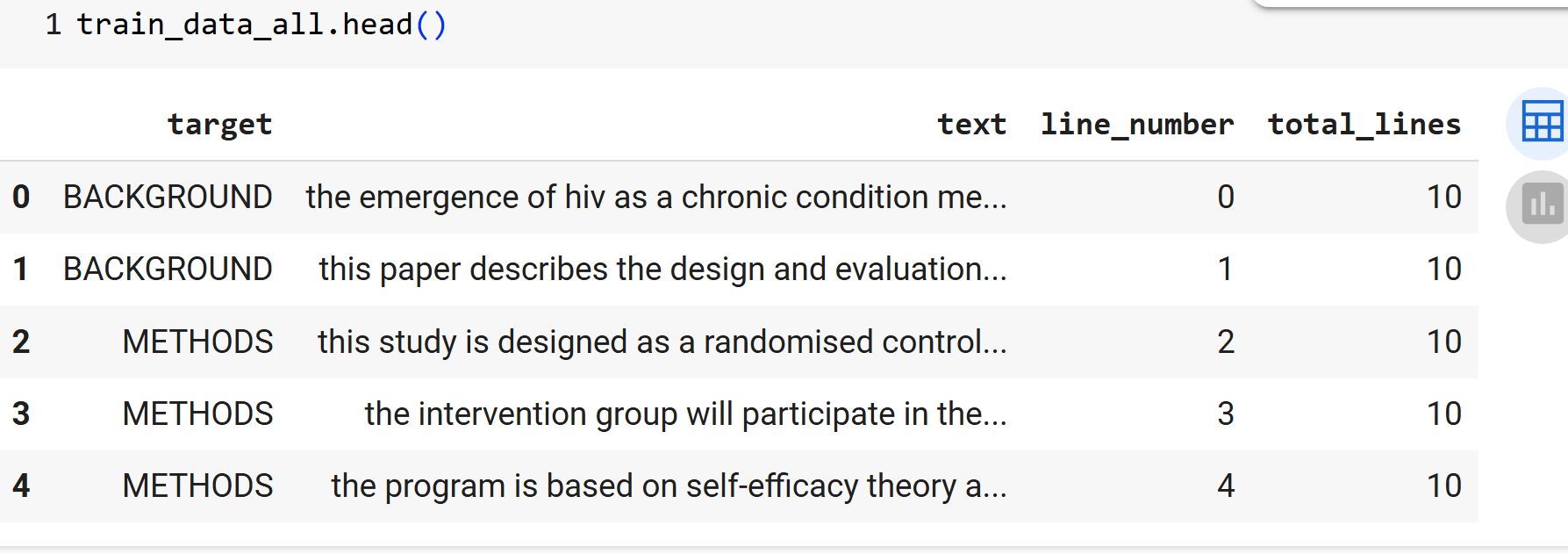


The Raw Data

The ###24562799\n in the image above appears to be an identifier (possibly a PubMed ID or reference number) with a newline (\n) at the end. The text contains tab characters (\t), indicating that fields might be separated by tabs rather than commas or spaces.

"BACKGROUND\tMany pathogenic processes and diseases..." suggests that this dataset is structured in sections, where "BACKGROUND" is a label indicating that the following text provides background information.

As part of preprocessing, remove the special characters (###, \n, \t) and extract meaningful text, label categorize different sections (



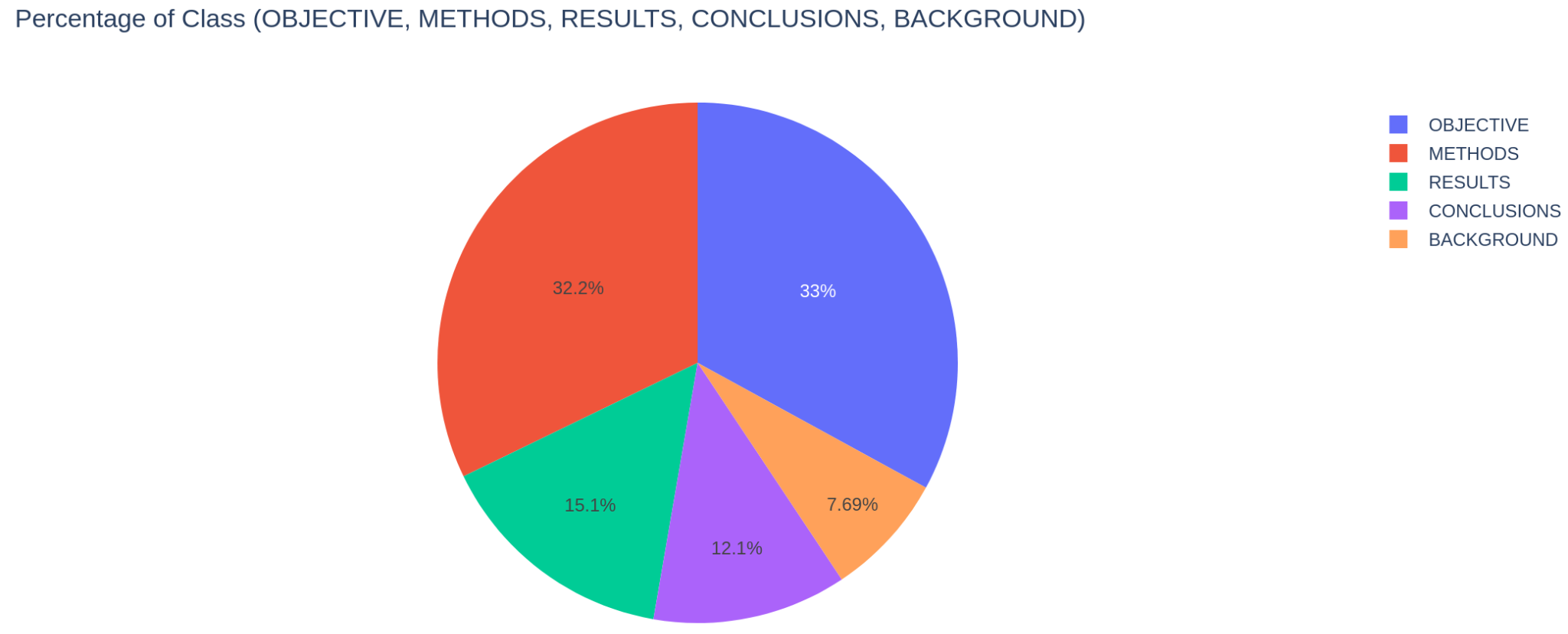
The Processed Data

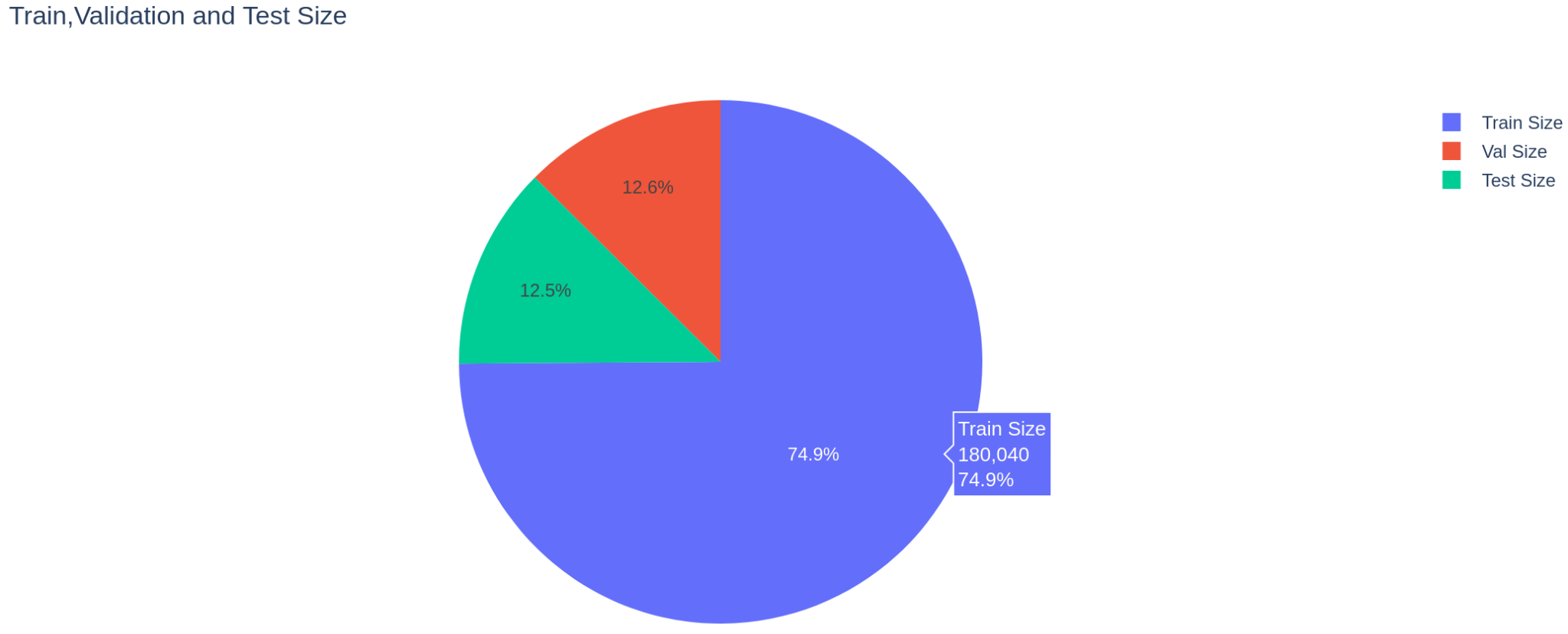
The dataset preview in the image consists of four columns:

* target: The category assigned to the sentence (e.g., BACKGROUND, METHODS).
* text: The actual sentence content.
* line\_number: The position of the sentence within the abstract.
* total\_lines: The total number of sentences in the abstract.

This structure suggests that the dataset is designed for sequential sentence classification, where sentence order plays a crucial role in classification.

**Data Visualization:**





**Data Preprocessing:**

The abstracts are split into an abstract identifier, labels, and text.

**Tokenization & Padding:**

DistilBertTokenizerFast tokenizer was used for the vocabulary tokenization with a padding of 128.. Scikit-Learn’s LabelEncoder() was used to encode the labels.

**The Base Model:**

The initial layer consists of a concatenation of two embedding layers, one for the vocabulary and another for positional embeddings. This output is then passed to a 12 head multihead attention layer whose output is run through a Normalization Layer. That output is sent through a Linear ->GELU->Linear sequence before being normalized.

**The Updated Model with Dropout:**

**Comparison & Conclusion:**

References:

1. The dataset: https://github.com/Franck-Dernoncourt/pubmed-rct
2. GitHub repository link: [**AyerDaniel/ds677: Deep Learning Project 2025**](https://github.com/AyerDaniel/ds677)

Daniel Ayer, da632@njit.edu

Fahriye Gurel, fg245@njit.edu