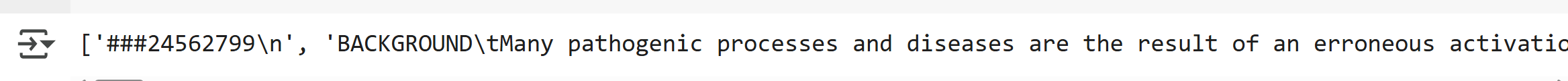
This project aims to develop a DistilBERT-based model for sequential sentence classification. After model development and training, performance will be evaluated using standard classification metrics. In a subsequent phase, dropout regularization will be introduced to assess its impact on accuracy and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

**A screenshot of a computer

AI-generated content may be incorrect.**

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.

**Data Visualization:**

This pie chart below visualizes the distribution of different sections in a dataset labeled with five scientific article components: OBJECTIVE, METHODS, RESULTS, CONCLUSIONS, and BACKGROUND. The largest proportion belongs to the OBJECTIVE class, comprising 34.6% of the data, followed closely by METHODS at 32.7%. The RESULTS section makes up 15.4%, while CONCLUSIONS and BACKGROUND constitute smaller shares of 8.89% and 8.44%, respectively. This indicates that the dataset places the most emphasis on the goals and methodological approaches of scientific texts, with less representation in summary or introductory components.

**A pie chart with different colored circles

AI-generated content may be incorrect.**

The Percentage of Classes in the Dataset

This histogram below displays the distribution of abstract word counts within a dataset. The x-axis represents the number of words per abstract, while the y-axis shows the frequency of abstracts with that word count. The distribution is right-skewed, with the majority of abstracts containing between 10 and 60 words. The peak occurs around the 25-word mark, indicating that this is the most common abstract length in the dataset. As the word count increases, the frequency drops sharply, with very few abstracts exceeding 100 words. This suggests that most abstracts in this dataset are relatively concise, possibly due to formatting or publication constraints.

A graph with a line

AI-generated content may be incorrect.

This bar chart presents the top 20 most common words found in medical or clinical research abstracts, excluding stop words (like "the", "and", "is", etc.). The top word is “results”, followed by “methods” and “were”, suggesting a structured format typical of research abstracts: Objective → Methods → Results → Conclusions. Words like “patients”, “group”, “treatment”, and “study” emphasize the clinical and experimental nature of the texts.

A graph with different colored bars

AI-generated content may be incorrect.

**Data Preprocessing:**

**Tokenization & Padding:**

**The Base Model:**

**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](mailto:fg245@njit.edu)

Jaime Corea, jdc53@njit.edu