Data Collection and Preprocessing Phase

Date	15 October 2024
Team ID	739823
Project Title	Spooky Author Identification Using Deep Learning
Maximum Marks	2 Marks

Data Quality Report Template

The Data Quality Report for spooky author identification provides an analysis of the text dataset's integrity and suitability for the project. It includes checks for missing or incomplete entries, duplicates, and inconsistent formatting. The report highlights text length variations, outliers, and class imbalances among authors. Additionally, it evaluates noise levels, such as irrelevant characters or excessive punctuation, ensuring the dataset is clean, consistent, and representative of spooky writing styles for effective model training.

Data Source	**Data Quality Issue**	**Severity**	**Resolution Plan**
Dataset	Missing authors or incomplete text data for certain authors.	High	Collect additional data from reliable sources to complete the dataset.
Dataset	Duplicated text excerpts from the same author.	Moderate	Remove duplicate entries by checking for identical text excerpts.
Dataset	Irregular formatting or inconsistent encoding in text files.	Moderate	Apply text cleaning procedures like removing special characters and correcting encoding errors.
Dataset	Class imbalance among spooky authors (e.g., some authors have significantly more data).	High	Use oversampling (e.g., SMOTE) or undersampling techniques to balance the dataset.
Dataset	Excessive punctuation or irrelevant symbols.	Moderate	Remove non-alphanumeric characters using regular expressions or custom filters.

Dataset	Inconsistent text length and tokenization issues.	Low	Standardize text length by truncating or padding text and fix tokenization.
Dataset	Lack of metadata such as genre or writing style.	Low	Manually label additional metadata or extract relevant information from the text.
Dataset	Noise in text, such as random strings or HTML tags.	Moderate	Use a denoising process to filter out irrelevant content like HTML tags or random symbols.