Data Appendix: COVID-19 Sentiment Analysis

This project analyzes public sentiment on social media during the early stages of the COVID-19 pandemic (March–April 2020). Using a dataset of tweets, we aim to compare the performance of different sentiment classification models, including lexicon-based approaches (TextBlob, VADER) and transformer-based models (BERT).

Data Source & Collection

- Dataset: Coronavirus Tweets NLP Text Classification.
- Source: Publicly available Kaggle dataset.
- Initial Scale: Corona_NLP_train.csv: 41,157 records.
    - Corona_NLP_test.csv: 3,798 records.
- Timeframe: March 2020 – April 2020.

Data Schema (Raw Data)

Column	Data Type	Description
UserName	Integer	Randomized unique ID for the user (Anonymized).
ScreenName	Integer	Randomized unique ID for the display name (Anonymized).
Location	String	Self-reported geographical location of the user.
TweetAt	Date/String	The date the tweet was posted (DD-MM-YYYY).
OriginalTweet	String	The full text of the tweet as originally posted.
Sentiment	String	Human-labeled ground truth (5-level: Extremely Negative to Extremely Positive).

Data Cleaning & Preprocessing Pipeline

To ensure the integrity of sentiment predictions, we implemented a multi-stage preprocessing pipeline using Python and Pandas:

- Data was loaded using latin-1 encoding to handle special characters common in social media text. The training and testing sets were concatenated to ensure consistent transformation across the entire corpus.
- Removed UserName and ScreenName columns as randomized IDs serve no predictive purpose for sentiment analysis.
- Rows with null values in OriginalTweet or Location were dropped to maintain a complete dataset for comparative analysis.

The following regex-based transformations were applied to the OriginalTweet column:

- Restored HTML entities (e.g., converting & to &) to their readable format.
- Stripped all hyperlinks starting with http or www to prevent the model from focusing on non-sentimental strings.
- Stripped @username tags, as specific user mentions do not contribute to general sentiment.
- Removed non-renderable Unicode characters and emojis by converting text to ASCII, ensuring compatibility with standard tokenizers.
- Removed redundant line breaks, tabs, and multiple spaces to standardize the text structure.