Yashraj Parmar

Data Glacier

26 May 2025

Week 8: Deliverables

1) **Information:**
    a) Group Name: NLP Sentinels
    b) Name: Yashraj Parmar
    c) Email: yparmar2024@gmail.com
    d) Country: United States of America
    e) College: Stevens Institute of Technology
    f) Specialization: Natural Language Processing
    g) Internship Batch: LISUM44
2) **Problem Description:**
    a) Social media platforms often use hate speech detection to get monitor posts which may harm others online through cyberbullying. Although, not everything gets caught by their hate speech detection, therefore, the objective is to build a natural language processing machine learning model which correlates a text with a hate speech level.
3) **Data Understanding:**
    a) The data that we have received is two files, one that allows for testing, and the other which allows for training. Both files share an id and text column which allows for matching texts between files through a unique identification value and the text which has the string of text. The training file, however, has the label

column which provides the true value for if there is hate speech in the text or not. From these variables, we can train the model and test it on the respective files.

4) **Types of Data for Analysis:**

a)

Feature Name	**Type**	**Data Type**	**File(s)**
id	Numerical	int	test_tweets_anuFYb8.csv & train_E6oV3IV.csv
text	Categorical	str	test_tweets_anuFYb8.csv & train_E6oV3IV.csv
label	Numerical	int	train_E6oV3IV.csv

5) **Problems in Data:**

a) There exist no NA values or outliers in the data. As for skewed distributions, there's an overwhelming amount of 0 labels than 1 labels, 29,720 to 2,242 respectively. But, there does exist "@user" strings inside some text, representing some arbitrary user on the social media platform that they are referring to. It's much better to remove this "@user" with nothing to remove any noise from the patterns that the machine learning model will recognize. We don't want the model to attribute "@user" as a pattern to certain text. Some other problems that are present are the special characters in some tweets which can be attributed to extra noise for the model and the uneven capitalization of some tweets which can become different tokens. There are also hashtags available which can be converted to just the word and the extra spaces in some tweets. One last thing

that's very minimal but can greatly improve the performance of the model is turning slang words into comprehensive English for the model.

6) **Solutions to Problems in Data:**

   a) To fix the skewed distribution of 0 labels, we can apply a class weight feature, class_weight = "balanced", this will penalize the minority class more heavily than mistakes on the majority class. To fix the problem of removing noise from unwanted strings in the text, we can start by doing text.lower() to make everything lowercase and then continue by doing re.sub(r"@[\w_]+", "", text) for "@user", re.sub(r"#", "", text) for the hashtags, re.sub(r"[^a-z0-9\s]", "", text) for the special characters and punctuation, and lastly re.sub(r"\s+", " ", text).strip() for any extra spaces. Lastly, to fix any slang which may be incomprehensible by the model, we can import the slang library to convert to text which can be tokenized.