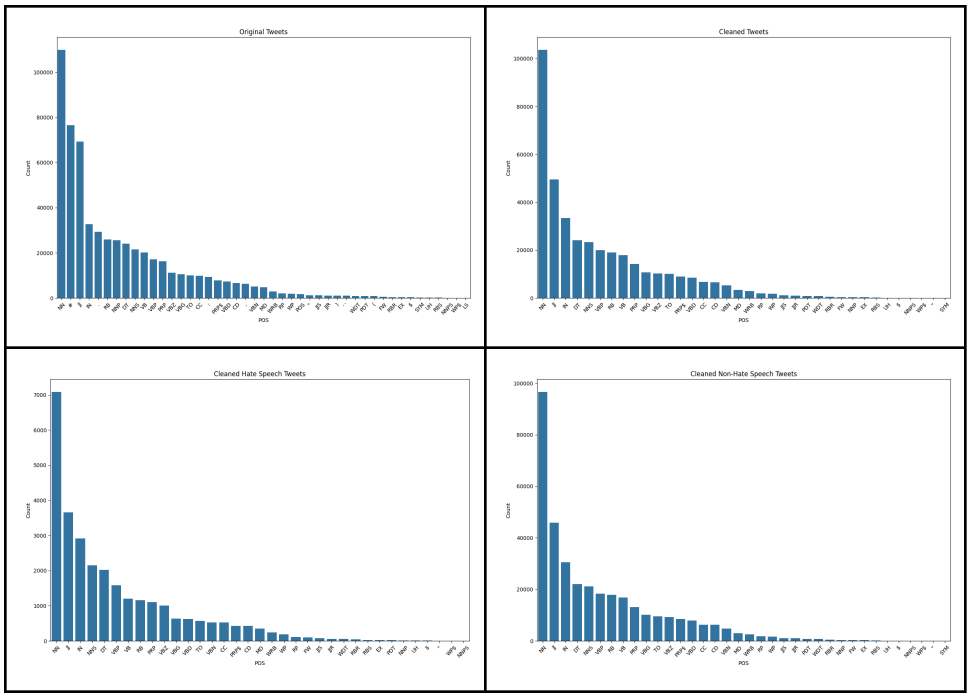
Yashraj Parmar

Data Glacier

9 June 2025

Week 10: 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 use hate speech detection systems to monitor and filter content that could lead to online harm, such as cyberbullying. Despite these efforts, such systems are not always completely effective, and harmful content can sometimes slip through. This project aims to build a machine learning model leveraging natural language processing (NLP) techniques to accurately evaluate and classify the degree of hate speech present in a given text.
3) **Exploratory Data Analysis:**
    a) We want to see the different frequencies of different part-of-speech (POS) tags for the original tweets, cleaned tweets, cleaned hate speech tweets, and cleaned non-hate speech tweets to see how cleaning the dataset changes it.



i) To go briefly go over the main POS tags we've encountered: NN (Noun, singular or mass such as dog, car, freedom, etc), # (captures hashtags such as #love, #hate, etc), JJ (Adjective such as happy, blue, angry, etc), IN (Preposition or subordinating conjunction such as in, on, because, that, etc), NNS (Noun, plural such as dogs, cars, ideas, etc), and DT (Determiner such as the, a, some, every, etc). We can see that the main POS tags seen in all of the datasets is NN which represents a singular noun which is expected since tweets often mention people, topics, or objects. Hashtags are the second highest in only the uncleaned tweets, which makes sense since after cleaning, there should be no hashtags, providing no insight. After cleaning, JJ which represents adjectives is the second highest which is also expected since after declaring a noun, the next step is usually to provide some sort of description for the noun, hateful or non-hateful. After JJ comes other tags like IN, NNS, and DT, or

prepositions, plural nouns, or determiners, respectively. These provide general support to proper sentences other than the plural nouns which just indicate hateful or non-hateful emotion towards groups of people or topics. From the POS tagging analysis, we can conclude that cleaning the tweets effectively removes any noisy tokens like hashtags, allowing for words with higher significance such as nouns, singular or plural, and adjectives to demonstrate emotion towards those nouns to represent higher importance in the sentiment analysis.

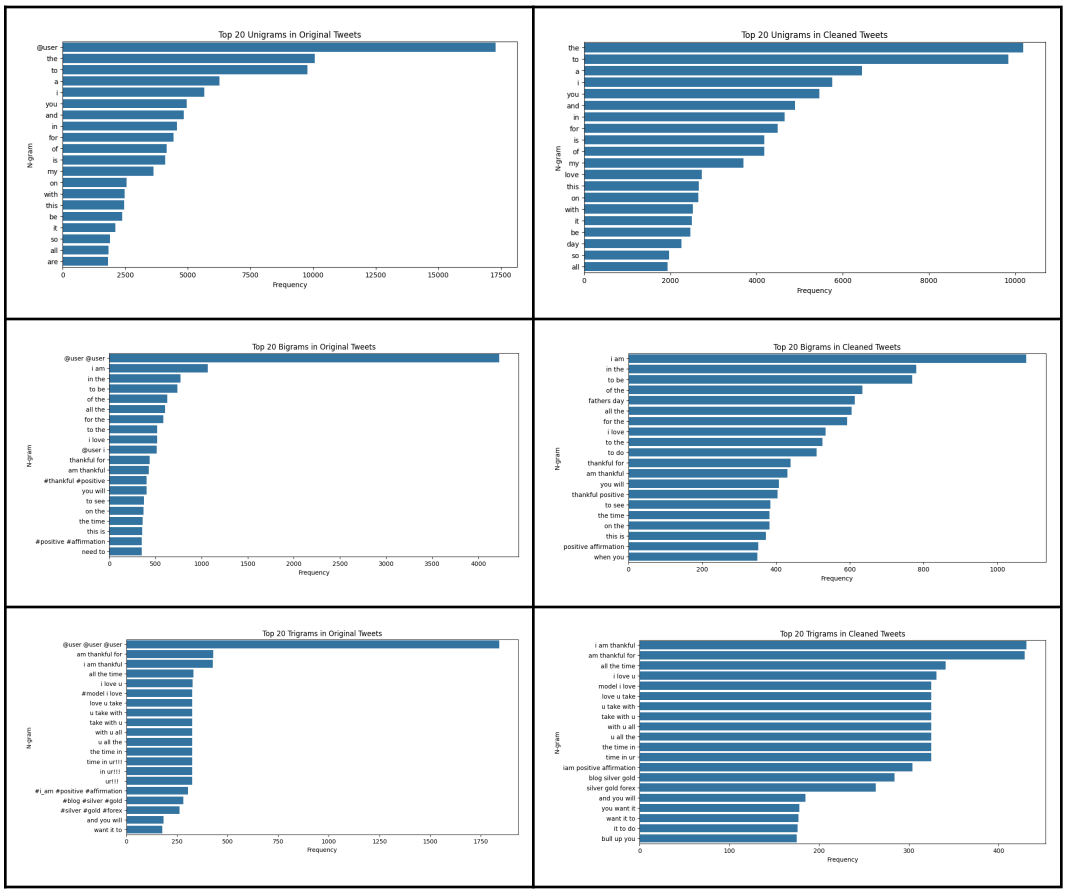
b) We want to find the average word count in a tweet before being cleaned and after being cleaned to display the average difference which will demonstrate the amount of noisy tokens in a tweet before and after being cleaned.

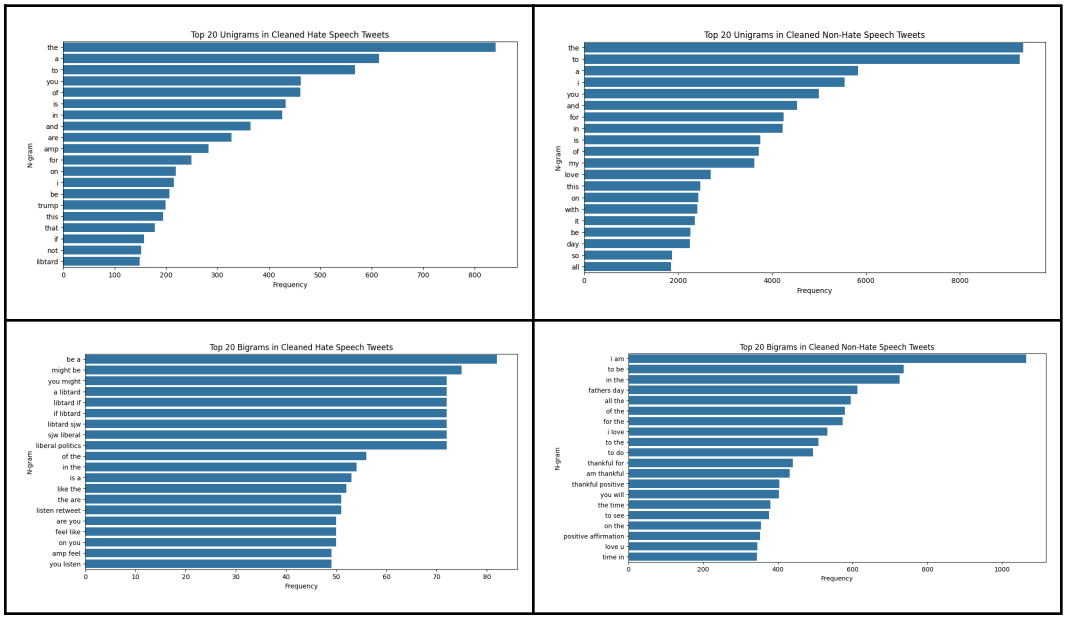
```
Average Word Count before Cleaning: 13.16
Average Word Count after Cleaning: 12.11
Average Word Count Difference before and after Cleaning: 1.04
```

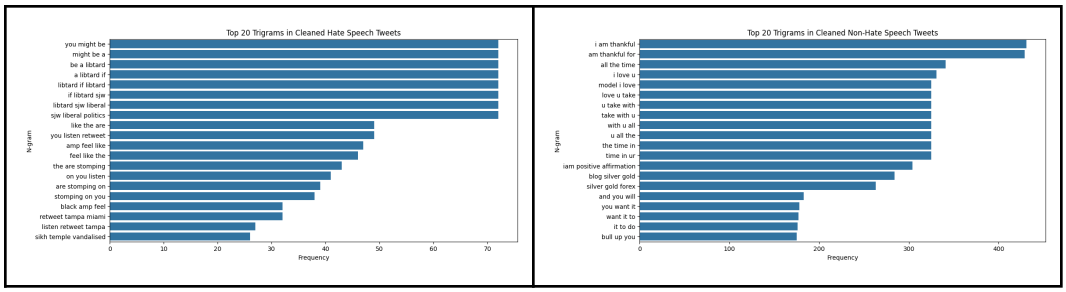
i) From the above statistics, we observe that, on average, about one word is removed from each tweet after cleaning. Based on our cleaning process, this removed word is typically a user mention (e.g., “@user”), a link, or a special character. Since we do not consider user annotations, links, or special characters important for sentiment analysis, these removals effectively eliminate noise from the tweets. Therefore, we can conclude that removing this average of one word per tweet does not negatively impact the overall sentiment analysis or its conclusions.

c) We want to check the frequency of groups of words, specifically, one word, two words, and three words for original tweets, cleaned tweets, cleaned hate speech

tweets and cleaned non-hate speech tweets. This will help us determine which forms of words or part sentences provide hate or non-hate to the tweets.



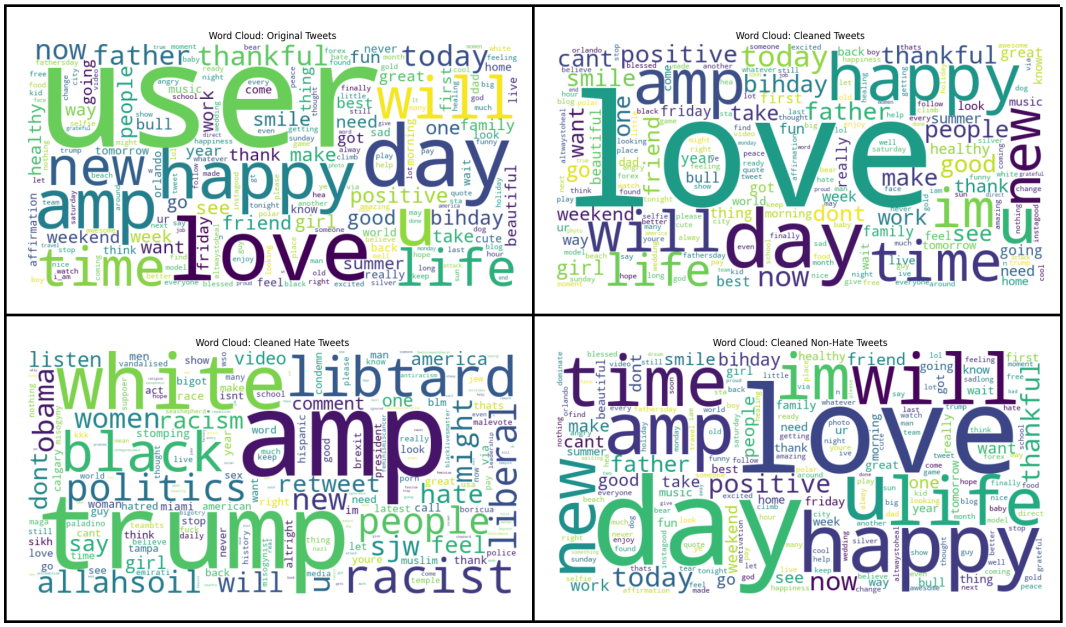




i) Splitting the visualizations up from cleaned tweets and uncleaned tweets, to then uncleaned hate speech tweets and uncleaned non-hate speech tweets, we can see the difference that cleaning provides to our dataset. For original tweets, without any cleaning or organization between hate speech and non-hate speech, it's very clear that "@user" is the most common word(s) since it's the highest in the unigrams, bigrams and trigrams. For cleaned tweets, after cleaning and no organization between hate speech and non-hate speech, the initial noise of "@user" is now gone away, and we can see that the top words or pairs of words are either parts which make up a sentence such as "the", or they are parts which make some sort of a claim such as "i am" and "i am thankful" which shows that there are sides being taken in these tweets, creating more controversy leading to more hate speech tweets or non-hate speech tweets. For cleaned tweets, with splitting up hate speech and non-hate speech, we can see that hate speech generally seems to take on pairs of words such as "be a" or a "you might be" which shows an attacking sentiment onto another user generally displayed in hate speech. Another thing to note is that the cleaned hate speech tweets seem to be more political and racist which is usually the stem of most arguments in the social media scene. For the cleaned tweets with non-hate speech tweets, we can see that this follows very closely to

the cleaned tweets without any organization between hate speech or non-hate speech. We know that this stems from the original problem in the training dataset where it's skewed towards non-hate speech tweets. We can conclude that all of these frequent words or pairs of words in each n-gram correlates directly with the already known issues in social media, usually stemming from political disagreements or racism.

d) We want to find the most frequent words being used in each original, cleaned, cleaned hate speech, and cleaned non-hate speech tweets in order to find out what words are generally considered hateful by terms of the true label provided in the training set.



i) From the original tweets, we can see that there's a lot of noise involved which is normal since the tweets aren't cleaned, some display of this noise is "user" from the "@user". The second most frequent word used in original tweets is "love" which displays the class imbalance of non-hate

speech tweets to hate-speech tweets. For the cleaned tweets, we can see that the most frequent word in this cleaned dataset is “love” which demonstrates that the cleaning process of removing tokens like “@user” works. For the cleaned hate speech tweets, we can see that clearly there are numerous words which relate to issues such as racism or politics even combining words to create offensive words. For the cleaned non-hate speech tweets, we can see that it looks very similar to the cleaned tweets, yet there is less frequency in the word “love” for the cleaned non-hate tweets. This shows that although the word “love” is used mainly in non-hate speech tweets, there is also some usage of it in hate speech tweets declaring some idea of sarcasm. We can conclude that there is a mix of words being thrown around in both the original and cleaned, as well as the hate speech and non-hate speech tweets therefore concluding the sentiment of a tweet will be harder to accurately retrieve.

4) **Final Recommendation:**

a) From the analysis of the visualizations and statistics provided above, we can declare a few statements which prove the efficiency of processes such as cleaning, tagging, dataset imbalance and final sentiment analysis. The cleaning process effectively removes noisy tokens such as user mentioned (“@user”), links, hashtags, and special characters, which do not contribute meaningful information for sentiment or hate speech detection. We know this since in numerous visualizations, we can see that the frequency of these words drops and allows for words with sentiment to be shown next. The POS tagging process and N-gram

analysis provides the idea that hate speech tweets tend to use certain attacking phrase patterns (e.g., “be a,” “you might be”) and often revolve around political and racial topics. The dataset is heavily skewed towards non-hate speech tweets, this influences frequent word distributions and model bias, this can be fixed through techniques such as balanced sampling or class weighting during model training to better learn to distinguish hate speech from non-hate speech. The context of each tweet is harder to be determined since from the word cloud analysis, we’ve seen that the word “love” is also involved in the hate speech tweets, meaning that there exists a hint of sarcasm in some tweets. Lastly, the model needs to be constantly retrained with new slang, cleaning rules, etc in order to maintain detection effectiveness over time.