Data Science Intern at Data Glacier

Project: Hate Speech Detection using Transformers (Deep Learning)

Team Member: Manhui Zhu

Email: zmanhui09@outlook.com

Country: China

College: University of Southern California

Specialization: Data Science

Table of Contents

1. Project Plan

Weeks	Date	Deliverables
Week 7	June 19, 2024	Problem Statement, Data Intake Report, Project Plan
Week 8	June 26, 2024	Data Preprocessing
Week 9	July 2, 2024	EDA (Exploratory Data Analysis)
Week 10	July 9, 2024	Feature Extraction
Week 11	July 16, 2024	Model Building and Training
Week 12	July 23, 2024	Model Performance Evaluation
Week 13	July 30, 2024	Final Submission (Slides + Report + Code)

2. Problem Statement

The term hate speech is understood as any type of verbal, written or behavioral communication that attacks or uses derogatory or discriminatory language against a person or group based on what they are, in other words, based on their religion, ethnicity, nationality, race, color, ancestry, sex or another identity factor. In this problem, we will take you through a hate speech detection model with Machine Learning and Python.

Hate Speech Detection is generally a task of sentiment classification. A model that can classify hate speech from a certain piece of text can be achieved by training it on a data that is generally used to classify sentiments. For the task of hate speech detection model, we will use the Twitter tweets to identify tweets containing Hate speech.

3. Data Intake Report

Name: Twitter Hate Speech
Report date: 06/19/2024
Internship Batch: LISUM33
Version: 1.0
Data intake by: Manhui Zhu
Data Intake reviewer: Data Glacier
Data Storage location: https://github.com/Manhui-z/Data-Glacier-Internship/tree/0083a551094656a2b96e6b2b64fd353394d34756/Week%207

Tabular data details:

Name of data	hate_speech.csv
Total number of observations	31962
Total number of features	3
Base format of the file	.csv
Size of the data	2.95 MB

Proposed Approach:

- The full dataset is consisting of 3 features: `id` with data type int64, `label` with data type int 64, and `tweet` with data type object.
- There is no missing value in the dataset.

4. Data Preprocessing

There are mainly 4 approaches to transform raw text into a structured format, making it easier for models to analyze and learn from the data. They are text cleaning, removing stop words, tokenization, and lemmatization.

4.1 Text Cleaning

The usual tweets have many causal colloquial expressions, special characters, and emojis. These messy texts make it difficult for the model to learn the underlying pattern and classify the hate speech and non-hate speech. Therefore, we need to clean the text first before we fit data into the model for training. Here are detailed steps.

- **Lowercasing:** For the same words apple and Apple, the computer will recognize them as different words. To avoid this from happening, we need to all words in lowercase.
- **Removing User Mentions:** @Users is used when we mentioned someone in our tweets. It usually doesn't have any special meanings, so we remove @Users by using `re` (regular expression) package.
- **Removing URLs:** Since the model cannot directly interpret whether the content represented by the URLs is problematic, it is not helpful for the model training, so we remove it.
- **Removing Special Characters:** For better text understanding, we remove special characters in tweets by keep only letters, digits, and whitespace.
- **Removing leading and trailing whitespace:** we remove meaningless whitespace before and after each tweet.

4.2 Removing Stop Words

Stop Words are some high-frequency common words in English language expression like 'and', 'the', 'is', but they may not contribute to the meaning and context of the sentence. We use `nltk` library to help remove the stop words. This reduces number of words the model needs to handle so that models can focus on more meaningful words, which improves efficiency and reduce noise.

4.3 Tokenization

Tokenization splits text into smaller units called token, it is usually in units of words. It is the foundation for other steps in NLP task, like stemming, lemmatization, and vectorization, which take tokens as their input.

4.4 Lemmatization

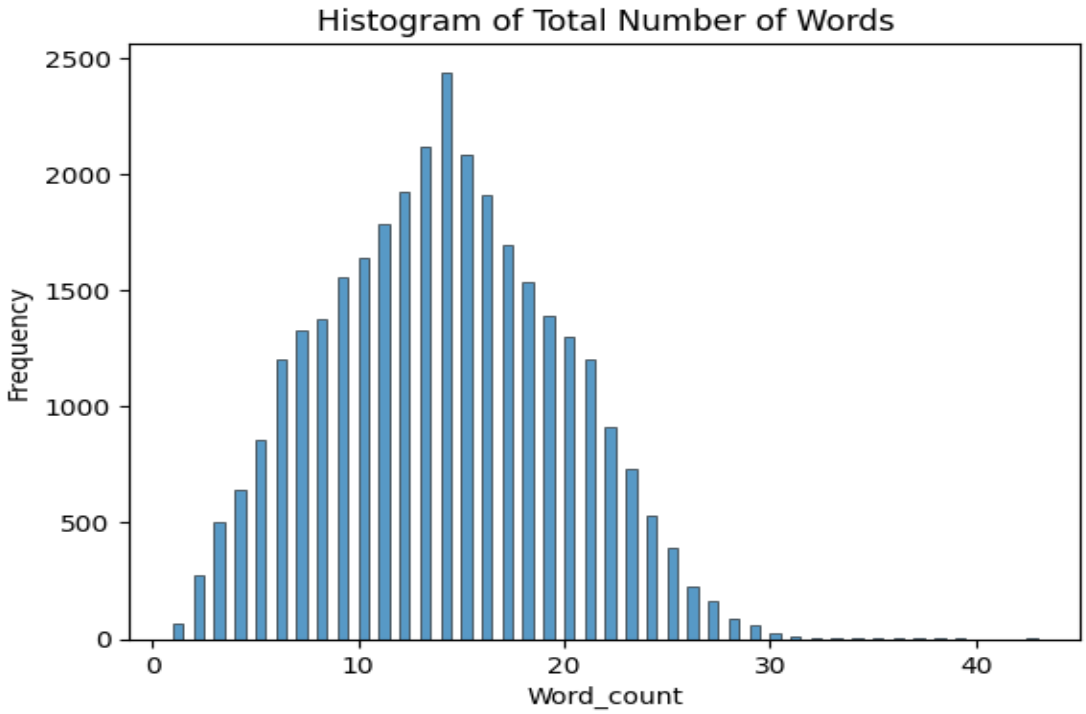
Lemmatization reduces words to their base or root form by considering the context and meaning of the word. It leads to better consistency in features, reduces redundancy, and helps in normalizing text, which is useful for classification task that requiring semantic understanding.

5. EDA (Exploratory Data Analysis)

In this part, we do the simple analysis to understand the data's structure and patterns. We check the number of total words in tweets and find the most frequents words appeared in hate and non-hate tweets. Then, we check the class distribution and use resampling technique to solve the imbalance issue.

5.1 Number of Total Words and Stop Words

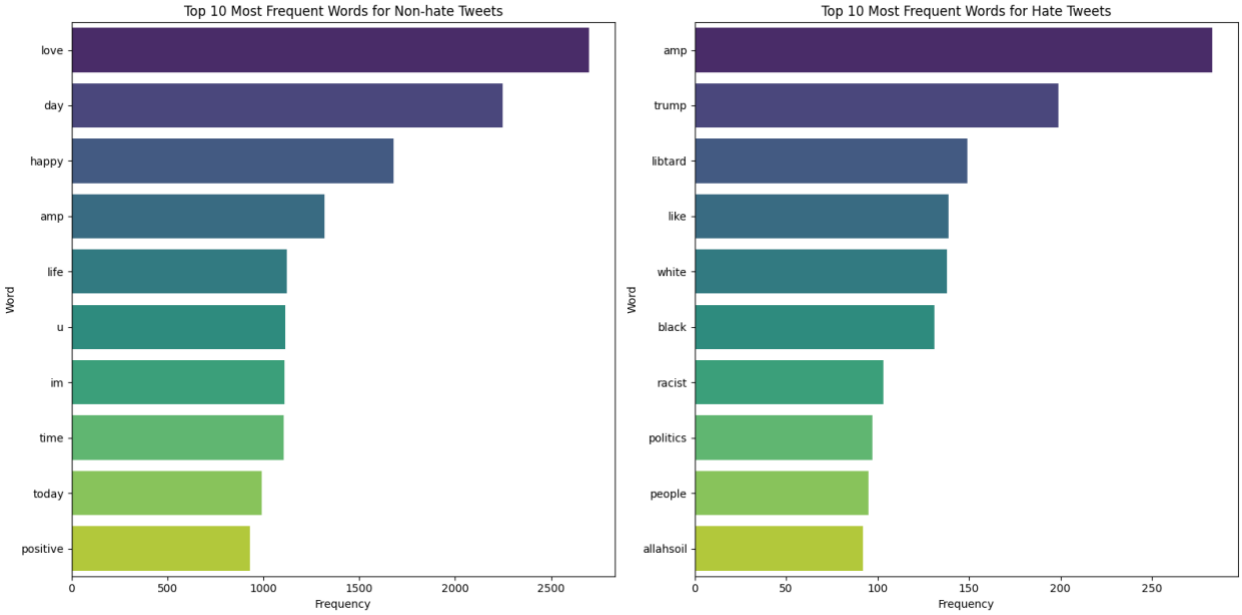
We check total number of words and stop words in each tweet. More than half of tweets have number of words in between 10 and 20.



5.2 – 5.3 The Most Frequent Words in the Whole Dataset, Hate Tweets, and Non-hate Tweets

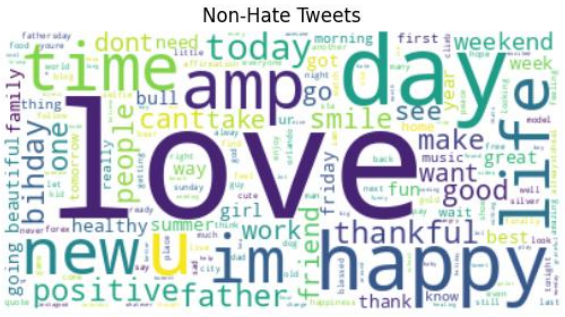
After removing the stop words, the top 10 frequent words in the whole dataset are 'love', 'day', 'happy', 'amp', 'im', 'u', 'life', 'time', 'like', 'today'. The top 10 frequent words in hate and non-hate speech are in the following graph. The word 'amp' appeared frequently in both classes. But it only means the symbol & (an ampersand), which usually appeared in HTML codes. Since

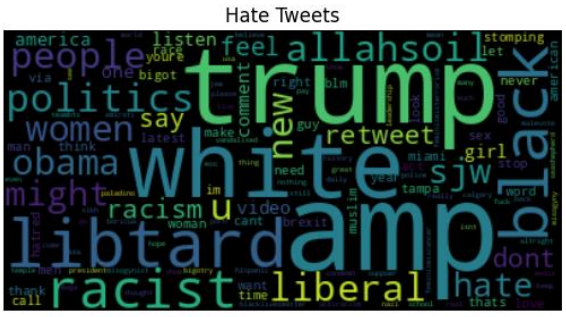
our dataset may collected by directly web scraping from Twitter, it may convert symbol '&' into string 'amp'. Therefore, this word does not have significant impact for the model's understanding of the context and meaning of tweet. We can either keep it or remove it.



5.4 Word Clouds for Each Label

To better visualize the high-frequent words in each class, we plot the word cloud for each class. The larger the size of word, the more frequently it occurs. The graph below shows that 'love', 'time', 'day', 'happy', 'life' are high frequent words in non-hate tweets. These words are either positive or at least neutral. While in hate tweets, the high-frequent words are 'trump', 'white', 'politics', 'black', and 'racist'. These words are associated with politics and race, which is usually controversial and prone to disputes.





5.5 – 5.6 Data Distribution & Solving Imbalance Issue

There are 29720 non-hate tweets and 2242 hate tweets. It indicates the imbalance in our dataset. The models trained on imbalanced data often become biased towards the major class, resulting in incorrectly classifying minority class instances. It will also cause underfit of the minority class, leading to poor model performance on minority instances. To avoid this problem, we resample the minority samples (label = 1) with replacement so that it reaches the same number of samples of majority class (label = 1). After resampling, we have 29720 samples in each class.

6. Feature Extraction

In this section, we transform raw text data into structured numerical representations that machine learning models can process. By extracting meaningful features, we can reduce the data dimension, making the model more efficient and faster to train. Here are techniques we used to do the feature extraction.

6.1 Bag of Words (BoW)

This method converts text into a matrix of token counts (numerical representations), ignoring grammar and word order but keeping multiplicity. The intuition behind this method is that similar text fields will contain similar kind of words, therefore having similar bag of words. It has three steps:

- **Tokenization**: Splitting the text into individual words and tokens
- **Vocabulary**: Creating a set of unique words (tokens) from the entire text corpus
- **Vectorization**: Converting each document into a vector where each dimension corresponds to a word from the vocabulary and the value represents the word's frequency in the document.

6.2 TF-IDF (Term Frequency – Inverse Document Frequency)

This method evaluates the importance of a word in a document relative to a collection of documents (corpus), reducing the weight of common words.

- **Term Frequency (TF):** Measures how frequently a term occurs in a document. The assumption is that terms that appear more frequently within a document are more important.

$$TF(t, d) = \frac{\text{Number of term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

- **Inverse Document Frequency (IDF):** Measures how important a term is within the entire corpus. It helps to reduce the weight of terms that occur very frequently in many documents and increase the weight of terms that occur rarely.

$$IDF(t, D) = \log\left(\frac{\text{Total Number of documents}}{\text{Number of documents containing term } t}\right)$$

- The TF-IDF score is the product of TF and IDF:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

6.3 Word Embeddings (*Word2Vec*)

This method represents words in a dense vector space, capturing semantic relationships. Words with similar meanings are located close to each other in this vector space.

Word embeddings are typically trained on large text corpora using neural networks. Here, we use the pre-trained `Word2Vec` models developed by Google (`GoogleNews-vectors-negative300.bin`), which contains vectors for a large vocabulary of around 3 million words and phrases. Each word vector is represented in a 300-dimensional space.

Twitter language is full of slang, abbreviations, and emojis. Word2Vec can learn these informal and evolving expressions, leading to accurate detection of hate tweets.

6.4 BERT (Bidirectional Encoder Representations from Transformers)

BERT is a pre-trained transformer-based model designed for a wide range of NLP tasks. It is known for its ability to understand the context of a word in a sentence, reading text in both direction (left-to-right and right-to-left). Unlike traditional models that only look at a word's surrounding context in a single direction, BERT looks at the words that come before and after it.

- **Transformer:** it applies self-attention mechanism, allowing the model to weight the importance of different words in a sentence when encoding a particular word. It enables the model to capture long-range dependencies.
- **Positional Encoding:** Since transformers do not have any built-in notion of word order, positional encodings are added to the input embeddings to provide information about the position of each word in the sequence.
- **Encoder:** it processes the input text and generates a contextualized representation. The original transformer model consists of an encoder and a decoder, BERT only use the encoder, its architecture involves stacking multiple transformer encoders, which allows it to build rich contextual representations of text. Each encoder layer consists of:
    - Multi-head self-attention mechanism: it enables the model to focus on different parts of the sentence simultaneously, capturing various aspects of the relationships between words.

- Feed-Forward Neural Networks (FFNN): it applies to each position separately and identically. FFNN layers introduces non-linearities and transforms the feature extracted by the self-attention mechanism, increasing the model's capacity to capture intricate relationships in the data.
- Add and normalize operations for layer normalization and residual connections: it stabilizes training and allows gradients to flow through the network more effectively.

If receiving a batch of string that are too large, the BERT tokenizer cannot process at once. So we first break the large dataset into small batch with size 32. The final embedding vector for each tweet has 768 dimensions, which aligns with the expectation of BERT base model,