# Data Glacier Internship

## Data Science: Natural Language Processing

# Hate Speech Detection: Transformer Based Neural Networks

Siobhan Hwang, USA
siobhan.hwang@outlook.com

Andrew O'Drain, USA
andrewodrain@outlook.com

Yawo Eklou, USA
eklouya@gmail.com

September 26, 2023

# Contents

# 1 Data Understanding

## 1.1 Type of Data

**train_tweets.csv**: Labeled, short texts (tweets) taken from Twitter.

1. `id: int64`
2. `label: int64`
   (a) Binary labels: 0 → no hate speech, 1 → hate speech present
3. `tweet: object`

**test_tweets.csv**: Unlabeled tweets

1. `id: int64`
2. `tweet: object`

## 1.2 Problems

This data does not contain any NA values, nor does it contain any traditional outliers that would skew the data. This is due to the fact that the data is primarily text and the purpose of the model is to categorize texts based on their contents. However, there are inherent phenomena that are obstacles when processing natural language — in this case tweets.

A tweet is a short text posted on the social networking platform, Twitter, now known as X, which has specific qualities of text as opposed to other forms of human-written texts. These qualities are the limited length, hashtags, emojis and emoticons, and acronyms and abbreviations. Tweets can also include images and video, but our goal is to categorize the text only.

In addition to those listed above, additional qualities that will need to be handled are:

- Spelling errors
- Punctuation
- Contractions
- Numbers
- Usernames
- Different languages
- Capitalization
- Characters in HTML or other formats

### 1.3 Approaches

Before our model is trained on the tweets, we must perform cleaning and other processing tasks. In our case, categorizing the tweets based on the presence of hate speech, which is basically a binary sentiment analysis, it is beneficial to keep emoticons and emojis, as well as words written in all capitalized characters. For this, emoticons and emojis could be replaced with a written description.

Punctuation has to be handled carefully so as not to lose meaning in the text. If we simply remove punctuation, contractions are no longer spelled correctly or will take on a different meaning. For example, "she'd" would become "shed" instead of the intended "she would." We will make sure to expand all contractions before removing any punctuation.

Hashtags are denoted by the hash symbol, #, which can simply be removed, leaving the word. This would happen during punctuation removal.

Numbers and usernames, as well as any URLs or other extraneous data, do not contribute to our task and can, therefore, be removed.

According to [**saif-etal-2014-stopwords**], stop-words, while canonically removed during preprocessing for NLP tasks, actually carry meaning in sentiment analysis tasks. The authors found that only removing a dynamic list of infrequently appearing stopwords, compiled during processing, improved the performance of their model.

Tentative tweet example: @user #tgif #ff to my #gamedev #indiedev #indiegamedev #squad! @user @user @user @user @user

Tentative tweet example after cleaning, preprocessing, and tokenization: [tgif, ff, to, my, gamedev, indiedev, indiegamedev, squad]

After tokenization, the processed tweets will need to be vectorized to allow the model to interpret and find relationships between them. To accomplish this task, the vectors need to be of the same length. The above example is only of length 6 tokens, whereas the longest tweet in the set is 84 characters and could potentially yield more tokens. Our team will decide on a set length before vectorizing the processed tweets.

# References

[1] Anderson, Luvell and Barnes, Michael. "Hate Speech." In: *The Stanford Encyclopedia of Philosophy.* Ed. Edward N. Zalta. 2022. `https://plato.stanford.edu/archives/spr2022/entries/hate-speech/`.

[2] Saif, Hassan and Fernandez, Miriam and He, Yulan and Alani, Harith. "On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* May 2014. Reykjavik, Iceland. European Language Resources Association (ELRA). `http://www.lrec-conf.org/proceedings/lrec2014/pdf/292_Paper.pdf`. pp. 810–817. Abstract: "Sentiment classification over Twitter is usually affected by the noisy nature (abbreviations, irregular forms) of tweets data. A popular procedure to reduce the noise of textual data is to remove stopwords by using pre-compiled stopword lists or more sophisticated methods for dynamic stopword identification. However, the effectiveness of removing stopwords in the context of Twitter sentiment classification has been debated in the last few years. In this paper, we investigate whether removing stopwords helps or hampers the effectiveness of Twitter sentiment classification methods. To this end, we apply six different stopword identification methods to Twitter data from six different datasets and observe how removing stopwords affects two well-known supervised sentiment classification methods. We assess the impact of removing stopwords by observing fluctuations on the level of data sparsity, the size of the classifier's feature space, and its classification performance. Our results show that using pre-compiled lists of stopwords negatively impacts the performance of Twitter sentiment classification approaches. On the other hand, the dynamic generation of stopword lists, by removing those infrequent terms appearing only once in the corpus, appears to be the optimal method to maintain a high classification performance while reducing the data sparsity and shrinking the feature space."