1 - Problem

1.1 - The problem we are trying to solve

Our problem is a classification task. We want to assess and detect the sentiment of social media posts. That is, we want to create a language model that will be able to automatically classify a post as negative or harmful. This problem can be used in various ways to address issues such as flagging harmful content or gauging customer reviews of products, or opinions depending on how it's trained and used. Sentiment analysis has been used effectively and "increased a lot of acceptance among various zone like politics, business and marketing/selling and advertisement (to estimate sales of specific products)"[3]

1.2 - Why this problem is important

With anonymity so enabled on the internet, hate speech and discrimination has become all too common on the internet. This problem has only gotten worse in recent years, and "the problem is now rampant"[1]. With how many users exist on social media, and how many posts each user has, it's nearly impossible for people to moderate these posts within a reasonable timeframe. They may catch a harmful post, but only after 5 days after it was posted. This would not be an issue if these posts stayed on the internet, however, "In some cases, toxic comments online have even resulted in real life violence"[1]. Being able to automatically classify a post as negative or harmful is extremely useful for social media moderators. If we filtered out the posts that were not classified as negative or harmful, the moderators would have significantly less posts to look at, allowing them to moderate more quickly and effectively.

1.3 - Project goals

Our first goal of this project is to create a simple web interface which allows a user to type in a tweet, and the site will classify that tweet as either positive or negative. Our next goal, if we have the time, is to enhance the already existing user interface. We will allow a user to upload a text file, which should contain a comma separated list of tweets, and output them separately into 2 different text files; One for storing the tweets classified as positive, and one for those classified as negative.

2 - Data

For this project, we will use the Sentiment140 dataset [2]. This dataset contains 1.6 million tweets. With each tweet, there is a classification of either positive, neutral, or negative. With how large this dataset is, it should be more than enough to train a language model to classify sentiment. Due to the nature of twitter a tweet's limited size "influences the use of abbreviations, irregular expressions and infrequent words."[4] This needs to be taken into account in the use, or lack thereof, stopword lists since it may be detrimental to the performance of the model if not careful.

3 - References

[1] Hanu, Laura, James Thewlis, and Sasha Haco. "How AI is learning to identify toxic online content." Scientific American 8 (2021)
www.scientificamerican.com/article/can-ai-identify-toxic-online-content/

[2] “Sentiment140 | TensorFlow Datasets.” TensorFlow
www.tensorflow.org/datasets/catalog/sentiment140

[3] Mehta, Pooja, and Sharnil Pandya. "A review on sentiment analysis methodologies, practices and applications." International Journal of Scientific and Technology Research 9.2 (2020): 601-609
https://www.researchgate.net/profile/Pooja-Mehta-26/publication/344487215_A_Review_On_Sentiment_Analysis_Methodologies_Practices_And_Applications/links/5f7bfb2992851c14bcb16528/A-Review-On-Sentiment-Analysis-Methodologies-Practices-And-Applications.pdf

[4] Saif, Hassan, et al. "On stopwords, filtering and data sparsity for sentiment analysis of twitter." (2014): 810-817
https://www.researchgate.net/publication/306364792_On_stopwords_filtering_and_data_sparsity_for_sentiment_analysis_of_twitter