



Disaster Tweets Identification

By Waya Piyapanopas (Jumbo)

Project

An NLP classification project. Using machine learning to create models and identify text-based tweets that talks about real disaster from a mixture of tweets with varying topics and fake disaster tweets.

Our priority will be on the target variable which is the disaster tweets. Additionally, we want to identify words that are strong predictors for our model.

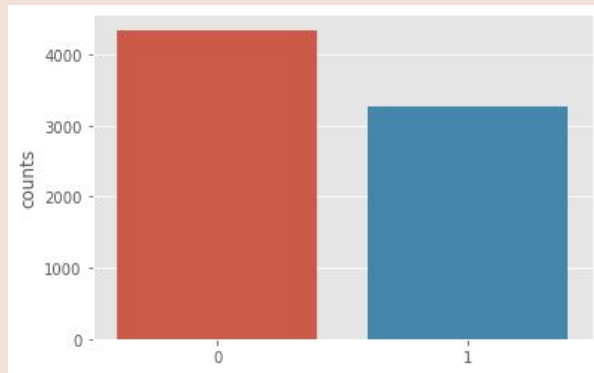
DATA

Location

The location the tweet was sent from. A lot of the tweets do not have their location identified. In cases where they are, the locations varies too greatly in each tweets.

Target

43% of the tweets belong in the target class



Text

The text of the tweet. Could include website links, emoji, and typos.

Keyword


a particular keyword from the tweet. This could be blank for some tweets.



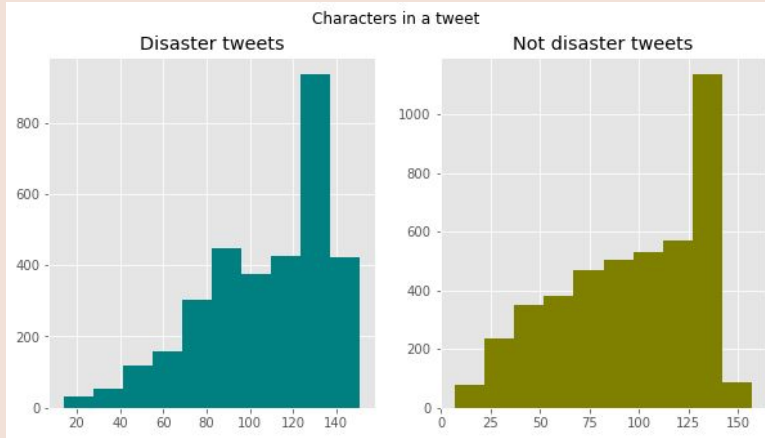
Keywords

Keywords have a huge range of impacts. Some keywords are common within both tweets which means they are not great predictors, but some are also clear indications that the tweet is talking about real disaster (or not a disaster).

For example, derailment, debris, and wreckage are keywords used only in real disaster tweets, and the word aftershock is not contained within any real tweet. Additionally, these words occur often.

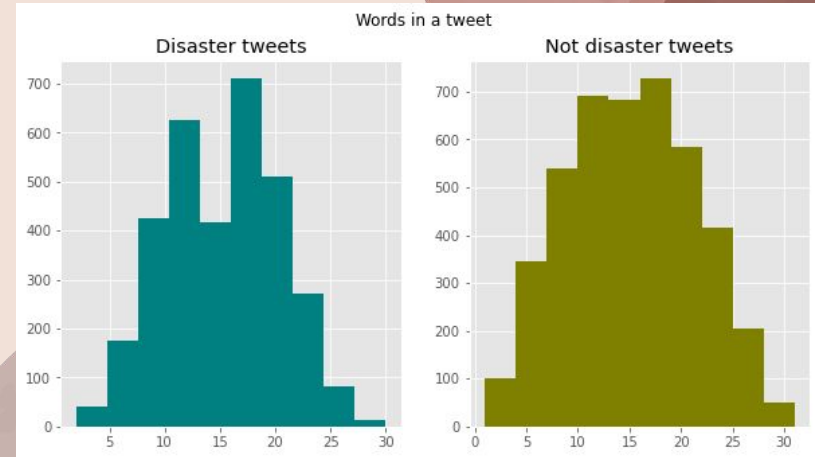


Text: Word Length



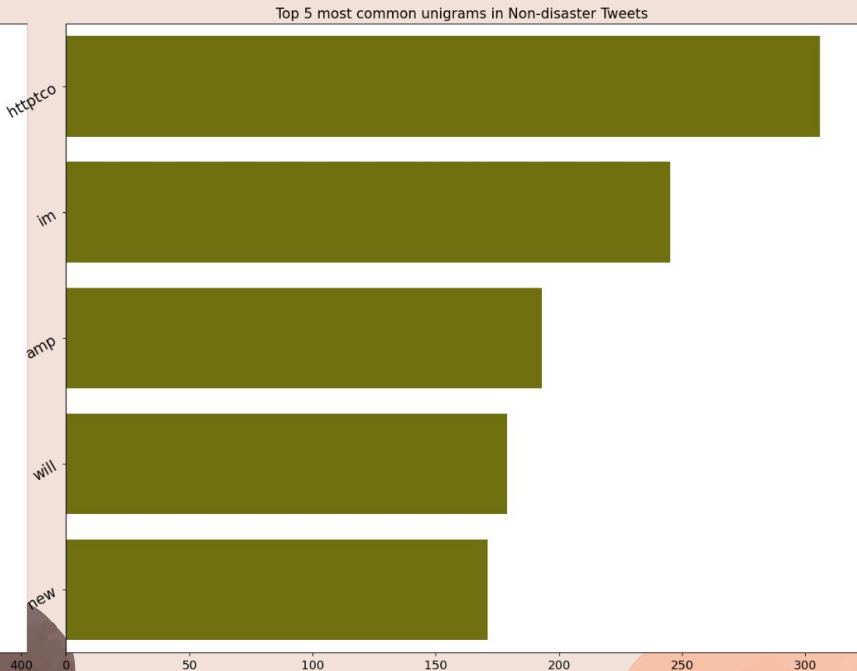
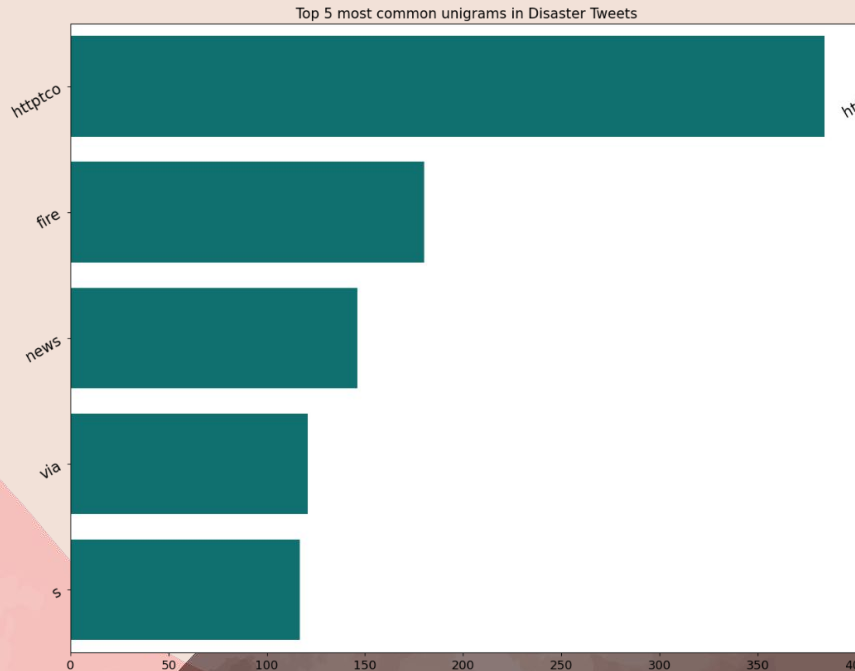
The distributions of the characters are nearly the same. Note that 120 to 140 characters in tweets are common in both. This indicate that the length of the tweet won't help to determine our outcome

The distribution of the number of words in a tweet is similar in both types of tweets. They both follow a bell curve and peak and around 13 to 18 words per tweet.



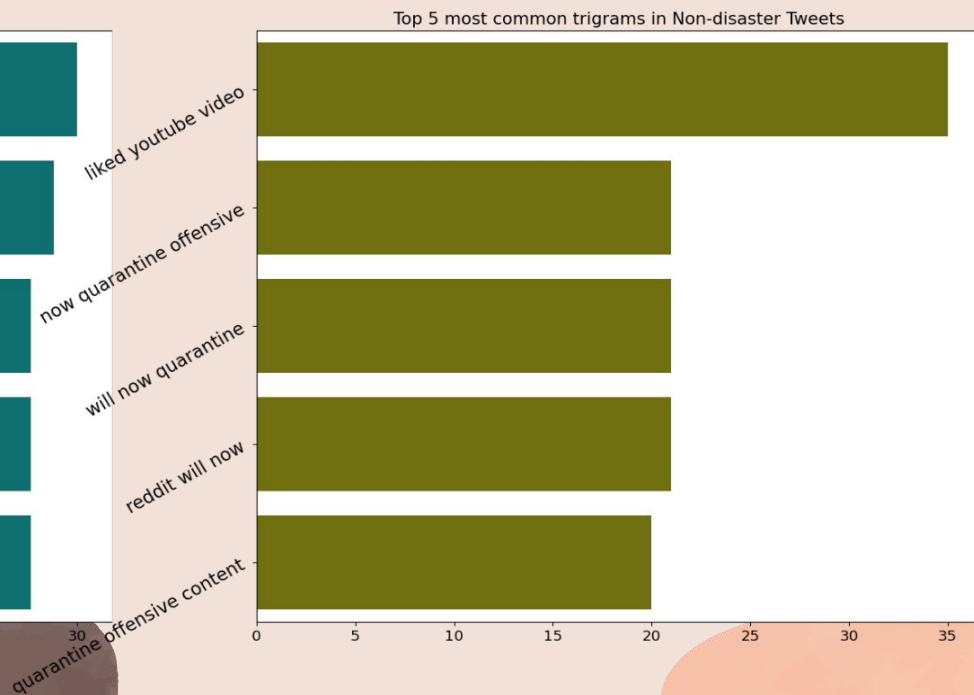
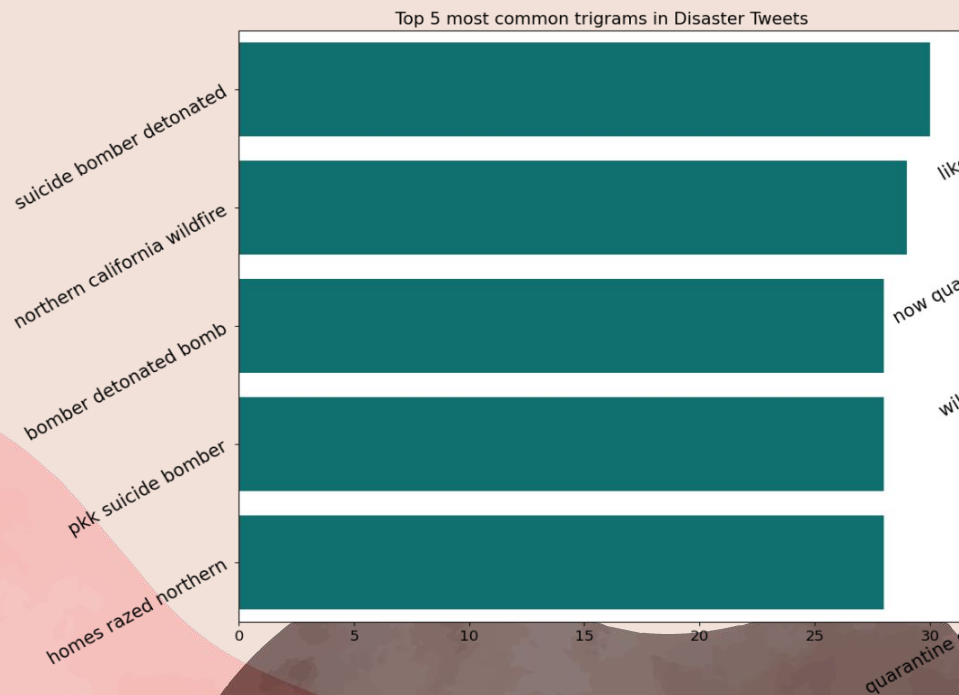
Text: Word Occurrence - Singular Words

lots of letters that are not formed into words. http, com, co, and all other indication of a website should be remove as they occur regularly in both types of tweets as well as stopwords.



Text: Word Occurrence - Pairs

Pair of words show a clearer picture of what words occurs often in non disaster tweet but not on real disaster tweets. This will help us when optimising our model



Text Preprocessing

Removing URL

[https. www. co. com.](https://www.co.com)

Removing HTML Tags

Removing Emoji

Emoticons, pictographs,
transport & map
symbols, and flags (for
iOS tweets)

Removing Punctuation

Spelling Correction

Corrections to small
typos

Feature Engineering

Word Count

Unique Word Count

URL Count

Character Count

Punctuation Count

Stop Word Count

Mention Count (@)

Mean Word Length

Hashtag Count (#)

The Best Models

Countvectorizer

Logistic Regression

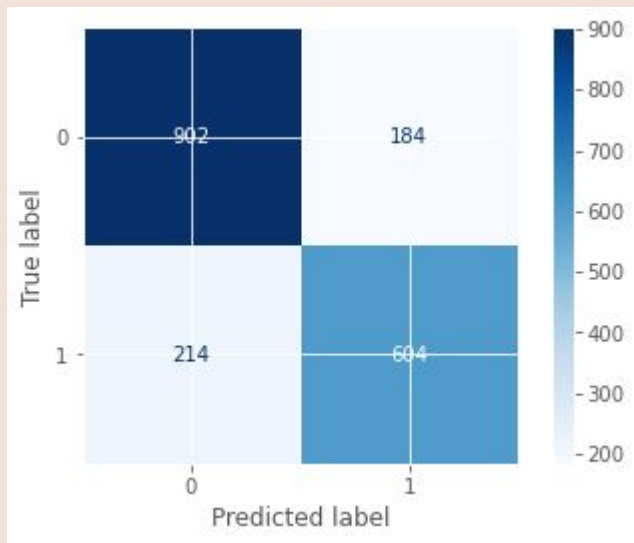
Multinomial Naive
Bayes

Bernoulli Naive
Bayes

Despite all the models performing at around 80% accuracy, each model have different strength. From our evaluation, logistic regression is able to locate more of the target variable than the other models. In this case, logistic model best help solves our problem, which is to identify disaster tweets among tweets with varying topics.

The Product model: Logistic Regression

Predictions:



Specificity: 0.8306

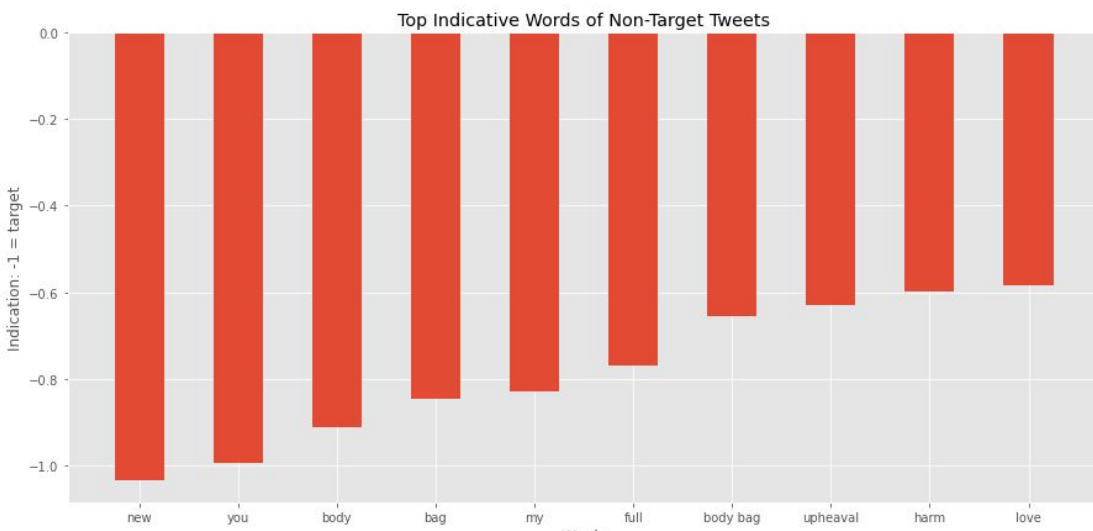
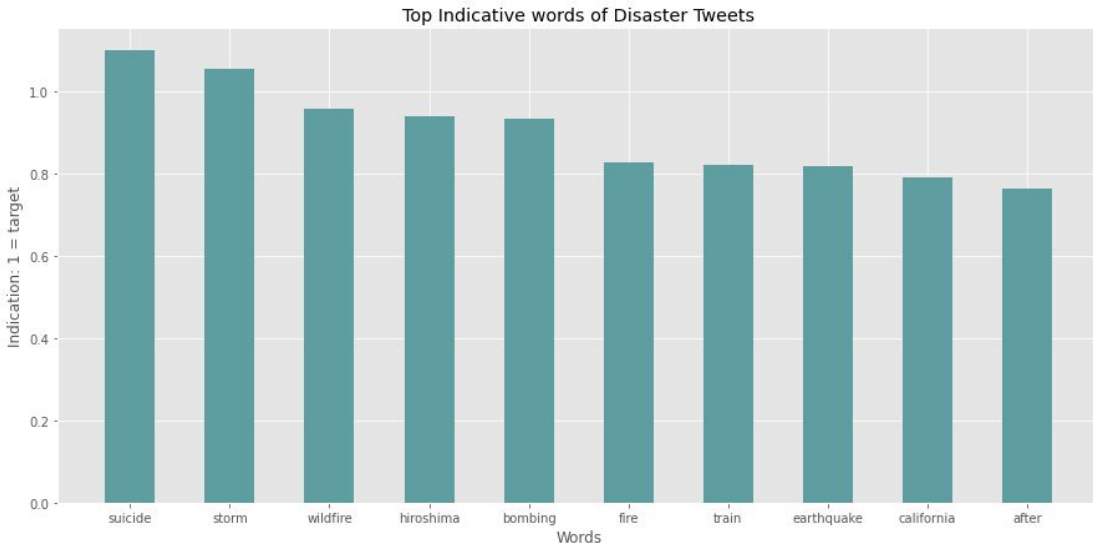
Sensitivity: 0.7384

Correctly Predict Target (Disaster) (TP): 604

Correctly Predict non-Target (TN): 902

Incorrectly Predict that the post is Target (Disaster) (FP): 184

Incorrectly Predict that the post is non-Target (FN): 214



Strong Predictive Words

Top Indicators of Disaster Tweets:
Suicide, Storm, Wildfire, Hiroshima,
Bombing, Fire, etc.

Top Indicators of Non-Target
Tweets: New, You, Body, Bag, My,
Full, etc.

Conclusion

Overall all the models have roughly 80% accuracy after the texts have been preprocessed. However, the logistic Regression model was able to correctly identify around 74% of all of the disaster tweets which is the highest of all the models.

This model associate topic words such as, Suicide, Storm, Wildfire, Hiroshima, Bombing, and Fire with disaster tweets and descriptive words like New, You, Body, Bag, My, and Full with other types of tweets.

Thank You for your Attention
Questions?

The background features abstract, organic shapes. A large peach-colored shape is in the top right. A dark brown shape is in the bottom right. A small pink shape is in the bottom center. The text is centered on the left side of the image.