# Tweeting About Disaster

Using Neural Networks and NLP to Detect Tweets About Real Crises

Flatiron School // Data Science // Project 4 // Zaid Shoorbajee

# Overview

- Business Understanding
- Data Understanding
  - Dataset
  - NLP
- Modeling
- Evaluation
- Recommendations
- Future Work

# Business Understanding

A news outlet, *The Flatiron Post*, wants to report on stories of natural disasters and other kinetic events as they happen.

- Plane crashes, earthquakes, hurricanes, wildfires, terrorist attacks, etc.
- Such events can happen without warning
- Twitter can help monitor for events, but there's too much noise for humans to parse through
- Can we use Machine Learning to predict if a tweet is about a disaster or not?

# Data Understanding: Dataset
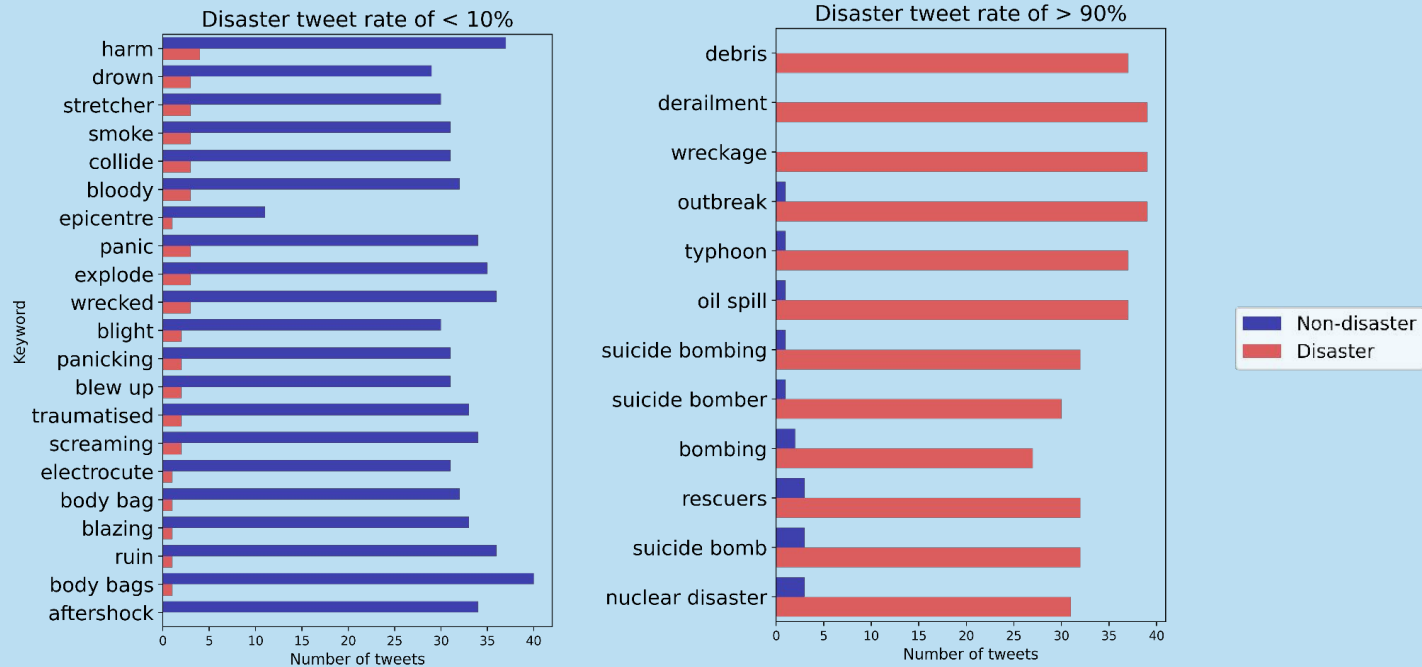
**Kaggle's Disaster Tweets dataset**

- **7,163 tweets**
- **42% disaster tweets,**
- **58% non-disaster tweets**

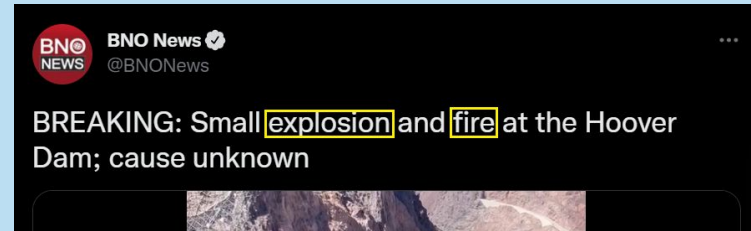| | |
|---|---|
| **Keyword** | Search term used to search for tweets |
| **Location** | User-generated location on the account |
| **Text** | Actual text of tweet |
| **Target** | Binary label<br>1: Disaster   //   0: Non-disaster |

# Data Understanding: Dataset
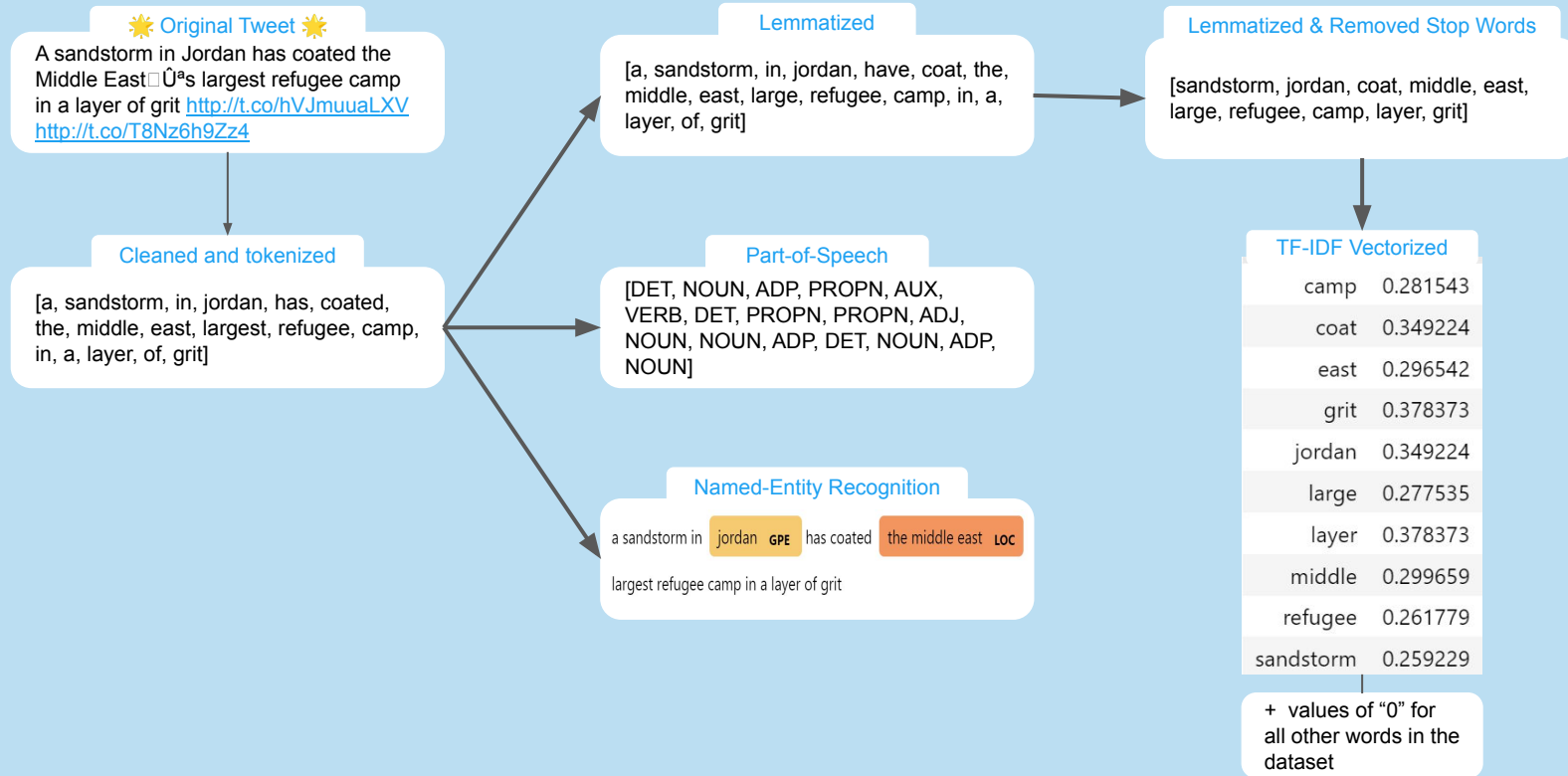


## Class Distribution by Keyword

# Data Understanding: NLP

## Natural Language Processing

- Language is messy
  - Easy for humans to understand; not for machines
- Unstructured data
- Must be heavily processed

# Data Understanding: NLP

## 🌟 Original Tweet 🌟

A sandstorm in Jordan has coated the Middle East□Ûªs largest refugee camp in a layer of grit http://t.co/hVJmuuaLXV http://t.co/T8Nz6h9Zz4

## Cleaned and tokenized

[a, sandstorm, in, jordan, has, coated, the, middle, east, largest, refugee, camp, in, a, layer, of, grit]

## Lemmatized

[a, sandstorm, in, jordan, have, coat, the, middle, east, large, refugee, camp, in, a, layer, of, grit]

## Lemmatized & Removed Stop Words

[sandstorm, jordan, coat, middle, east, large, refugee, camp, layer, grit]

## Part-of-Speech

[DET, NOUN, ADP, PROPN, AUX, VERB, DET, PROPN, PROPN, ADJ, NOUN, NOUN, ADP, DET, NOUN, ADP, NOUN]

## Named-Entity Recognition

a sandstorm in  jordan  GPE  has coated  the middle east  LOC

largest refugee camp in a layer of grit

## TF-IDF Vectorized

| | |
|---|---|
| camp | 0.281543 |
| coat | 0.349224 |
| east | 0.296542 |
| grit | 0.378373 |
| jordan | 0.349224 |
| large | 0.277535 |
| layer | 0.378373 |
| middle | 0.299659 |
| refugee | 0.261779 |
| sandstorm | 0.259229 |

+ values of "0" for all other words in the dataset

# Data Understanding: NLP

## Effects of Lemmatization & Removing Stop words

This version has fewer words in common between disaster and non-disaster tweets →



Top 20 tokens

Top 20 lemmas (no stop words)

# Data Understanding: NLP

**Meta-Features**

Aside from linguistic features, we can extract seemingly arbitrary information about each tweet

- Character count
- Number of stop words
- Proportion of stop words
- Character count of non-stop words divided by total character count
- Number of tokens
- Average length of lemmas
- Number of unique lemmas
- Proportion of words that are hashtags (#)
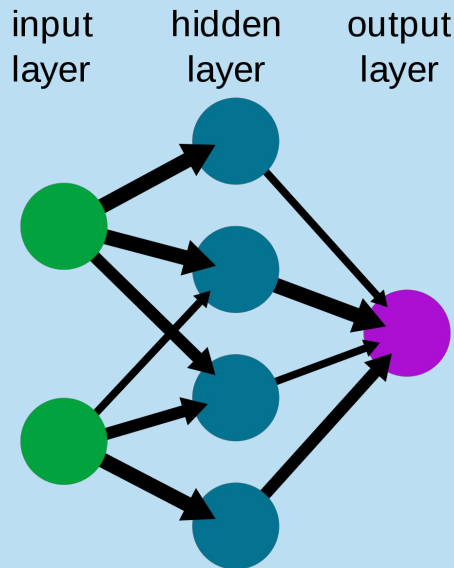- Proportion of words that are mentions (@)
- Has URL

**Examples:**
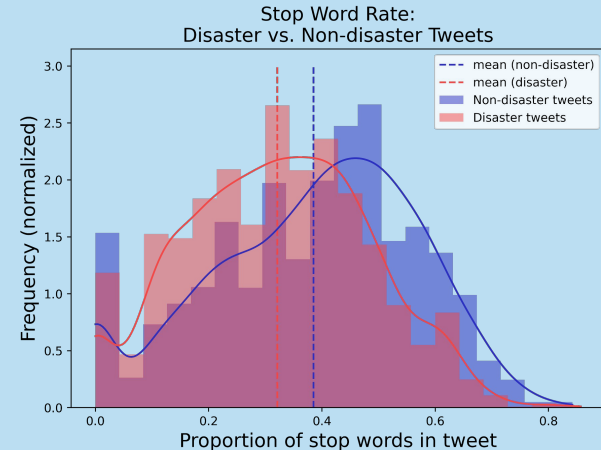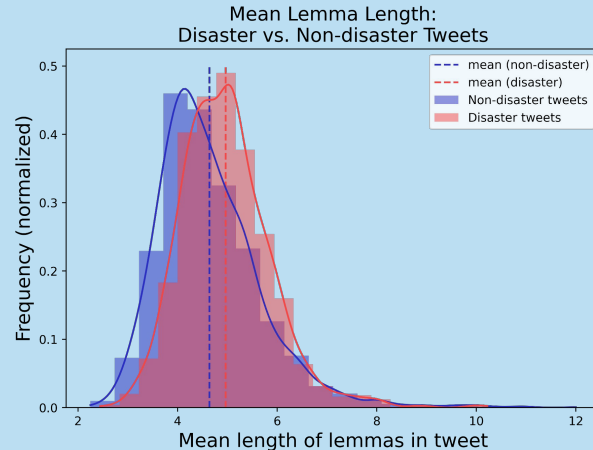
# Modeling

## Neural Networks:

- Machine learning practice mimicking the neurons of a brain. They can be tuned:
  - Inputs, layers, nodes
  - Simple baseline model; five additional models
- Task: Binary classification
- Training, validation, and test datasets

input
layer

hidden
layer

output
layer

# Evaluation

*The Flatiron Post* is concerned with false negatives, i.e. neglected disaster tweets.

| Accuracy | How many tweets does it get right? | 78% |
|----------|-----------------------------------|-----|
| Recall | How many **disaster** tweets does it get right? | 72% |



Final Test Results
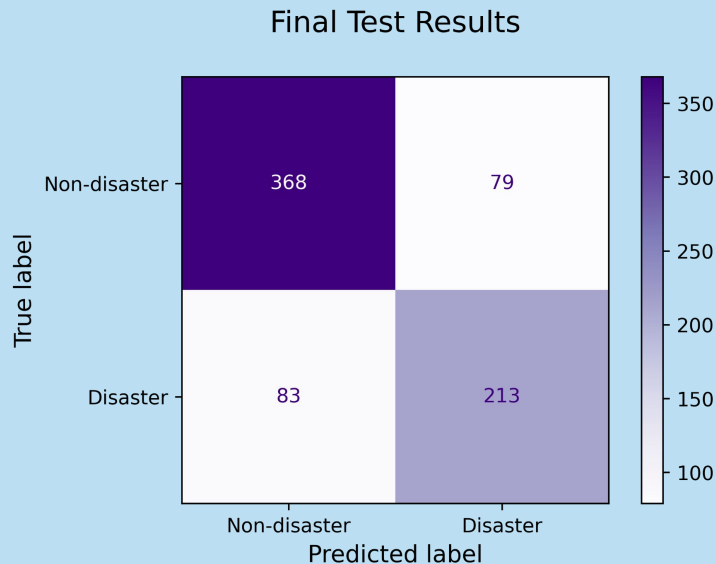
# Recommendations

- False negatives are still prominent, so don't completely filter out non-disaster predictions. Instead, run all tweets in a feed that also displays the model's predicted probability.

- Discard search terms that yield very few disaster tweets, like "harm," "bloody," "screaming," "ruin," etc.

- Narrow the criteria for what constitutes a "disaster." Focus on kinetic events and more immediate crises (bombings, earthquakes, crashes, etc.).

  - Requires relabeling or gathering new data

Example Feed with Probabilities

| Tweet | Probability |
|---|---|
| Watch This Airport Get Swallowed Up By A Sandstorm In Under A Minute http://t.co/H84R1TIh8J | 97.85% |
| 1.9 earthquake occurred 15km E of Anchorage Alaska at 00:11 UTC! #earthquake #Anchorage http://t.co/QFyy5aZIFx | 96.99% |
| Two giant cranes holding a bridge collapse into nearby homes http://t.co/9asc1hhFNJ | 94.70% |
| Slash-and-burn blamed for bush fires in western St Thomas - http://t.co/5dJ6cHjFZP | 76.26% |
| Truck crash on 40w at US70 in Lebanon is a fatality. Very sad. Expect long delays through the morning. | 68.47% |
| British bake off was great pretty hilarious moments #mudslide | 40.98% |
| Super loud thunder woke me up from my very nice nap | 27.90% |
| I PUT MY CHICKEN NUGGETS IN THE MICROWAVE FOR 5 MINUTES INTEAD OF 1 ON ACCIDENT AND THEY FUCKING BURNED | 15.30% |
| On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE http://t.co/qqsmshaJ3N | 14.90% |
| *screams internally* | 11.47% |
| I rate Hazard very highly but his fanboys are among the worst accounts on Twitter. | 4.38% |

# Future Work

- Training this model was limited by the search terms used to obtain these tweets. Training on a less biased sample might yield better results.

- Missing feature: Does the tweet contain a photo or video?

- This model is just one piece of a larger application. Other pieces:

  - Tool that automatically requests tweets through Twitter API
  - User-friendly feed for journalists

# Thank you

zshoorbajee@gmail.com

@zshoorbajee

linkedin.com/in/zshoorbajee/

zaid.fyi