Detecting Disaster Tweets

By Jeff Spagnola

Introduction

- Accurately determining the context of a tweet is useful in many areas.
- Social media outlets increasingly becoming a platform for breaking news.
- Being able to identify "disaster tweets" can lead to better rescue response and more widespread awareness.

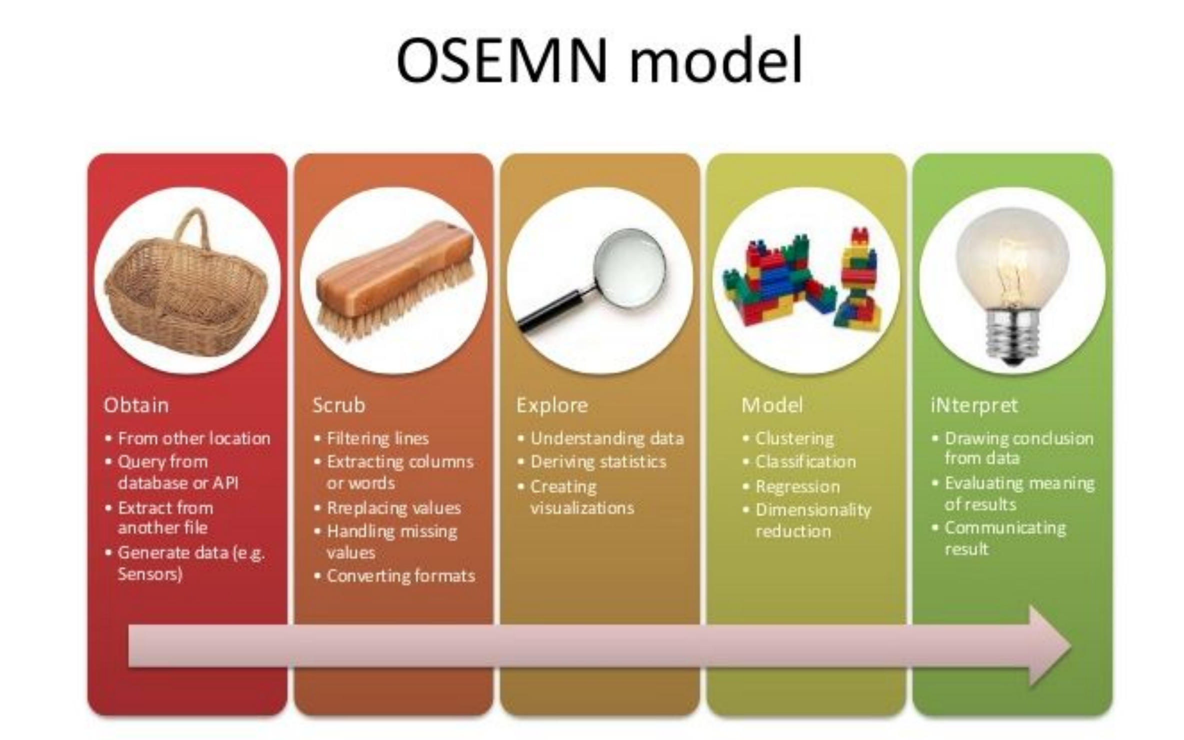
Introduction

- Create a model to accurately identify disaster-related tweets.
- Using natural language processing, supervised machine learning, deep learning.
- Difference between “The plane crashed.” and “We crashed the party.”

1. The Process

Process

Throughout this presentation, we will be following the OSEMN Data Science Process.

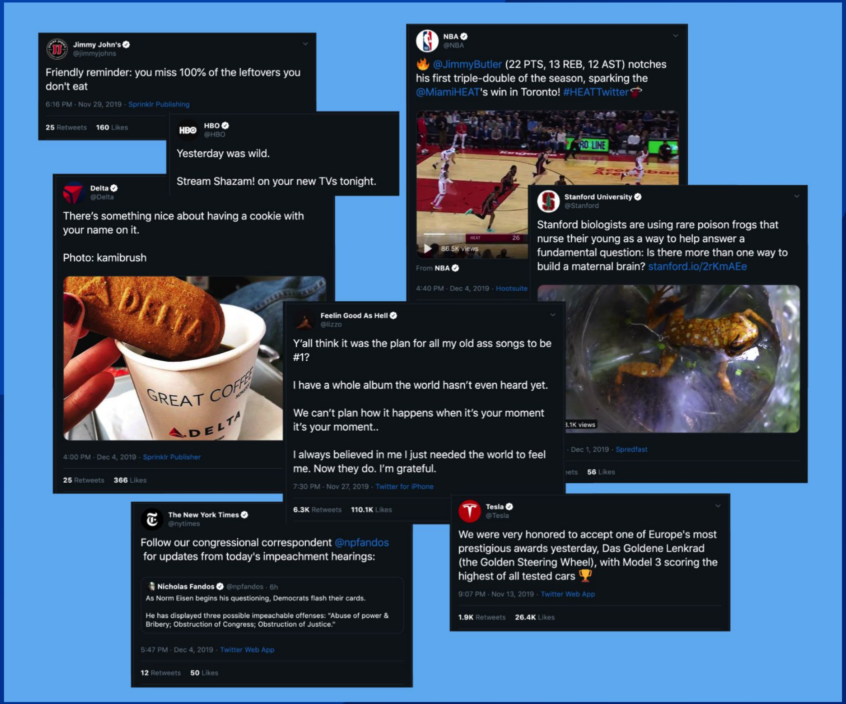


2. The Data

Data

- The data was obtained from the Disaster Tweets Dataset from Kaggle.
- Contains over 11,000 tweets
- Collected Jan. 14th, 2020
- Topics include Taal Volcano, Coronavirus, Flight PS752

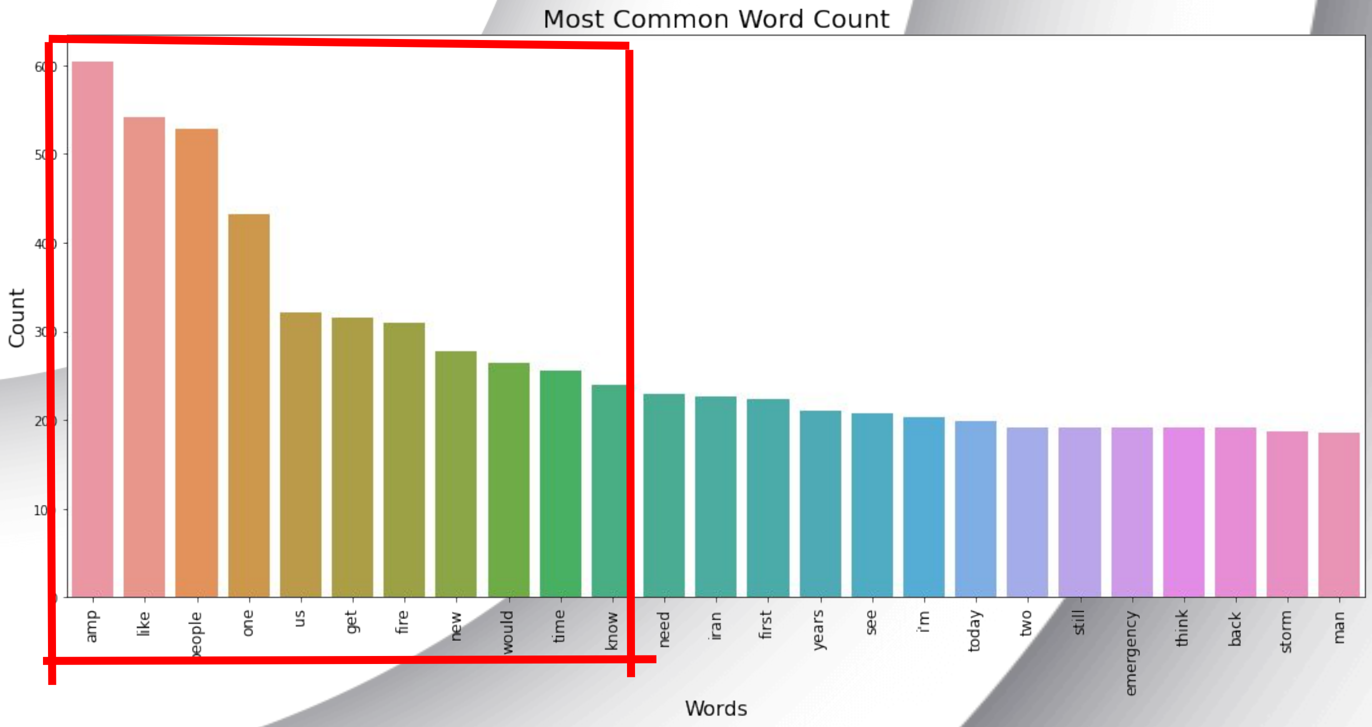
Scrub



- Removed punctuation, emojis, HTML, and URL's
- Removed Stop Words (the, an, and, etc.)
- Removed capital letters
- Tokenized and Created Word Counts
- Created word vectors & sequences

3. Explore

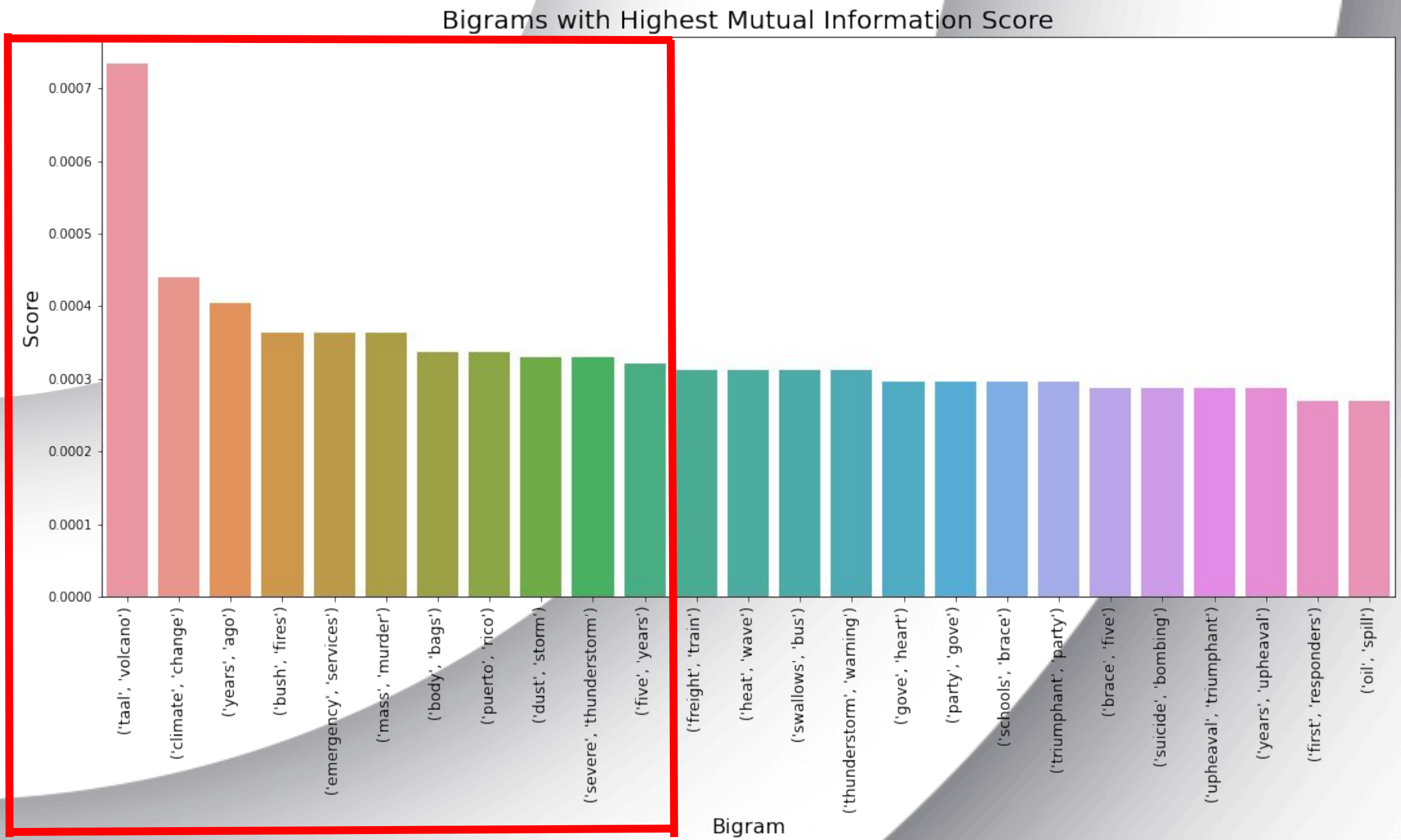
Word Count



<u>10 Most Common Words:</u>

- Amp
- Like
- People
- One
- Us
- Get
- Fire
- New
- Would
- Time
- Know

Bigrams



Top 10 Bigrams:

- Taal, volcano
- Climate, change
- Years, ago
- Bush, fires
- Emergency, services
- Mass, murder
- Body, bags
- Puerto, rico
- Dust, storm
- Severe, thunderstorm
- Five, years

Word2Vec

- Word2Vec allows us to explore relationships between words.
- Getting the 'most similar' for disaster-related words yields interesting results.

Fire

```
[('accident', 0.6279330253601074),
 ('iran', 0.6176088452339172),
 ('ukrainian', 0.59111421585083),
 ('shot', 0.5809152126312256),
 ('attack', 0.5783515572547913),
 ('airplane', 0.5768617391586304),
 ('like', 0.571844160556793 2),
 ('amp', 0.5559629201889038),
 ('blazing', 0.5528613924980164),
 ('army', 0.5281714200973511)]
```

Death

```
[('one', 0.5004619359970093),
 ('attack', 0.4937114119529724),
 ('army', 0.4733605086803436 3),
 ('attacked', 0.4692723751068115),
 ('know', 0.4676671028137207),
 ('bioterrorism', 0.4507286846637726),
 ('world', 0.4476444721221924),
 ('ambulance', 0.444823056459 4269),
 ('bleeding', 0.4361122548580169 7),
 ('accident', 0.4358267486095428 5)]
```

Accident

```
[('iran', 0.9049441814422607),
 ('shot', 0.9043893814086914),
 ('airplane', 0.8586768507957458),
 ('iranian', 0.8038906455039978),
 ('ukrainian', 0.8005987405776978),
 ('attack', 0.7705838084220886),
 ('shooting', 0.7558250427246094),
 ('people', 0.736716091632843),
 ('like', 0.7155638933181763),
 ('trump', 0.7147867679595947)]
```

4. Modeling

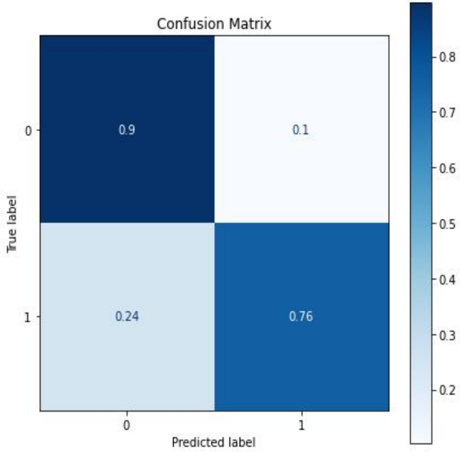
Classifiers

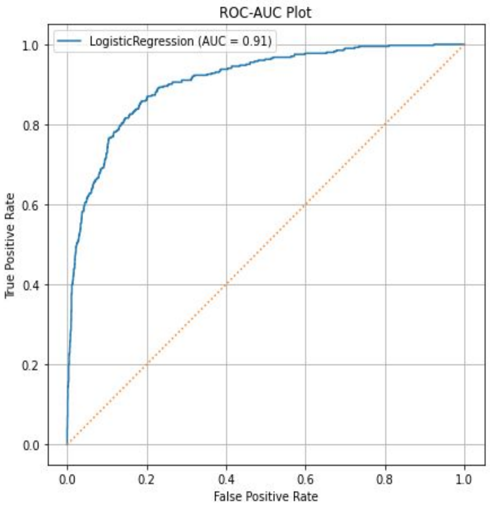
- Used a variety of classifiers
- Logistic Regression, Stochastic Gradient Descent, Random Forest, Stacking
- Logistic Regression outperformed all others.

```
Recall Scores for Each Model
--------------------------------
Logistic Regression:   0.76
SGD:   0.73
Random Forest:   0.53
Stacking Classifier:   0.68
```

Classification Report

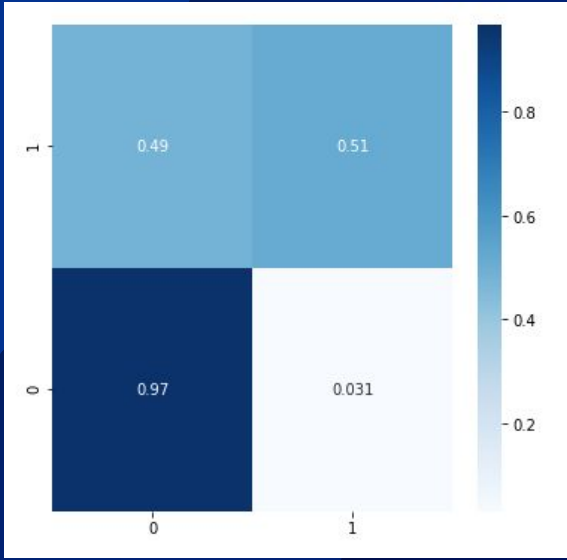
	precision	recall	f1-score	support
0	0.94	0.90	0.92	1872
1	0.61	0.76	0.68	402
accuracy			0.87	2274
macro avg	0.78	0.83	0.80	2274
weighted avg	0.89	0.87	0.88	2274

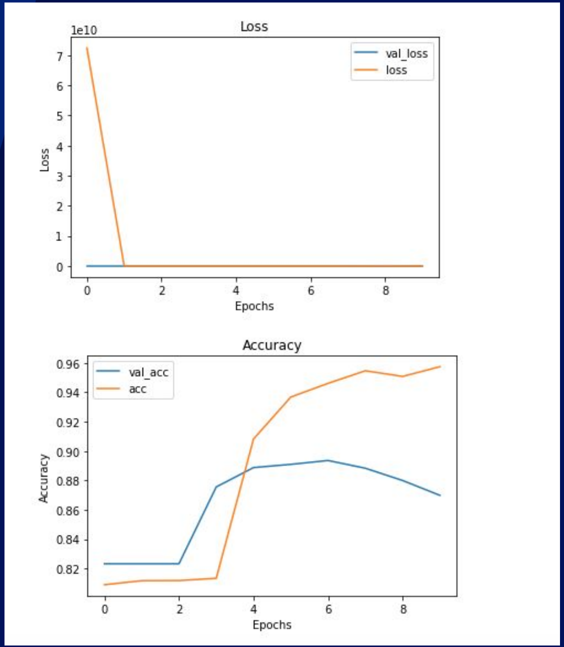




LSTM & GRU Neural Networks

- Powerful but extremely time consuming.
- Great accuracy but poor in terms of recall.
- Can be fine tuned and improved in the future.
- Scores are from best version of many iterations.

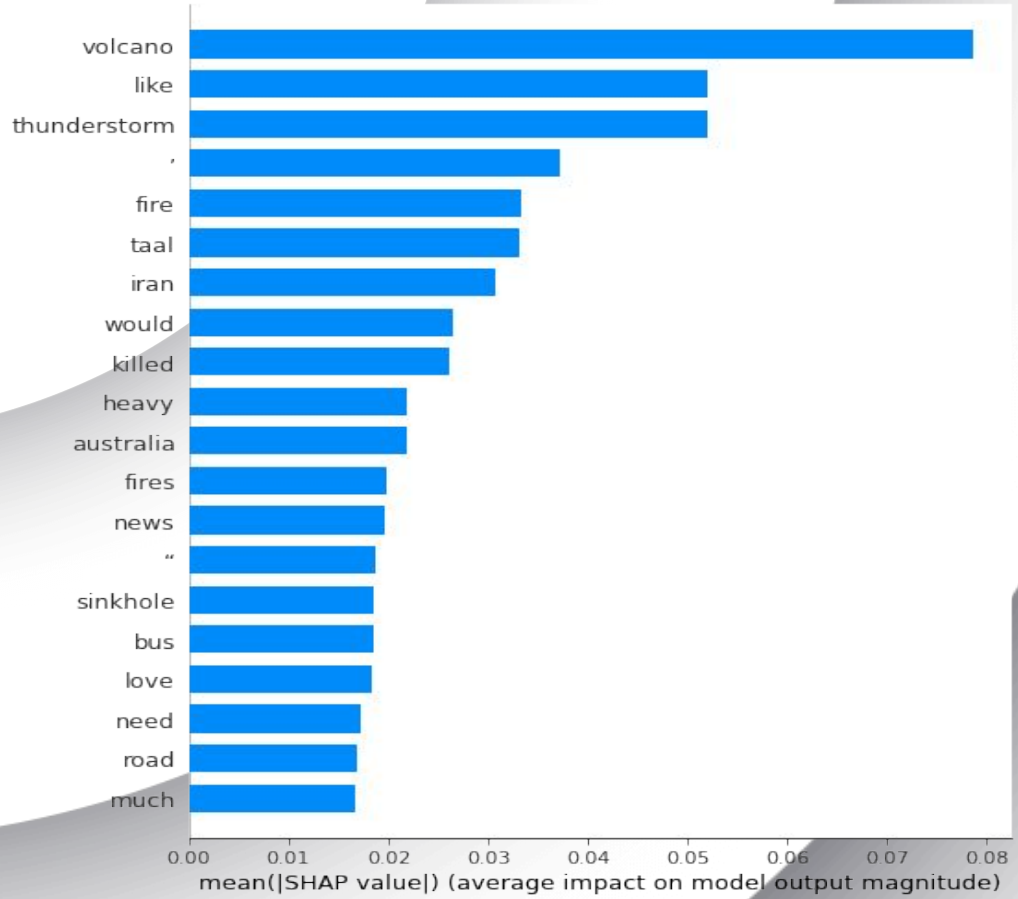





```
- 14s 188ms/step - loss: 0.4370 - acc: 0.8887
```

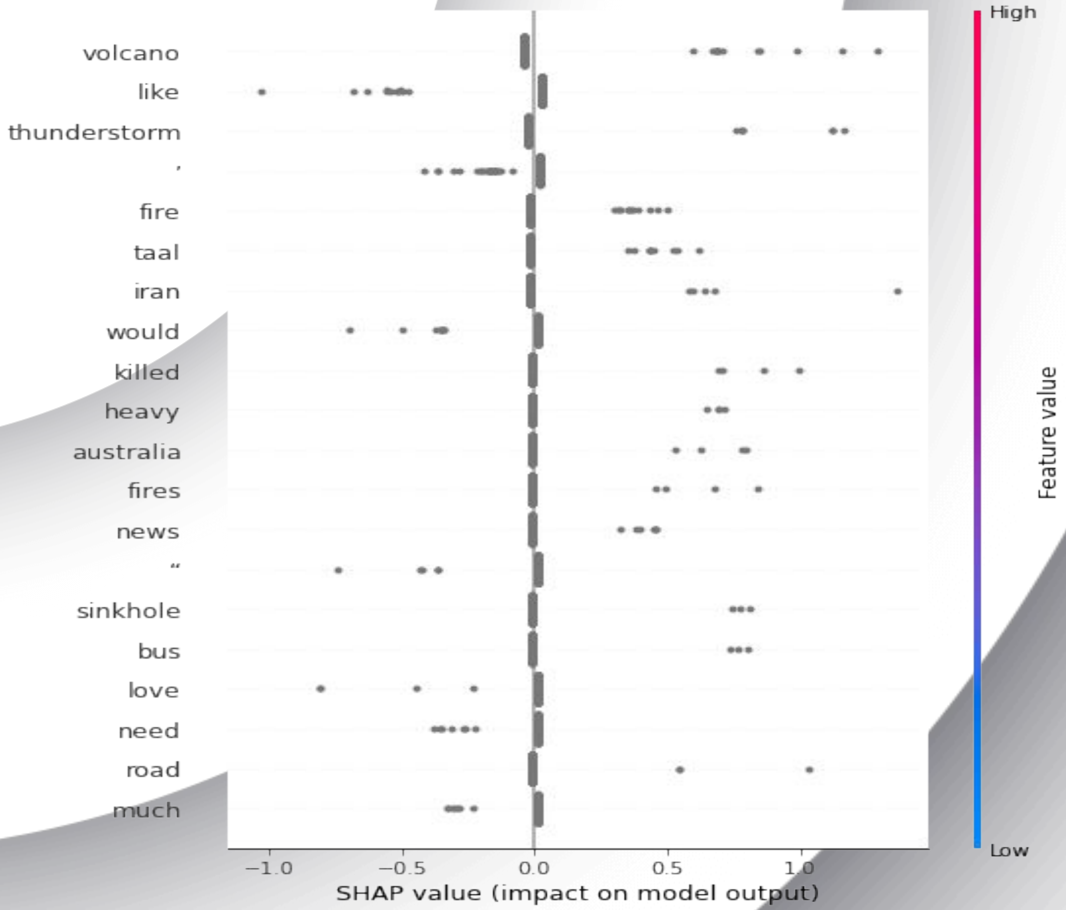
5. Results

Disaster Keywords



- **Top Words in Terms of Importance to Target**
- **Large portion refer to disasters**
- **Volcano, Thunderstorm, Fire, Killed, etc.**

Disaster



- These "disaster keywords" have highest level of importance in determining a positive disaster tweet.
- Few non-disaster words have positive importance

Results

- Basic NLP - Bigrams & Word2Vec "most similar" revealed interesting relationships between words in the dataset.

- The Logistic Regression model performed better than all other models that we attempted.

- Disaster-related keywords seem to have an extreme influence on whether a tweet is related to a disaster event.

6. Recommendations

Recommendations

Disaster Keywords

For basic NLP, focus on short phrases that include a “Disaster Keyword”. (fire, tornado, etc.)

Model Selection

In terms of both speed and accuracy, a Logistic Regression model is recommended.

Keep It Simple

While the power of the neural network is appealing, the accuracy and runtime make it unappealing for this particular task.

Future Work

With more time, we can improve our analysis in the following ways:

Fine Tuning

We can spend more time fine tuning our existing models in order to achieve better results.

Additional Models

There are several other types of models that can be applied in order to increase the accuracy of our results.

Additional Data

With more data, we can increase the accuracy of the models.

Thanks!

Any questions?

You can find me at:

- @spags093
- jeff.spags@gmail.com