NLP Text Classification of Disaster Tweets

Goals

1. Data Cleaning
2. EDA
3. Modeling

The Data

Source -> Kaggle.com

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

Data Cleaning

Data Cleaning

Keywords

	keyword	keyword cleaned
1345	burning%20buildings	burning buildings
4796	loud%20bang	loud bang
5115	nuclear%20reactor	nuclear reactor
7161	war%20zone	war zone
7318	wild%20fires	wild fires

Data Cleaning

Tweets

	text	pre_processed
0	Our Deeds are the Reason of this #earthquake M...	deeds reason earthquake may allah forgive us
1	Forest fire near La Ronge Sask. Canada	forest fire near la ronge sask canada
2	All residents asked to 'shelter in place' are ...	residents ask shelter place notify officer e...
3	13,000 people receive #wildfires evacuation or...	people receive wildfires evacuation order cal...
4	Just got sent this photo from Ruby #Alaska as ...	get send photo ruby alaska smoke wildfires p...

Data Cleaning

Keyword	target	
none	1	42
outbreak	1	38
debris	1	37
oil spill	1	37
wreckage	1	37
		..
blazing	1	1
body bag	1	1
epicentre	1	1
body bags	1	1
ruin	1	1

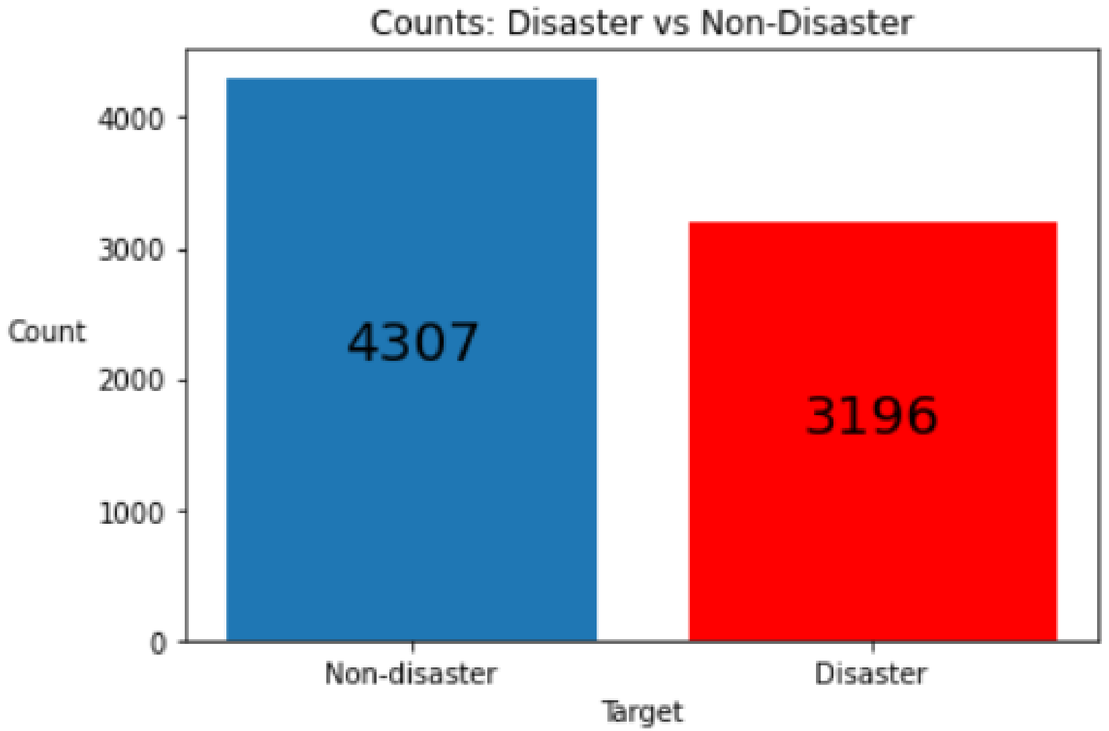
Length: 221, dtype: int64

Keyword	target	
typhoon	1	38
oil spill	1	38
outbreak	1	38
debris	1	38
wreckage	1	37
		..
electrocute	1	1
blazing	1	1
epicentre	1	1
body bags	1	1
died	1	1

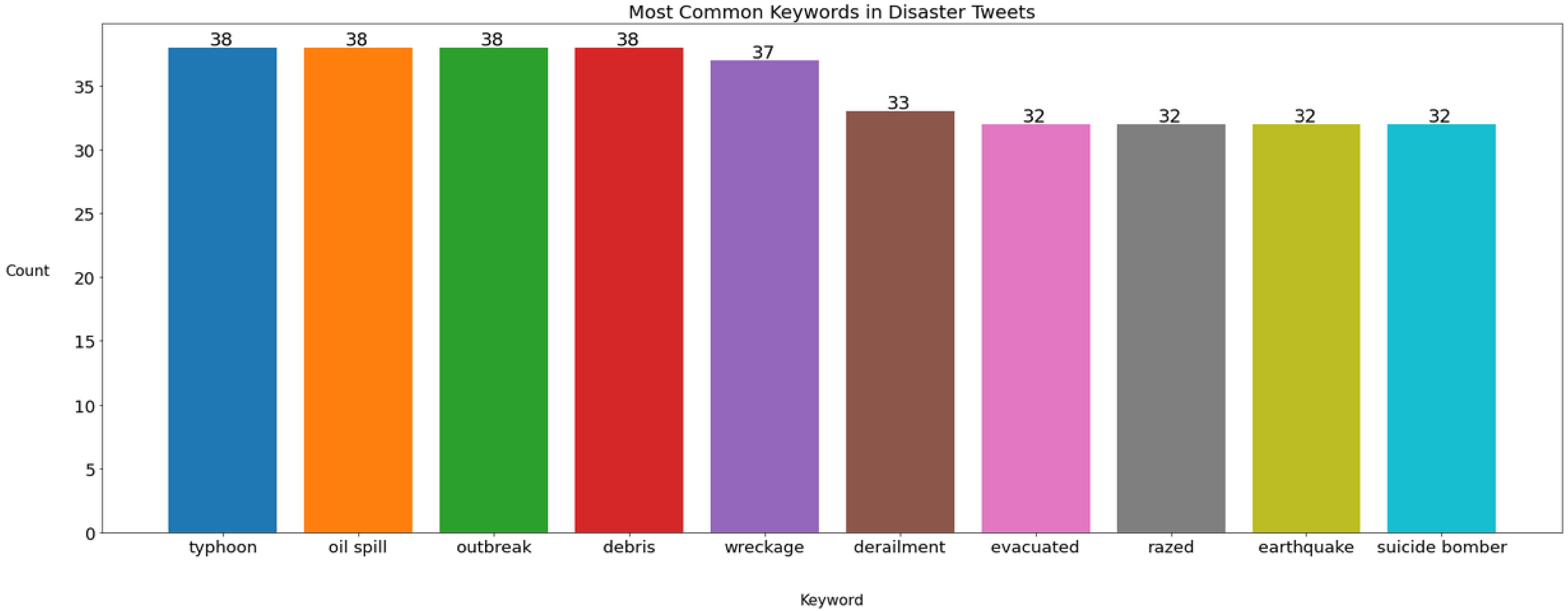
Length: 221, dtype: int64

EDA

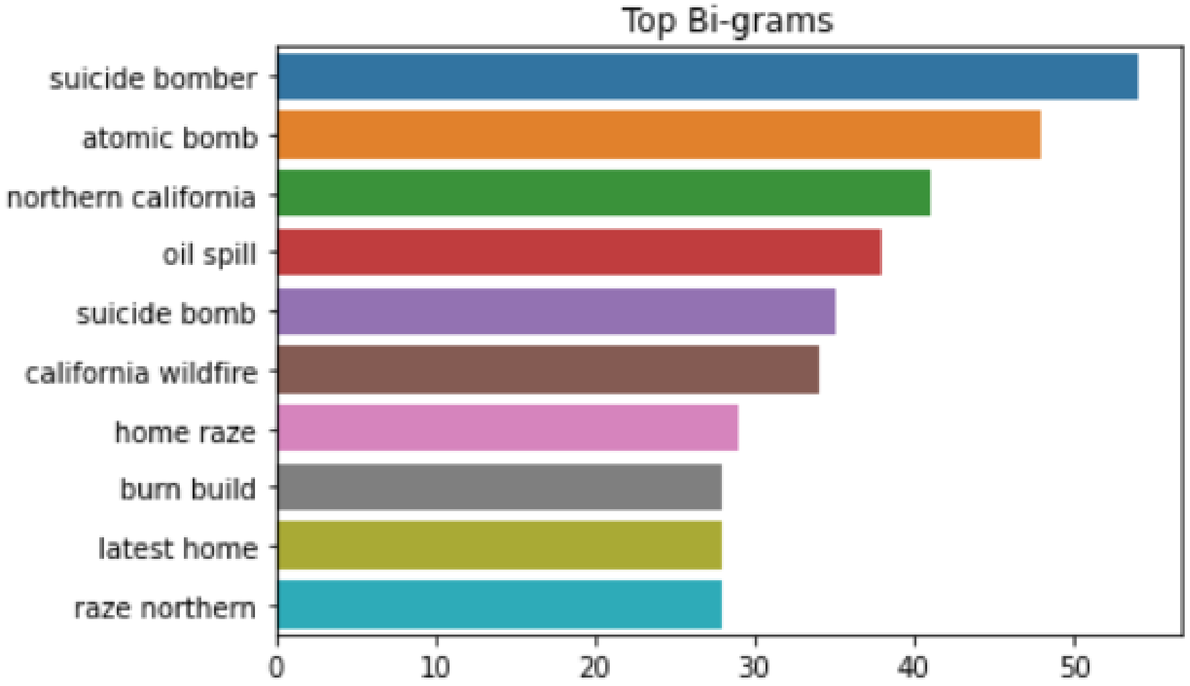
EDA

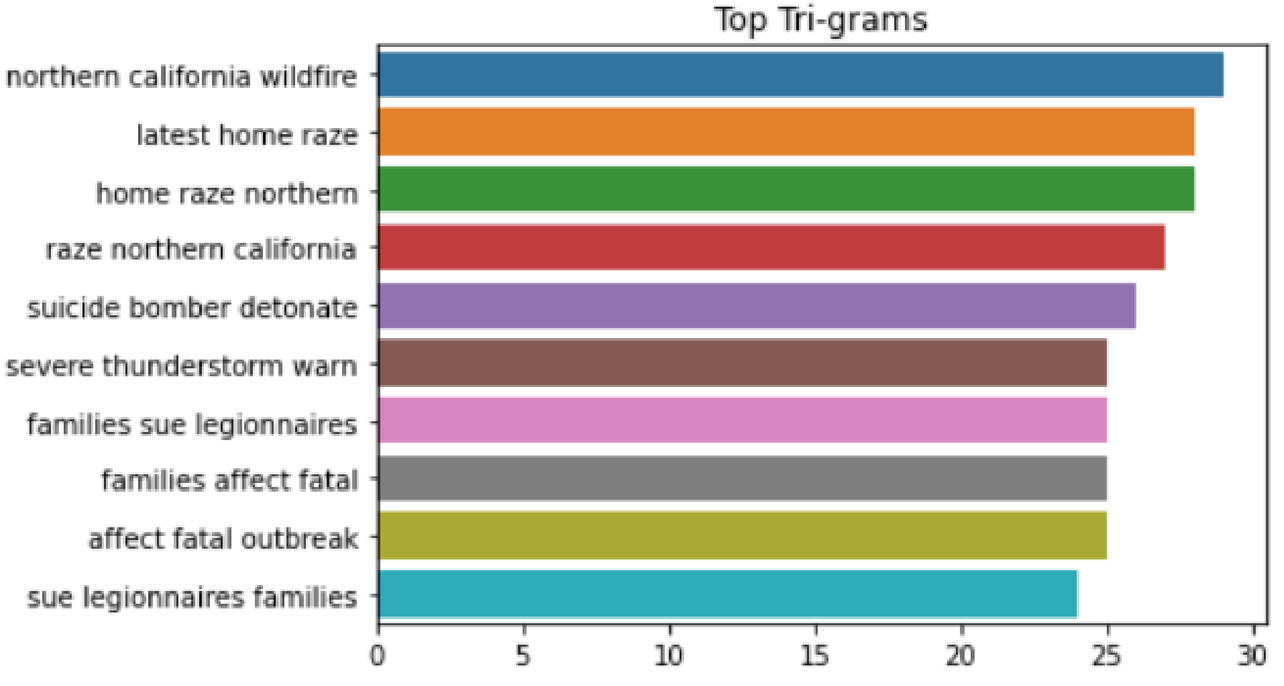

Counts: Disaster vs Non-Disaster
Count
4000
3000
2000
1000
0
4307
3196
Non-disaster
Disaster
Target

EDA


Most Common Keywords in Disaster Tweets
Count
Keyword
0
5
10
15
20
25
30
35
38
38
38
38
37
33
32
32
32
32
typhoon
oil spill
outbreak
debris
wreckage
derailment
evacuated
razed
earthquake
suicide bomber

EDA


Top Bi-grams
suicide bomber
atomic bomb
northern california
oil spill
suicide bomb
california wildfire
home raze
burn build
latest home
raze northern
0
10
20
30
40
50


Top Tri-grams
northern california wildfire
latest home raze
home raze northern
raze northern california
suicide bomber detonate
severe thunderstorm warn
families sue legionnaires
families affect fatal
affect fatal outbreak
sue legionnaires families
0
5
10
15
20
25
30

EDA


derailment
nuclear disaster
devastation
airplane accident
wild fires
mass murder
hostages
razed
debris
suicide bomber
casualties
sinkhole
oil spill
wildfire
evacuation
wreckage
sandstorm
flooding
earthquake
typhoon
rescuers
drought
collision
suicide bombing
evacuated
outbreak
bridge collapse

Pre-processing

Pre-processing

CountVectorizer

	aa	aaaa	aaaaaaallll	aaarrrgghhh	aaemiddleaged	aampb	aampw	aashiqui	ab	aba	...	zonewolf	zoom	zotar	zouma	zourryart	zss	zumiez	zurich	zxath
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	

5 rows × 17778 columns

Pre-processing

TfidfVectorizer

	aa	aaaa	aaaaaaallll	aaarrrgghhh	aaemiddleaged	aampb	aampw	aashiqui	ab	aba	...	zonewolf	zoom	zotar	zouma	zourryart	zss	zumiez	zurich	zxat
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

5 rows × 17778 columns

Modeling

Modeling

Surprisingly, Logistic Regression performed the best

	F1_Score	Accuracy
LogReg	0.789920	0.802772
RandFor	0.780207	0.794776
SVM	0.784138	0.801173

Improvements

I should show the predictions for
atleast the winning model

With better processing power, I'd love to be able to fine tune the models past a simple GridSearchCV.