Importing Libraries

In [135]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
from nltk.corpus import stopwords
from spellchecker import SpellChecker
from nltk.stem import WordNetLemmatizer
from tensorflow.keras.preprocessing.text import Tokenizer
from nltk.tokenize import word_tokenize
import nltk
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
from sklearn.linear_model import LogisticRegression
import tensorflow as tf
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.layers import Dense, Input
from tensorflow.keras.models import Model
from tensorflow.keras.callbacks import ModelCheckpoint
import re
import keras
from keras.models import Sequential
from keras.utils import to_categorical
from keras.layers import Embedding, LSTM
```

In [3]:
```
train_df = pd.read_csv('/Users/ashishpodar/Desktop/nlp-getting-started/train.csv')
```

In [4]:
```
test_df = pd.read_csv('/Users/ashishpodar/Desktop/nlp-getting-started/test.csv')
```

In [5]:
```
train_df.head()
```

Out[5]:

	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

Missing Values

In [6]:
```
train_df.isnull().sum()
```

Out[6]:
```
id             0
keyword       61
location    2533
text           0
target         0
dtype: int64
```

Visualizations

In [7]:
```
#Locations based on non disasters
```

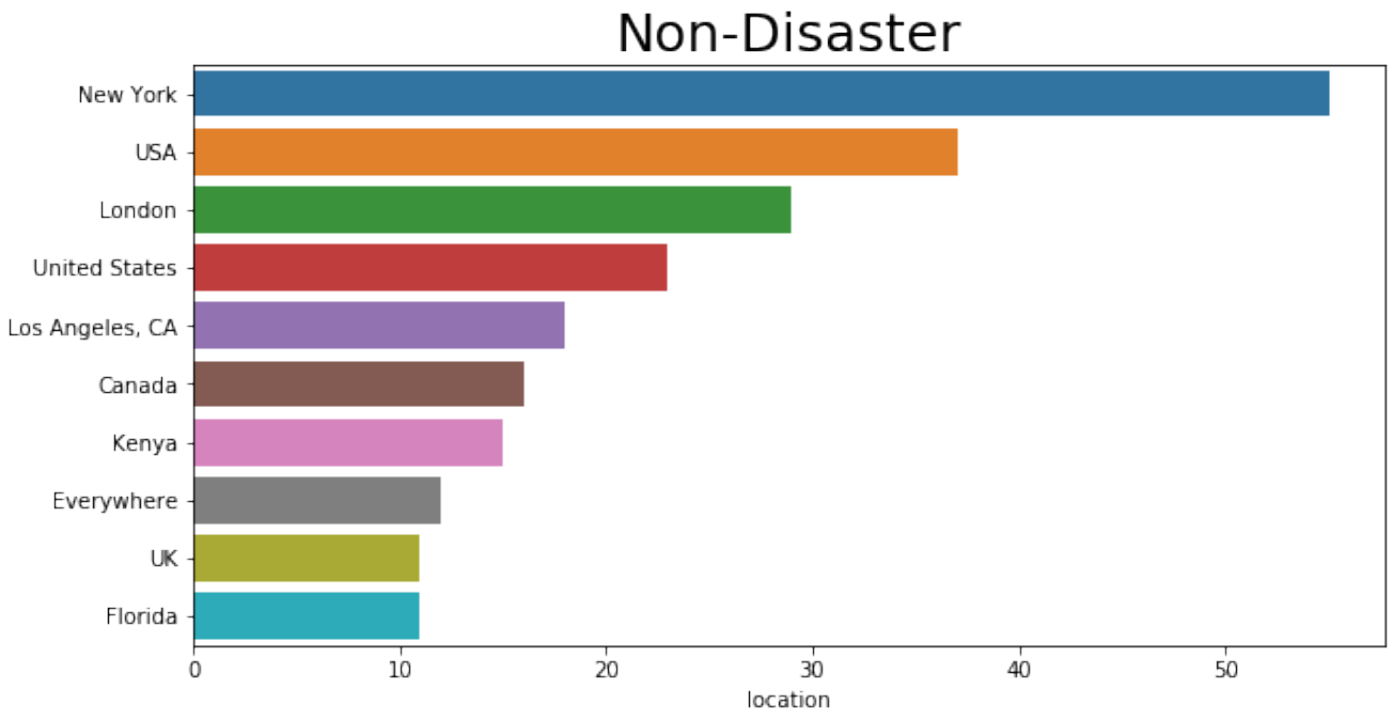
In [8]:
```
#Top 10 non-disaster locations
train_df[train_df['target']==0]['location'].value_counts()[:10]
```

Out[8]:
```
New York           55
USA                37
London             29
United States      23
Los Angeles, CA    18
Canada             16
Kenya              15
Everywhere         12
UK                 11
Florida            11
Name: location, dtype: int64
```

In [9]:
```
fig, axes = plt.subplots(1, figsize=(10, 5))

sns.barplot(ax=axes, x = train_df[train_df['target'] == 0]['location']
            .value_counts()[:10],
           y = train_df[train_df['target'] == 0]['location']
            .value_counts().index[:10])
            .set_title('Non-Disaster',size=25)
```

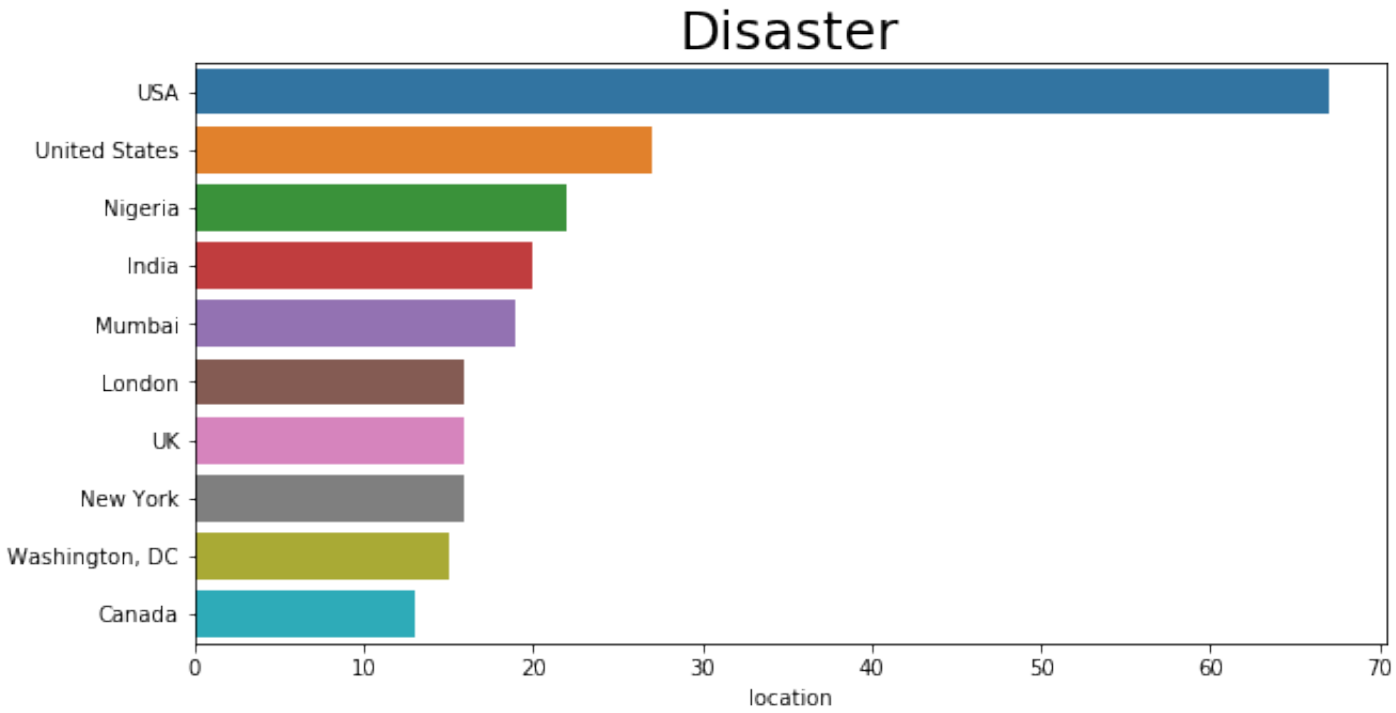
Out[9]: Text(0.5, 1.0, 'Non-Disaster')



In [10]:
```
#Locations based on disasters
fig, axes = plt.subplots(1, figsize=(10, 5))

sns.barplot(ax=axes, x = train_df[train_df['target'] == 1]['location']
            .value_counts()[:10],
           y = train_df[train_df['target'] == 1]['location']
            .value_counts().index[:10])
            .set_title('Disaster',size=25)
```

Out[10]: Text(0.5, 1.0, 'Disaster')

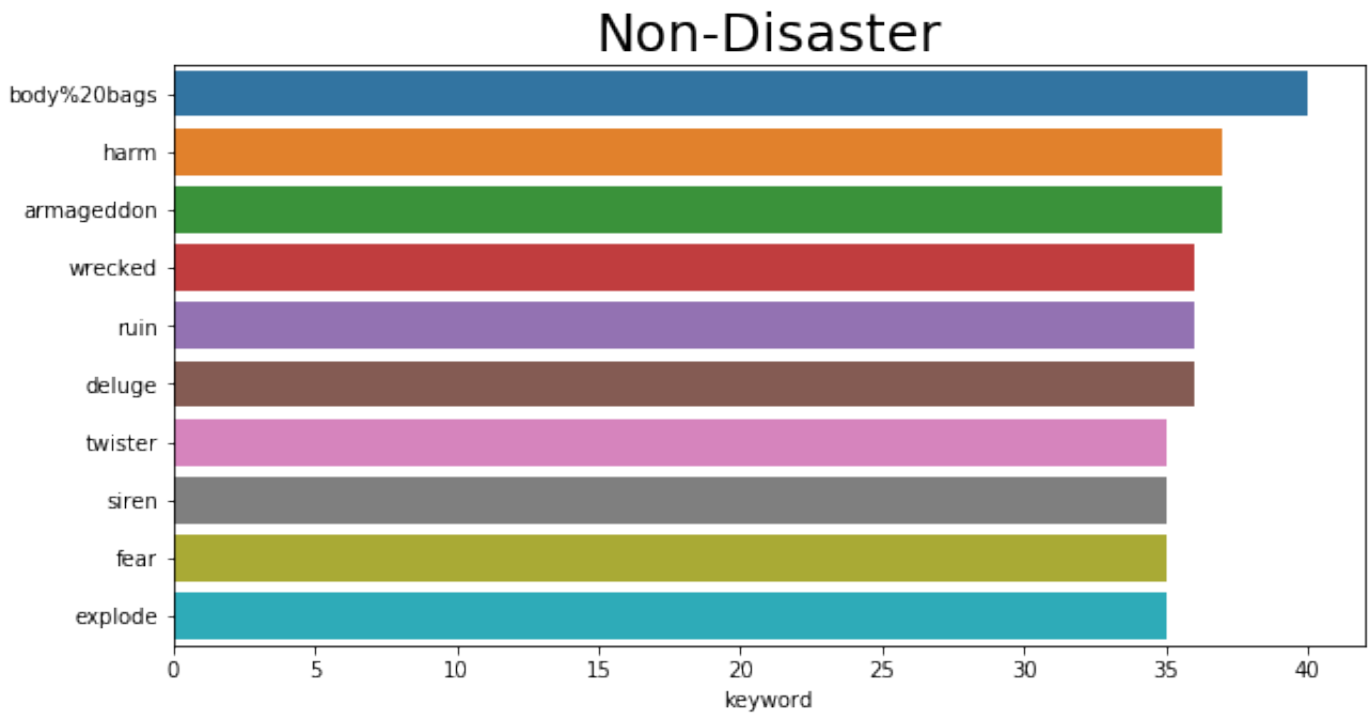


In [11]:
```
#Keywords based on non-disasters
```

In [12]:
```
fig, axes = plt.subplots(1, figsize=(10, 5))

sns.barplot(ax=axes, x = train_df[train_df['target'] == 0]['keyword']
            .value_counts()[:10],
           y = train_df[train_df['target'] == 0]['keyword']
            .value_counts().index[:10])
            .set_title('Non-Disaster',size=25)
```

Out[12]: Text(0.5, 1.0, 'Non-Disaster')

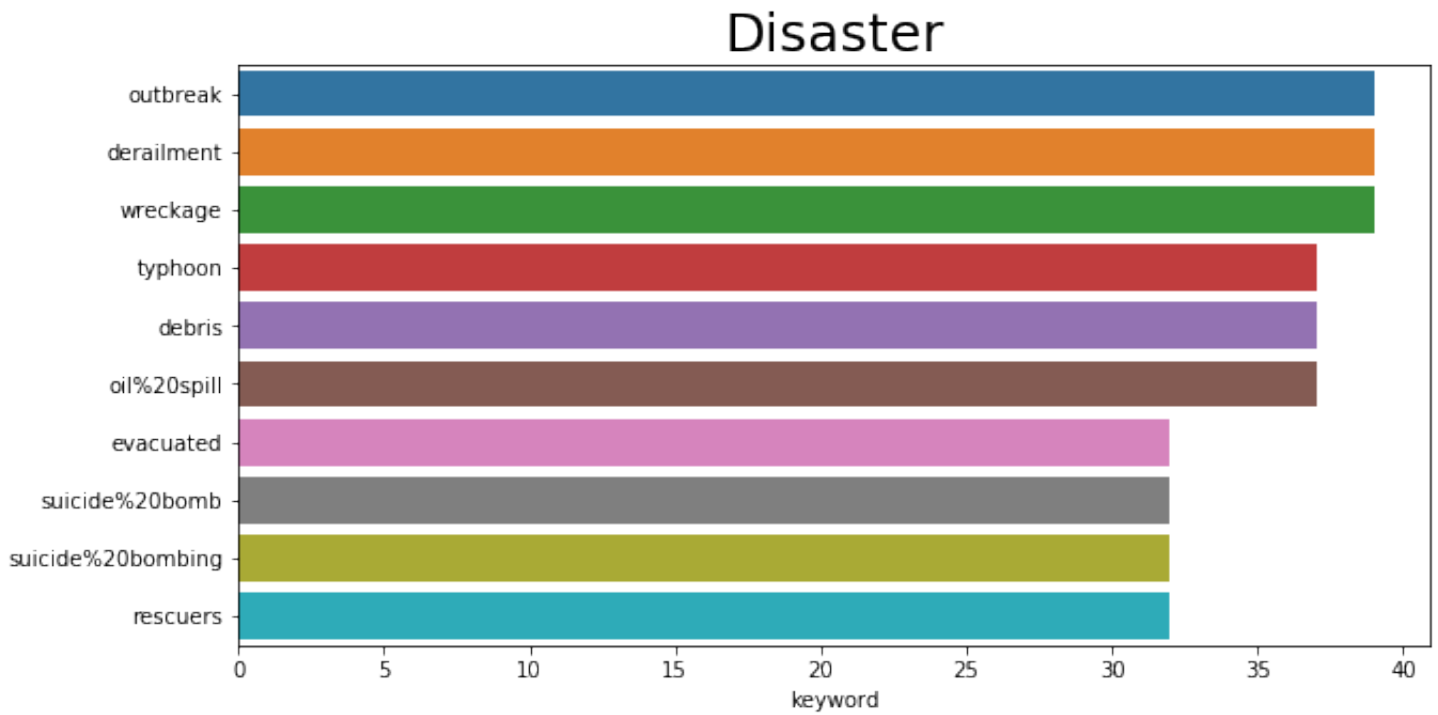


In [13]:
```
#Keywords based on Disasters
```

In [14]:
```
fig, axes = plt.subplots(1, figsize=(10, 5))

sns.barplot(ax=axes, x = train_df[train_df['target'] == 1]['keyword']
            .value_counts()[:10],
           y = train_df[train_df['target'] == 1]['keyword']
            .value_counts().index[:10])
            .set_title('Disaster',size=25)
```

Out[14]: Text(0.5, 1.0, 'Disaster')



Preprocessing

In [15]:
```
#Dropping Useless Columns
```

In [16]:
```
train_df = train_df.drop(['id','keyword', 'location'], axis = 1)
train_df
```

Out[16]:

	text	target
0	Our Deeds are the Reason of this #earthquake M...	1
1	Forest fire near La Ronge Sask. Canada	1
2	All residents asked to 'shelter in place' are ...	1
3	13,000 people receive #wildfires evacuation or...	1
4	Just got sent this photo from Ruby #Alaska as ...	1
...	...	...
7608	Two giant cranes holding a bridge collapse int...	1
7609	@aria_ahrary @TheTawniest The out of control w...	1
7610	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	1
7611	Police investigating after an e-bike collided ...	1
7612	The Latest: More Homes Razed by Northern Calif...	1

7613 rows × 2 columns

In [17]:
```
test_df = test_df.drop(['id','keyword', 'location'], axis = 1)
test_df
```

Out[17]:

	text
0	Just happened a terrible car crash
1	Heard about #earthquake is different cities, s...
2	there is a forest fire at spot pond, geese are...
3	Apocalypse lighting. #Spokane #wildfires
4	Typhoon Soudelor kills 28 in China and Taiwan
...	...
3258	EARTHQUAKE SAFETY LOS ANGELES ÛÒ SAFETY FASTE...
3259	Storm in RI worse than last hurricane. My city...
3260	Green Line derailment in Chicago http://t.co/U...
3261	MEG issues Hazardous Weather Outlook (HWO) htt...
3262	#CityofCalgary has activated its Municipal Eme...

3263 rows × 1 columns

In [19]:
```
#Removing Hashtags
```

In [24]:
```
train_df['text'] = train_df.apply(lambda row : re.sub("\B#\w+", "", row['text']),axis = 1)
train_df
```

Out[24]:

	text	target
0	Our Deeds are the Reason of this May ALLAH Fo...	1
1	Forest fire near La Ronge Sask. Canada	1
2	All residents asked to 'shelter in place' are ...	1
3	13,000 people receive evacuation orders in Ca...	1
4	Just got sent this photo from Ruby as smoke f...	1
...	...	...
7608	Two giant cranes holding a bridge collapse int...	1
7609	@aria_ahrary @TheTawniest The out of control w...	1
7610	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	1
7611	Police investigating after an e-bike collided ...	1
7612	The Latest: More Homes Razed by Northern Calif...	1

7613 rows × 2 columns

In [25]:
```
test_df['text'] = test_df.apply(lambda row : re.sub("\B#\w+", "", row['text']),axis = 1)
test_df
```

Out[25]:

	text
0	Just happened a terrible car crash
1	Heard about is different cities, stay safe ev...
2	there is a forest fire at spot pond, geese are...
3	Apocalypse lighting.
4	Typhoon Soudelor kills 28 in China and Taiwan
...	...
3258	EARTHQUAKE SAFETY LOS ANGELES ÛÒ SAFETY FASTE...
3259	Storm in RI worse than last hurricane. My city...
3260	Green Line derailment in Chicago http://t.co/U...
3261	MEG issues Hazardous Weather Outlook (HWO) htt...
3262	has activated its Municipal Emergency Plan.

3263 rows × 1 columns

In [26]:
```
#Removing @ mentions
```

In [27]:
```
train_df['text'] = train_df.apply(lambda row : re.sub("\B@\w+", "", row['text']),axis = 1)
train_df
```

Out[27]:

	text	target
0	Our Deeds are the Reason of this May ALLAH Fo...	1
1	Forest fire near La Ronge Sask. Canada	1
2	All residents asked to 'shelter in place' are ...	1
3	13,000 people receive evacuation orders in Ca...	1
4	Just got sent this photo from Ruby as smoke f...	1
...	...	...
7608	Two giant cranes holding a bridge collapse int...	1
7609	The out of control wild fires in California ...	1
7610	M1.94 [01:04 UTC]?5km S of Volcano Hawaii. htt...	1
7611	Police investigating after an e-bike collided ...	1
7612	The Latest: More Homes Razed by Northern Calif...	1

7613 rows × 2 columns

In [28]:
```
test_df['text'] = test_df.apply(lambda row : re.sub("\B@\w+", "", row['text']),axis = 1)
test_df
```

Out[28]:

	text
0	Just happened a terrible car crash
1	Heard about is different cities, stay safe ev...
2	there is a forest fire at spot pond, geese are...
3	Apocalypse lighting.
4	Typhoon Soudelor kills 28 in China and Taiwan
...	...
3258	EARTHQUAKE SAFETY LOS ANGELES ÛÒ SAFETY FASTE...
3259	Storm in RI worse than last hurricane. My city...
3260	Green Line derailment in Chicago http://t.co/U...
3261	MEG issues Hazardous Weather Outlook (HWO) htt...
3262	has activated its Municipal Emergency Plan.

3263 rows × 1 columns

In [29]:
```
#Uniform casing
```

In [30]:
```
train_df['text'] = train_df['text'].str.lower()
train_df
```

Out[30]:

	text	target
0	our deeds are the reason of this may allah fo...	1
1	forest fire near la ronge sask. canada	1
2	all residents asked to 'shelter in place' are ...	1
3	13,000 people receive evacuation orders in ca...	1
4	just got sent this photo from ruby as smoke f...	1
...	...	...
7608	two giant cranes holding a bridge collapse int...	1
7609	the out of control wild fires in california ...	1
7610	m1.94 [01:04 utc]?5km s of volcano hawaii. htt...	1
7611	police investigating after an e-bike collided ...	1
7612	the latest: more homes razed by northern calif...	1

7613 rows × 2 columns

In [31]:
```
test_df['text'] = test_df['text'].str.lower()
test_df
```

Out[31]:

	text
0	just happened a terrible car crash
1	heard about is different cities, stay safe ev...
2	there is a forest fire at spot pond, geese are...
3	apocalypse lighting.
4	typhoon soudelor kills 28 in china and taiwan
...	...
3258	earthquake safety los angeles ûò safety faste...
3259	storm in ri worse than last hurricane. my city...
3260	green line derailment in chicago http://t.co/u...
3261	meg issues hazardous weather outlook (hwo) htt...
3262	has activated its municipal emergency plan.

3263 rows × 1 columns

In [32]:
```
#removing extra white spaces
```

In [33]:
```
train_df['text'] = train_df['text'].str.strip()
train_df
```

Out[33]:

	text	target
0	our deeds are the reason of this may allah fo...	1
1	forest fire near la ronge sask. canada	1
2	all residents asked to 'shelter in place' are ...	1
3	13,000 people receive evacuation orders in ca...	1
4	just got sent this photo from ruby as smoke f...	1
...	...	...
7608	two giant cranes holding a bridge collapse int...	1
7609	the out of control wild fires in california ev...	1
7610	m1.94 [01:04 utc]?5km s of volcano hawaii. htt...	1
7611	police investigating after an e-bike collided ...	1
7612	the latest: more homes razed by northern calif...	1

7613 rows × 2 columns

In [34]:
```
test_df['text'] = test_df['text'].str.strip()
test_df
```

Out[34]:

	text
0	just happened a terrible car crash
1	heard about is different cities, stay safe ev...
2	there is a forest fire at spot pond, geese are...
3	apocalypse lighting.
4	typhoon soudelor kills 28 in china and taiwan
...	...
3258	earthquake safety los angeles ûò safety faste...
3259	storm in ri worse than last hurricane. my city...
3260	green line derailment in chicago http://t.co/u...
3261	meg issues hazardous weather outlook (hwo) htt...
3262	has activated its municipal emergency plan.

3263 rows × 1 columns

In [35]:
```
#Removing Punctuations
```

In [36]:
```
train_df["text"] = train_df['text'].str.replace('[^\w\s]','')
train_df
```

Out[36]:

	text	target
0	our deeds are the reason of this may allah fo...	1
1	forest fire near la ronge sask canada	1
2	all residents asked to shelter in place are be...	1
3	13000 people receive evacuation orders in cal...	1
4	just got sent this photo from ruby as smoke f...	1
...	...	...
7608	two giant cranes holding a bridge collapse int...	1
7609	the out of control wild fires in california ev...	1
7610	m194 0104 utc5km s of volcano hawaii httptcozd...	1
7611	police investigating after an ebike collided w...	1
7612	the latest more homes razed by northern califo...	1

7613 rows × 2 columns

In [37]:
```
test_df["text"] = test_df['text'].str.replace('[^\w\s]','')
test_df
```

Out[37]:

	text
0	just happened a terrible car crash
1	heard about is different cities stay safe eve...
2	there is a forest fire at spot pond geese are ...
3	apocalypse lighting
4	typhoon soudelor kills 28 in china and taiwan
...	...
3258	earthquake safety los angeles ûò safety fasten...
3259	storm in ri worse than last hurricane my citya...
3260	green line derailment in chicago httptcoutbxlc...
3261	meg issues hazardous weather outlook hwo httpt...
3262	has activated its municipal emergency plan

3263 rows × 1 columns

In [38]:
```
#Removing Stopwords
```

In [39]:
```
stop = stopwords.words('english')
```

In [40]:
```
train_df['text'] = train_df['text'].apply(lambda x: ' '.join([word for word in x.split() if word not i
train_df
```

Out[40]:

	text	target
0	deeds reason may allah forgive us	1
1	forest fire near la ronge sask canada	1
2	residents asked shelter place notified officer...	1
3	13000 people receive evacuation orders california	1
4	got sent photo ruby smoke pours school	1
...	...	...
7608	two giant cranes holding bridge collapse nearb...	1
7609	control wild fires california even northern pa...	1
7610	m194 0104 utc5km volcano hawaii httptcozdtoyd8ebj	1
7611	police investigating ebike collided car little...	1
7612	latest homes razed northern california wildfir...	1

7613 rows × 2 columns

In [41]:
```
test_df['text'] = test_df['text'].apply(lambda x: ' '.join([word for word in x.split() if word not in
test_df
```

Out[41]:

	text
0	happened terrible car crash
1	heard different cities stay safe everyone
2	forest fire spot pond geese fleeing across str...
3	apocalypse lighting
4	typhoon soudelor kills 28 china taiwan
...	...
3258	earthquake safety los angeles ûò safety fasten...
3259	storm ri worse last hurricane cityamp3others h...
3260	green line derailment chicago httptcoutbxlcbiuy
3261	meg issues hazardous weather outlook hwo httpt...
3262	activated municipal emergency plan

3263 rows × 1 columns

In [42]:
```
#removing special characters
```

In [43]:
```
def remove_special_characters(text):
    pattern = r'[^a-zA-Z]'
    text = re.sub(pattern, ' ', text)
    return text

train_df['text'] = train_df['text'].apply(remove_special_characters)
test_df['text'] = test_df['text'].apply(remove_special_characters)
```

In [44]:
```
#Spell Checking
```

In [45]:
```
spell = SpellChecker()
```

In [46]:
```
for line in train_df['text']:
    misspelled_words = spell.unknown(line.split())
```

In [47]:
```
misspelled_words
```

Out[47]: {'d', 'httptcoymy', 'rskq'}

In [90]:
```
for line in test_df['text']:
    misspelled_words = spell.unknown(line.split())
```

In [91]:
```
misspelled_words
```

Out[91]: set()

In [92]:
```
#Nothing Significant
```

In [96]:
```
#Tokenization and Lemmatization
```

In [48]:
```
w_tokenizer = nltk.tokenize.WhitespaceTokenizer()
lemmatizer = nltk.stem.WordNetLemmatizer()
```

In [49]:
```
def lemmatize_text(text):
    return [lemmatizer.lemmatize(w) for w in w_tokenizer.tokenize(text)]

df = train_df.text.apply(lemmatize_text)
```

In [50]:
```
df
```

Out[50]:
```
0                  [deed, reason, may, allah, forgive, u]
1            [forest, fire, near, la, ronge, sask, canada]
2       [resident, asked, shelter, place, notified, of...
3        [people, receive, evacuation, order, california]
4          [got, sent, photo, ruby, smoke, pours, school]
                              ...
7608    [two, giant, crane, holding, bridge, collapse,...
7609    [control, wild, fire, california, even, northe...
7610    [m, utc, km, volcano, hawaii, httptcozdtoyd, ebj]
7611    [police, investigating, ebike, collided, car, ...
7612    [latest, home, razed, northern, california, wi...
Name: text, Length: 7613, dtype: object
```

In [139]:
```

```

In [ ]:
```

```

In [51]:
```
def prepare_data(train_docs, test_docs, mode):
  tokenizer = Tokenizer()
  tokenizer.fit_on_texts(train_docs)
  Xtrain = tokenizer.texts_to_matrix(train_docs, mode=mode)
  Xtest = tokenizer.texts_to_matrix(test_docs, mode=mode)
  return Xtrain, Xtest
```

In [52]:
```
t1,t2 = prepare_data(df,test_df['text'],'count')
```

Train-Test Splitting

In [53]:
```
X = t1
y = train_df['target']
```

In [54]:
```
X_train, X_test, y_train, y_test = train_test_split(X,y)
```

Trying out ML Models

In [55]:
```
#RandomForestClassifier
rf = RandomForestClassifier()
rf_model = rf.fit(X_train, y_train)
y_pred = rf_model.predict(X_test)
```

In [56]:
```
metrics.accuracy_score(y_test, y_pred)
```

Out[56]: 0.7988445378151261

In [129]:
```
cross_val_score(rf_model, X_test, y_test, cv=5,scoring='accuracy')
```

Out[129]:
```
array([0.78896673, 0.76007005, 0.76707531, 0.7915937 , 0.77651183])
```

In [57]:
```
#Logistic Regression
lr = LogisticRegression()
lr_model = lr.fit(X_train, y_train)
lr_y_pred = lr_model.predict(X_test)
```

In [58]:
```
metrics.accuracy_score(y_test, lr_y_pred)
```

Out[58]:
```
0.8046218487394958
```

In [137]:
```
cross_val_score(lr_model, X_test, y_test, cv=5,scoring='accuracy')
```

Out[137]:
```
array([0.77427822, 0.76377953, 0.75328084, 0.79265092, 0.78157895])
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```