Real or Not? NLP with Disaster Tweets

Social media has become one of the most important part of our day-to-day life. We get information about all the topics, starting from news, stock markets, retail related etc. Since these applications also provide facility of trending important topics of the day, what is going on in the worlds is simply now one swap away!

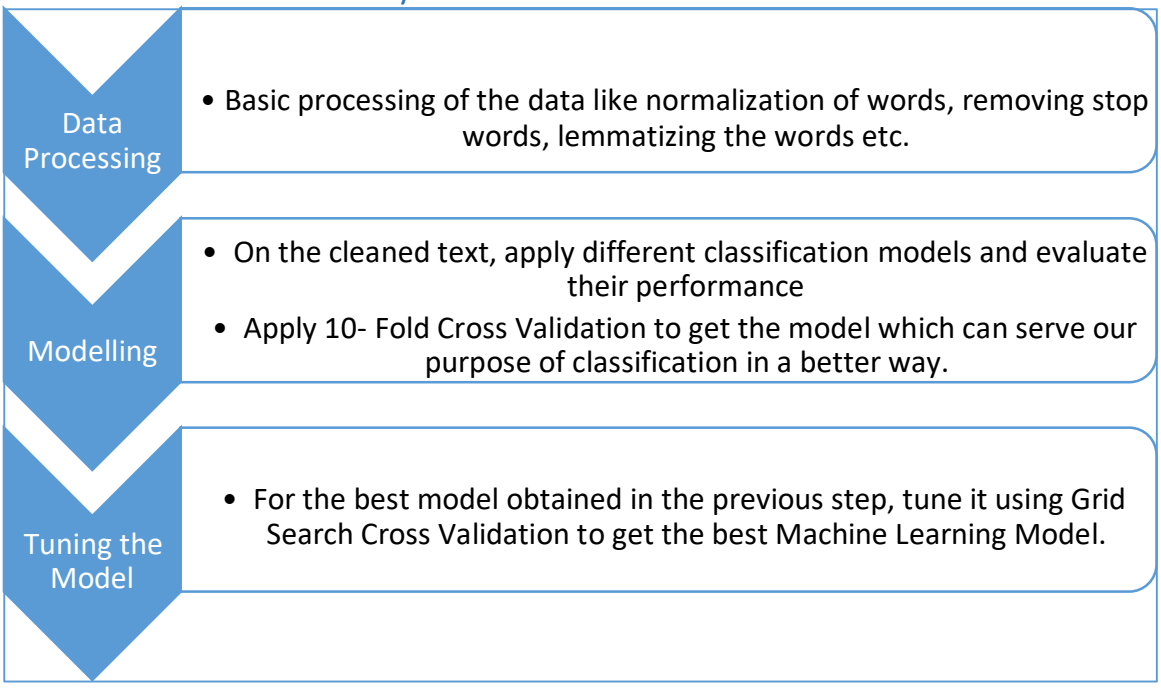
One such application is Twitter where we see the users sharing their opinions on variety of topics like, politics, movies, etc. It has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in programmatically monitoring Twitter (i.e. disaster relief organizations and news agencies).

But, it's not always clear whether a person's words are actually announcing a disaster.

In this project, we will be building a machine learning model that predicts which Tweets are about real disasters and which one's aren't.

Dataset is taken from Kaggle: https://www.kaggle.com/c/nlp-getting-started/data

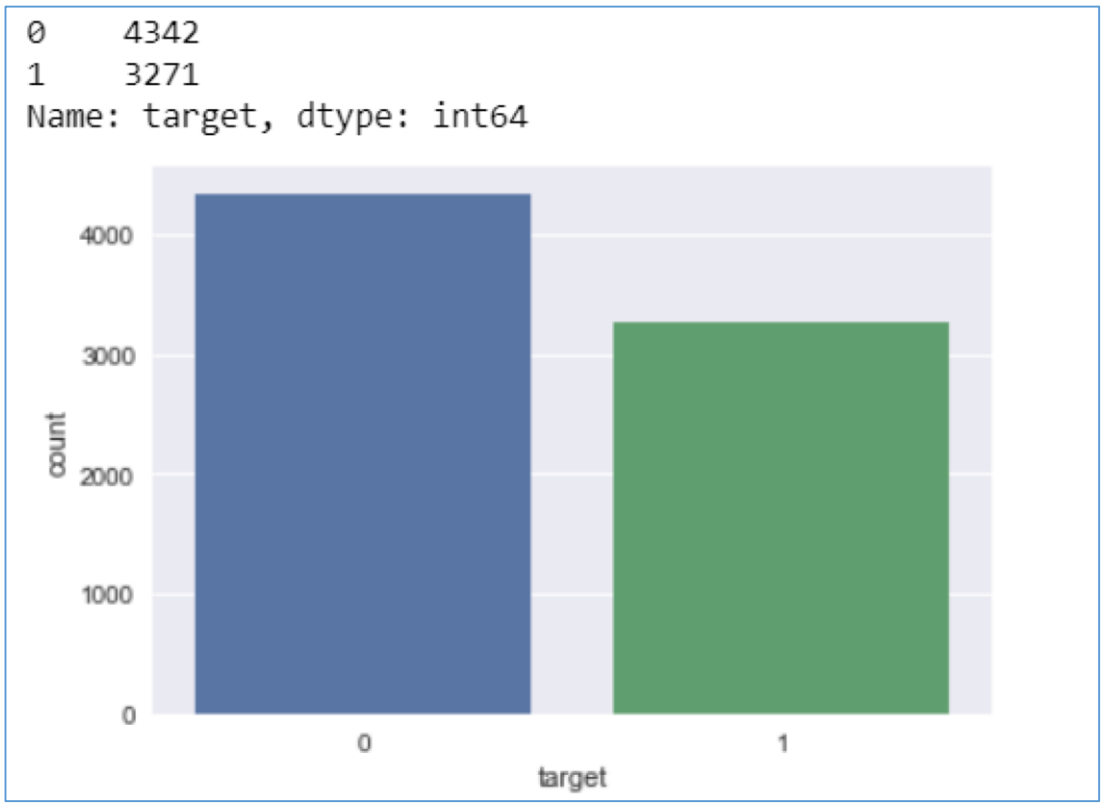
Process followed in the study:



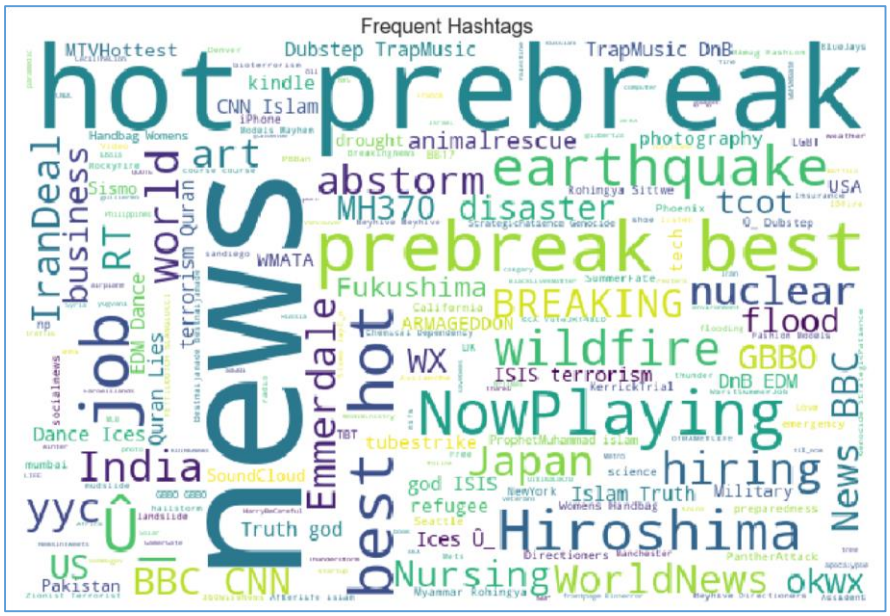
Exploratory Data Analysis on the dataset

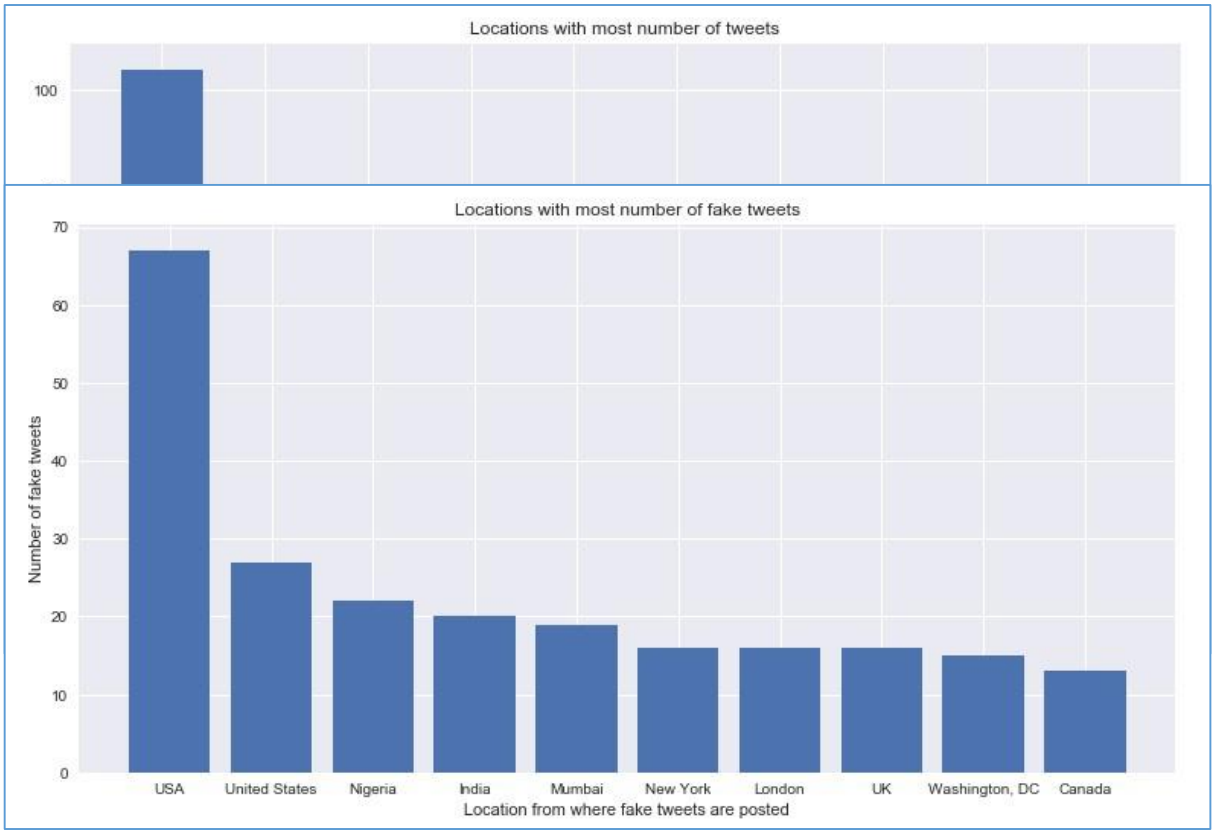
1. Target Label Count in the training dataset :

```
0    4342
1    3271
Name: target, dtype: int64
```



2. Most Frequent HashTags in the dataset :





Hmm, so we have quite a lot of hashtags flowing in our dataset. Most Frequent Ones are: news, prebreak, best, nowplaying etc.

3. Location with most number of tweets :
4. Location with most number of fake tweets :

We can infer from the above two plots that the most number of tweets are from USA and most number of fake tweets are also from USA

Data Processing

Since, here we are dealing with text data, hence there is a need to process these tweets in order to keep only relevant information which is best suited for modelling. In this data preparation process, we will:

1. Normalize the tweets to the lower format.
2. Remove tags like #, @ from the words in the text.
3. Remove punctuations.
4. Remove stop words like is, the, on etc.

5. Remove words which are of length less than 2.
6. Remove alpha-numeric string.
7. Reduce the words to their root forms using Lemmatization.

Sample execution from the method written for getting the cleaned text:

```
# Let's test the get_cleaned_text method
print(f'Original Text : {training_dataset.text[0]}')
print(f'Cleaned Text : {get_cleaned_text(training_dataset.text[0])}')
```

```
Original Text : Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all
Cleaned Text : deed reason earthquake may allah forgive
```

We can clearly see that:

1. Text is transformed into lower case.
2. # is removed from earthquake.
3. Stop words are removed.
4. Lemmatization of words : Reasons --> reason

Applying Different Machine Learning Algorithms on Balanced and Imbalanced Dataset

Algorithm	With Imbalanced Data		With Balanced Data	
	Accuracy Without Cross Validation	Accuracy Without Cross Validation	Accuracy Without Cross Validation	Accuracy Without Cross Validation
Decision Tree	59	51	57	73
Logistic Regression	58	63	56	78
Support Vector Machine	62	66	60	79
Naïve Bayes	55	60	56	71
K-Nearest Neighbor	62	58	64	67

We can see that Logistic Regressor and Support Vector Machine can be a go-to solutions to our problem and we can tune it to make them more accurate.

Tune the Logistic Regression Model

We will be using GridSearchCV to fine tune our Logistic Regression Model as below:

Parameters for tuning:

```
# Create regularization penalty space
penalty = ['l1', 'l2']

# Create regularization hyperparameter space
C = np.logspace(0, 4, 100)

# Create hyperparameter options for Logistic Regressor
lr_grid_params = dict(C=C, penalty=penalty)
```

Fitting the model with parameters for tuning:

```
lr_clf = GridSearchCV(lr_model, lr_grid_params, refit=True, verbose=2)
```

```
lr_clf.fit(x_train, y_train)
```

Score obtained by best fitted model:

```
lr_clf.best_score_
```

```
0.7809541331127946
```

Hence, with model tuning, we can increase the performance of the existing model, as in this case, after tuning, the accuracy was increase from **58% to 78%.**

Conclusion

1. Handling Data Imbalance can help increasing the performance of classification problems.
2. Logistic Regression Model and Support Vector Machine can be the go-to algorithms to solve our problem.
3. On tuning the model, we are able to increase the accuracy from 58% to 78%. We can do the same exercise for Support Vector Machine Model as well.