Natural Language Processing Project

Disaster Tweets Analysis

By Omran Fallatah - Samer Atwi- Abdalla Alnujaidy

Objective

In this project we'll be working on a dataset from Kaggle on Disaster tweets

The goal of this project is to develop a model that will predict whether a tweet is a real disaster tweet or not.

Tools & libraries

- Numpy
- Pandas
- Nltk
- Sklearn
- Matplotlib
- Seaborn

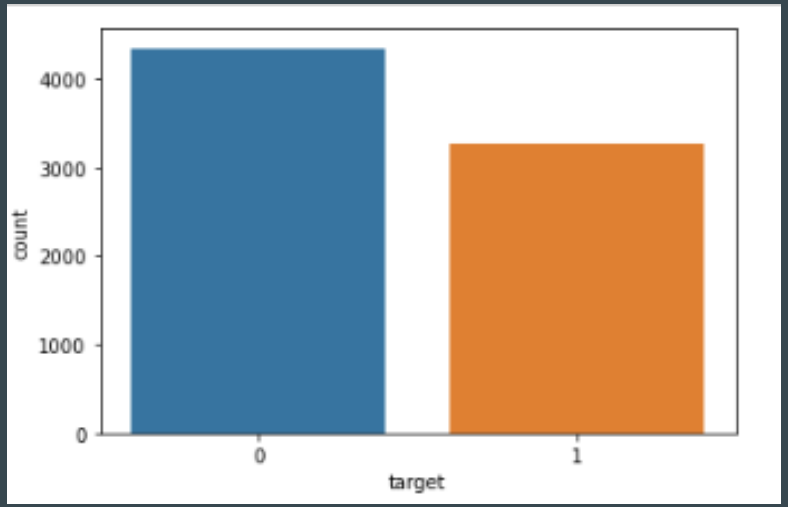
Dataset

This data set contain 7613 tweets

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all	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 being notified by officers. No other evacuation or shelter in place orders are expected	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation orders in California	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as smoke from #wildfires pours into a school	1

Balanced Data

There is no class imbalance in the distribution of target variable



Data Cleaning

* Removing urls from tweet
* Removing punctuations
* Removing stopwords
* Removing emoji
* Lemmatization

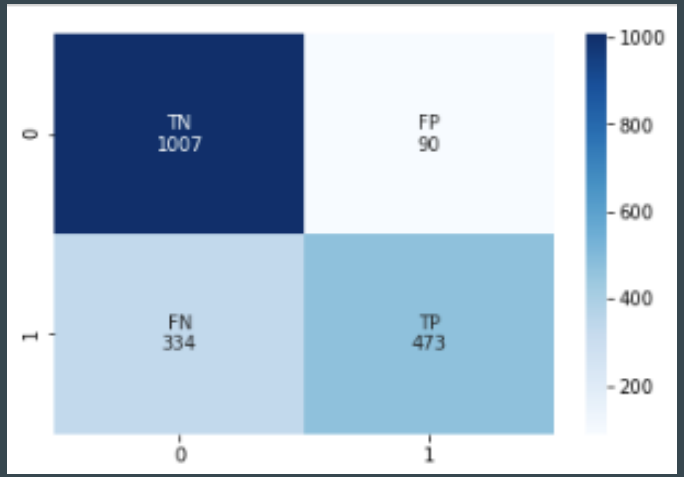
After cleaning

	text	target
	deed reason earthquake may allah forgive u	1
	forest fire near la ronge sask canada	1
	resident asked shelter place notified officer evacuation shelter place order expected	1
	people receive wildfire evacuation order california	1
	got sent photo ruby alaska smoke wildfire pours school	1

Random Forest classifier

	precision	recall	f1-score	support
0	0.92	0.75	0.83	1341
1	0.59	0.84	0.69	563
accuracy			0.78	1904

Confusion Metrix



Thank You .. Any Questions?