Natural Language Processing with Disaster Tweets

Goal

- Build a machine learning model that predicts which Tweets are about real disasters and which one’s aren’t.

Background

- According to omnicoreagency ,On average 500 million tweets are shared everyday. That means 6000 tweets per second 350,00 tweets per minute and around 200 Billion tweets every year.
- This shows us that many people use twitter.
- Many people use twitter as a source of News.

Natural Language Processing (NLP)

- Natural Language Processing (NLP) is **a subfield of artificial intelligence (AI). It helps machines process and understand the human language so that they can automatically perform repetitive tasks**. Examples include machine translation, summarization, ticket classification, and spell check

Procedure

- Data Exploration
- Data Visualization
- Training Model

Data Exploration

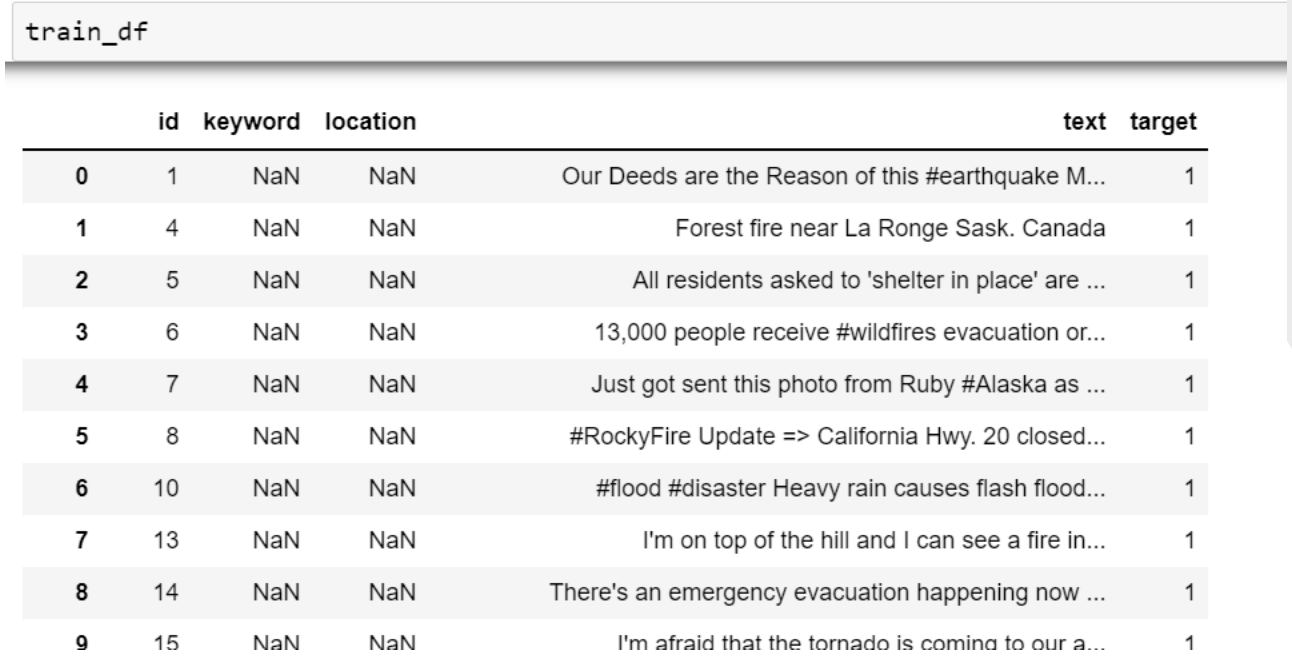
Data

Type: CSV file (train csv file and test csv file)

Input: CSV file of features, output: target ==> 1 or 0.

Size: How much data? 10,876

The train data contains ID , Keyword ,Location ,Text ,Target

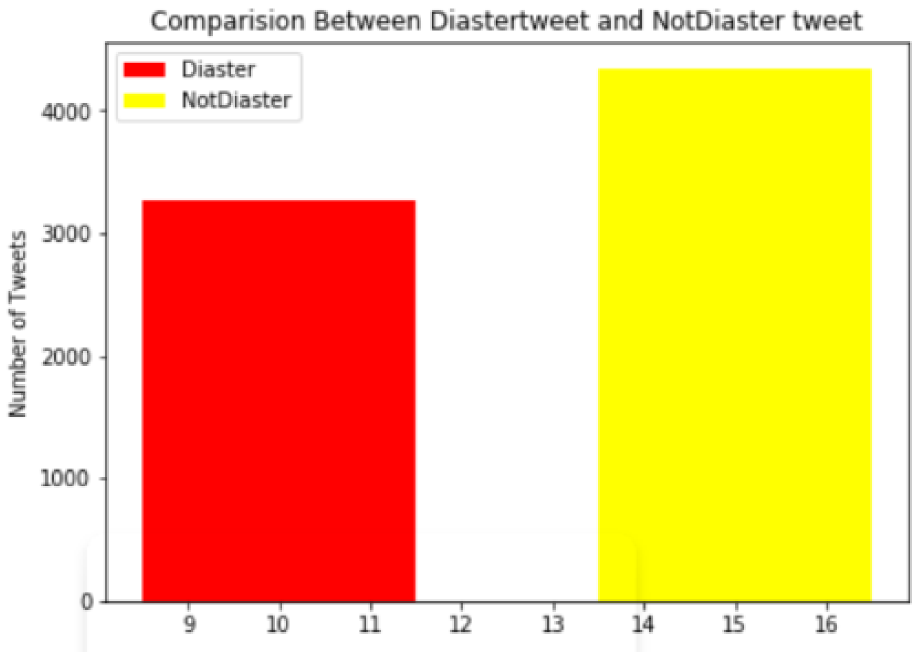


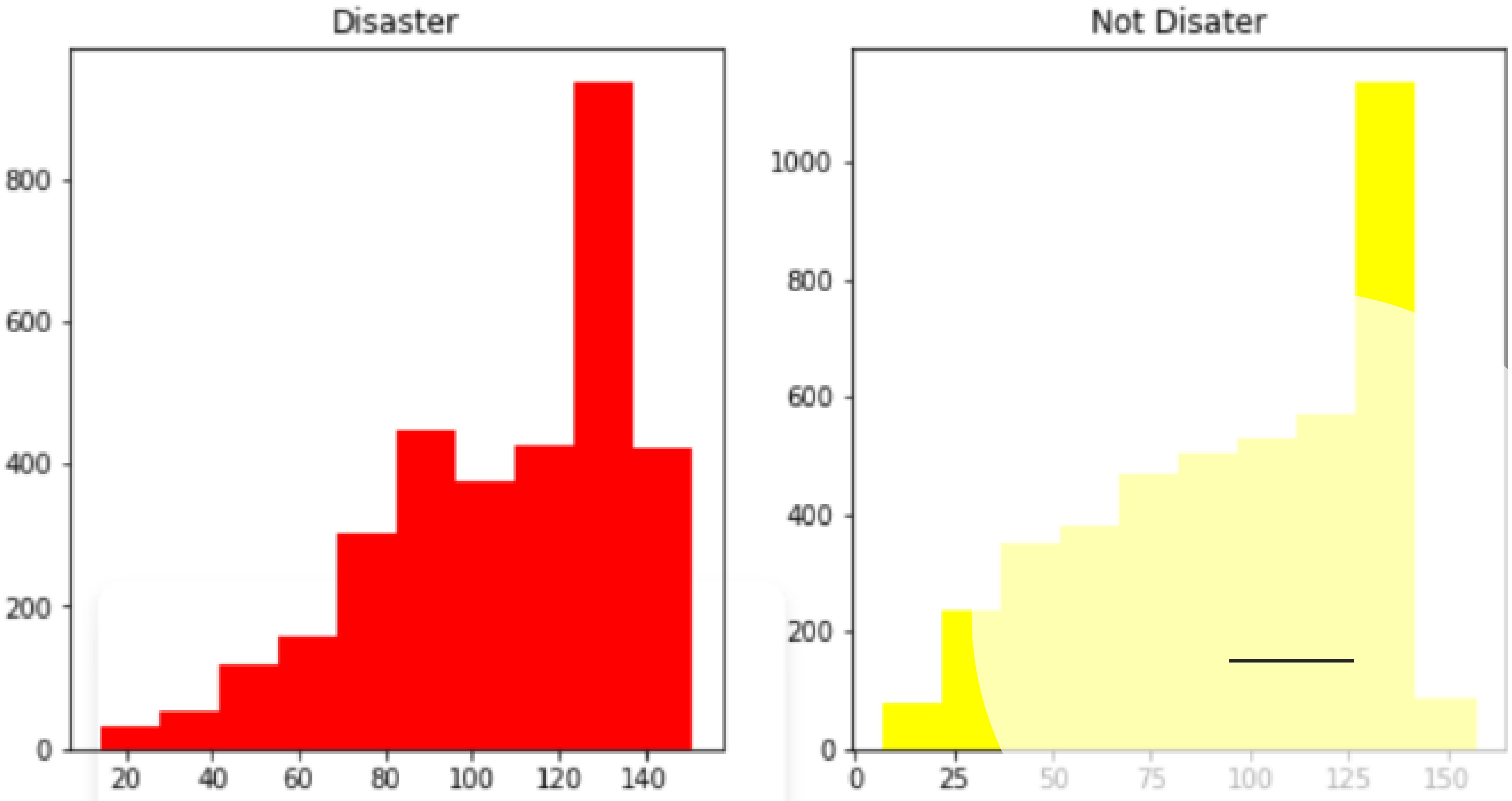
```
train_df
```

	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
5	8	NaN	NaN	#RockyFire Update => California Hwy. 20 closed...	1
6	10	NaN	NaN	#flood #disaster Heavy rain causes flash flood...	1
7	13	NaN	NaN	I'm on top of the hill and I can see a fire in...	1
8	14	NaN	NaN	There's an emergency evacuation happening now ...	1
9	15	NaN	NaN	I'm afraid that the tornado is coming to our a...	1

visualization

- **The first bar chart shows how many of the tweets are disaster or not and the second one shows the length of the tweets**




Disaster
0
200
400
600
800
20
40
60
80
100
120
140
Not Disater
0
200
400
600
800
1000
0
25
50
75
100
125
150

Data Cleaning

- **Data cleaning**
- **Remove URL from the**
- **Remove special characteristics**
- **Remove HTML Tags from the texts**

Build Model

- change the text to vectors
- It was trained using linear model

Result and Conclusion

- Used f1 score to see the accuracy the average was equal to 0.52632.
- The f1 score was not high which indicates linear model might not be the best model for this type of data