Classifying Real and Fake Disaster Tweets

Final Capstone
By Malak Mosly

Problem Statement

- ❑ It is critical to respond to disasters (fires, shootings, etc.) in a timely manner.
- ❑ Social media sites such as Twitter are a very effective way to report disasters in real-time.
- ❑ The goal of this project is to create a model that can quickly identify a disaster tweet.

Who can benefit from this model?

❏ News outlets
❏ Disaster relief organizations
❏ Grassroots networks
❏ Politicians
❏ Law enforcement agencies
❏ Fire departments, hospitals, etc.
❏ The general public

Dataset (Kaggle Competitions)

- Consisted of 7,613 rows of tweets.
- Columns:
  - ID
  - Keyword
  - Location
  - Text
  - Target

Project Steps

❏ Data wrangling and cleaning

❏ Exploratory data analysis

❏ Preprocessing and training

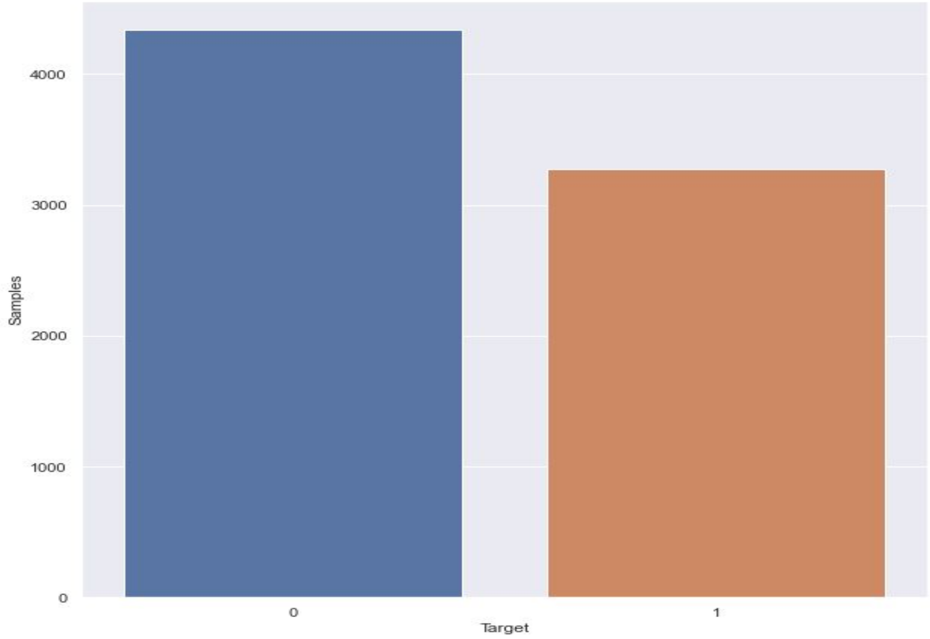
❏ Modeling

Data wrangling and cleaning

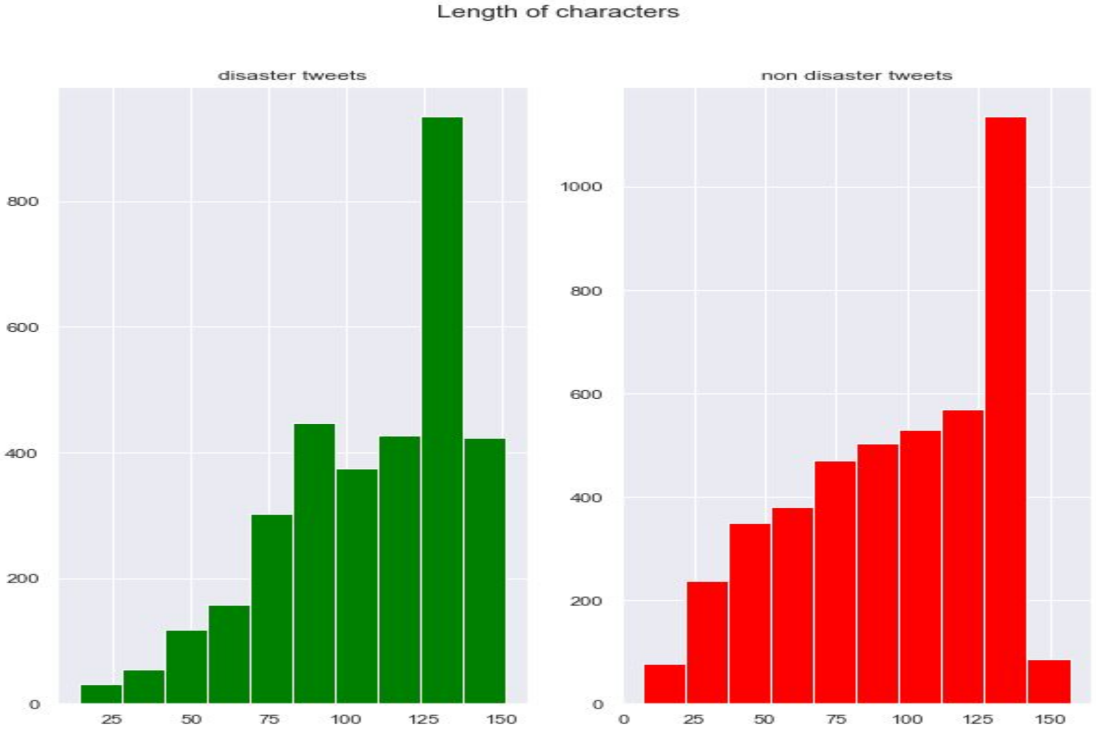
- ❑ Dataset was explored (shape, info, value counts)
- ❑ Text cleaning was the main focus of this step:
  - ❑ Punctuation removal
  - ❑ Emoji removal
  - ❑ Html link removal
  - ❑ Numbers removed
  - ❑ Square brackets removed
  - ❑ All text made lowercase
  - ❑ Stop word removal

Exploratory data analysis (I)

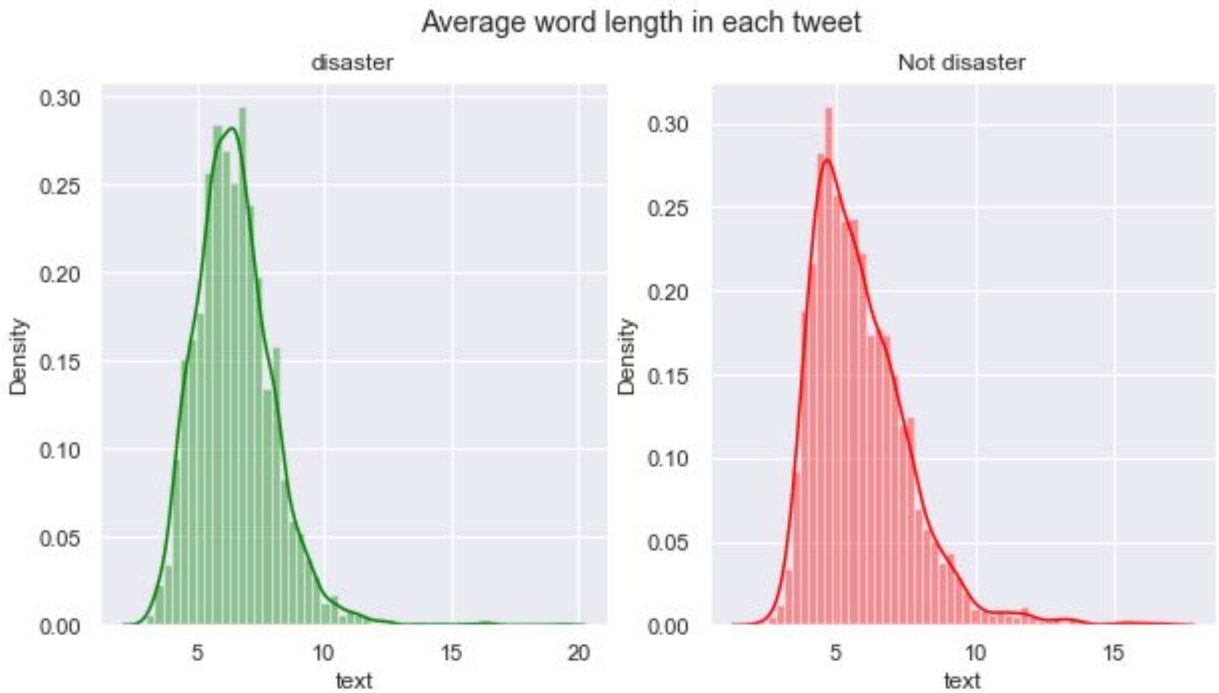
Distributions of target column



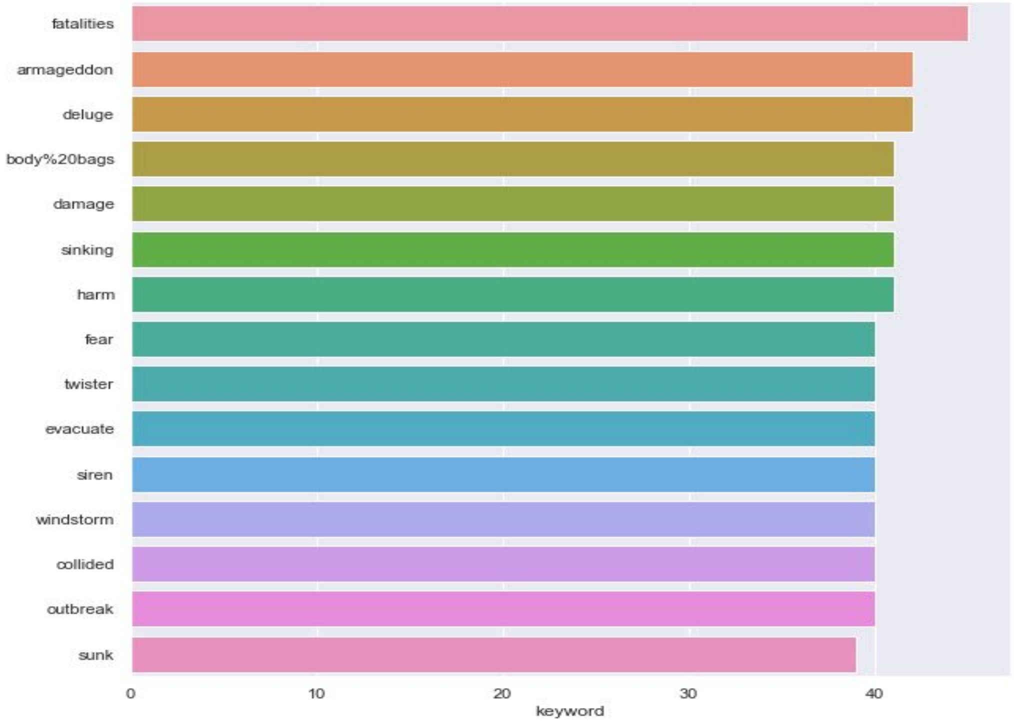
Exploratory data analysis (II)



Exploratory data analysis (III)



Exploratory data analysis (IV)



Preprocessing and training

- ❑ Defining X and y:
  - ❑ X ⟶ text column (predictor variable)
  - ❑ Y ⟶ target column (target variable, binary)
- ❑ Used CountVectorizer() on text column.
- ❑ Trian_test_split used to create training and testing datasets.
- ❑ TF-IDF transformation on text data.

Modeling (I)

- ❑ Three supervised models chosen:
    - ❑ Naive Bayes
    - ❑ Support-vector machines (SVM)
    - ❑ Logistic Regression
- ❑ Models were cross-validated and hyperparameter tuning done using GridSearchCV.

Modeling (II)

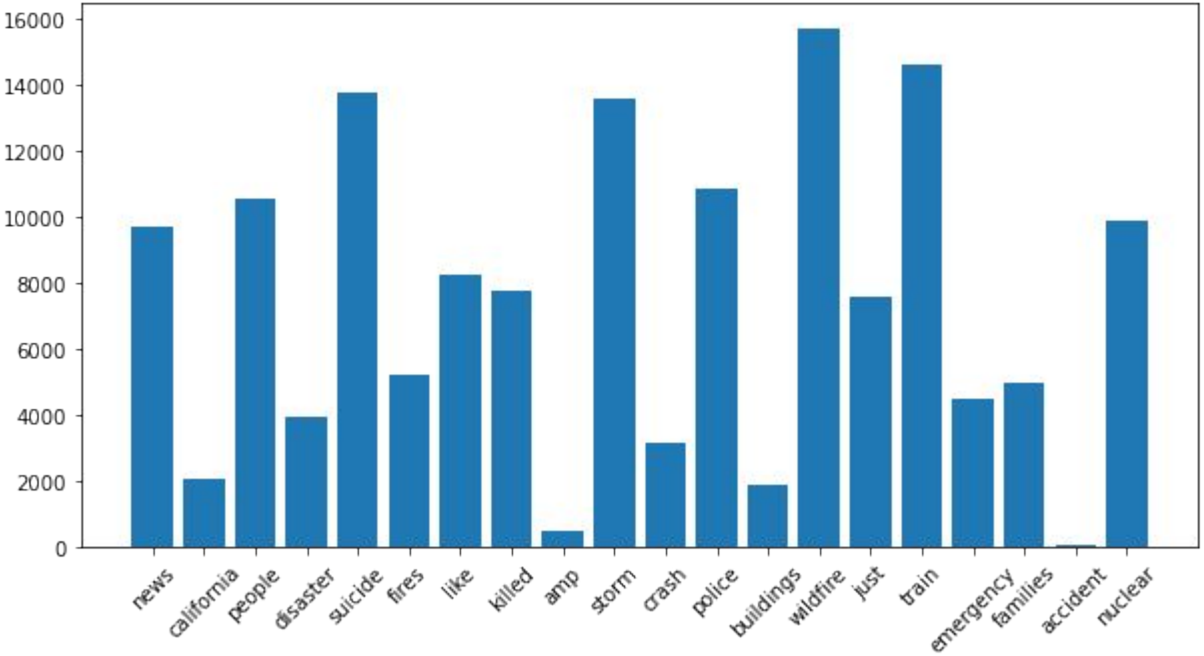
- ❑ Performance assessment metrics:
  - ❑ Classification report
  - ❑ Confusion matrix
  - ❑ Cross validation scores
  - ❑ Balanced accuracy scores
- ❑ Best performing model was Naive Bayes, with 80.22% testing accuracy.

Modeling (III)

Summary of performances:

Model	Testing Accuracy
Naive Bayes	80.22%
Logistic Regression	79.78%
SVM	78.95%

Most important features



Future research recommendations

❑ Expand model to note location and time of disaster

❑ Expand to all types of social media

❑ Model can go beyond fake/real disaster classification