Project Cymopolea: Natural Disaster detection from tweets

Author: Terence Yep

Image source: www.alrasub.com

Introduction

- On average, approximately 60,000 people globally died from natural disasters each year

- Low frequency, high impact events such as earthquakes cause significant but preventable loss of human life

- Why Twitter?
  - real-time microblogging platform
  - 186 million daily active users as of 2020
  - 150 million users worldwide, 36 million from US

Why Bother?

- Total $210 billion dollars in losses world-wide in 2020
  - Up 26% since 2019
  -

- Preventable loss of human life
  - Last year's natural disasters claimed approximately 8,200 lives

- Fraudulent insurance claims

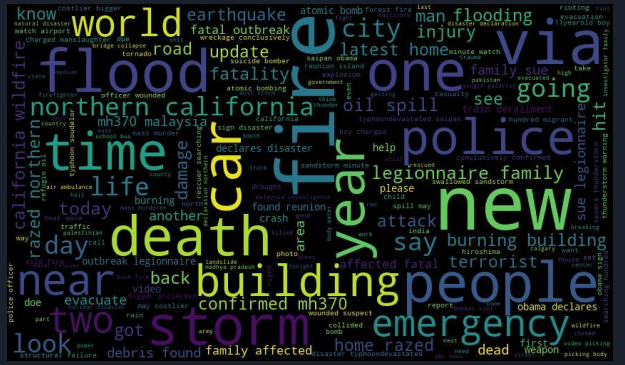
- Twitter is being used over 50 countries

Objectives

- Detect natural disaster tweets

- Detect natural disaster images from attached images

Data sources and Methodology



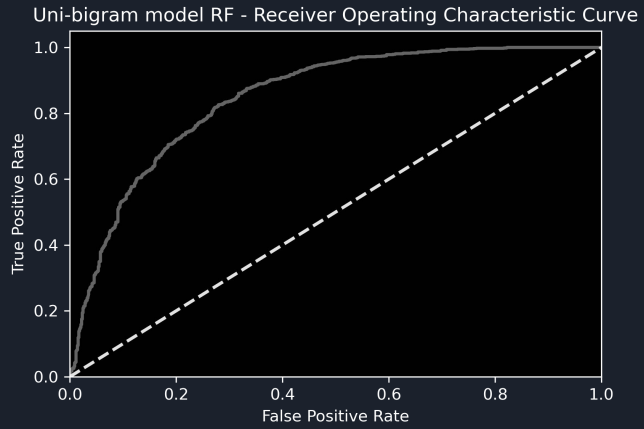
- Pre-labeled tweet dataset from Kaggle

- Python
  - NLTK to tokenize and lemmatize tweets
  - Pandas for handling data in dataframes and Numpy for transforming data.
  - wordcloud
  - Sci-kit learn:
    - Bag-of-Words and TF-IDF to vectorize corpus
    - Supervised classifiers: Random Forest, Logistic Regression, Gradient Boost

Results

- Best model based on accuracy
  - RandomForest classifier
  - ~78% accuracy
  - AUC ~0.85

- Worst model
  - Unigram BoW with Gradient-Boost classifier
  - ~65%
  - AUC ~0.85



Limitations

- Class imbalance

- Bias introduced during pre-processing
  - Suicide, war were excluded
  - Locations and keywords were excluded

Future Directions

- Include location and hashtags/keywords in modeling
- Improve accuracy of model (95% target)
- Word embeddings and deep learning (LSTM)
- Mobile app that tracks and sends email to subscriber

Fin