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

Objectives

- Detect natural disaster tweets

- Detect natural disaster images from attached images

Data sources and Methodology part 1

- Business case
  - Real-time information currently lacking/under-serviced

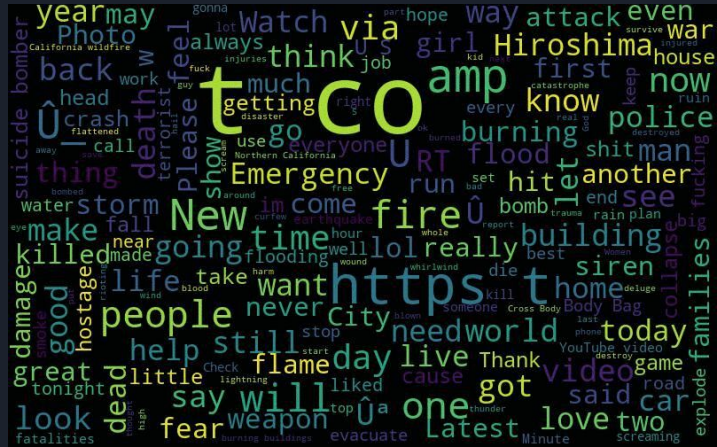
- Gather data
  - Pre-labeled tweet dataset from Kaggle
  - Pre-labeled images dataset from Kaggle
  - Generate dataframe

Data sources and Methodology part 2

- Preprocess data
  - Removed symbols, hyperlinks, punctuations, stop words using custom regex patterns
  - Lemmatized words
  - Changed target labels for tweets using non-natural disasters words
  - Tweet sentiment

- Visualization
  - Barplots (count-based features)
  - wordcloud

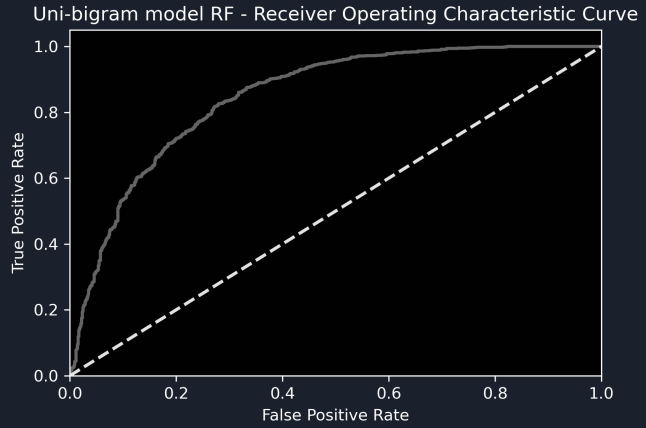
- Modeling
  - Logistic Regression
  - Ensemble (RandomForest, GradientBoost)
  - Model optimization : Random search Cross validation



Results

- Best model based on accuracy
  - RandomForest classifier
  - ~78% accuracy with an FPR ~6.47%
  - AUC ~0.85
  - Specificity of 94%
  - Sensitivity of 55%

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



Limitations

- Class imbalance
  - More non-disaster vs disaster tweets
  - Responsible for skew in sensitivity and specificity

- 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