Writeup.md

Natural Language Processing on Disaster Tweets

Abdulla Alnujaidy

Samer Atwi

Omran Fallatah

Abstract

The goal of this project was to use natural language processing to analyze disaster tweets and to be able to predict which group would future tweets fall into. Which in the future would help disaster relief organizations and news agencies.

Design

The data is downloaded from Kaggle and is used to build our model that'll be able to predict future disaster tweets.

Data

Disaster tweets contains 7613 tweets (data points) and a label(0,1) as not disaster and disaster respectively. Also, we have a test dataset of around 3000 unlabeled tweets.

Algorithms

Data manipulation and cleaning.

- Exploratory data analysis
  - Dropped unnecessary columns.
  - Dropped all duplicates data points.
  - Dropped all rows containing null or NaN values.
- Data Cleaning
  - Removed URLs
  - Removed punctuations (including @ for twitter tagging)
  - Removed Stop words
- Normalization
- Stemming
- Lemmatization

Model Evaluation and Selection

After modeling and applying regularization we found the following results:

Algorithm	Accuracy
Random Forest	0.78

Tools

- Numpy, Pandas and Regular Expression for data cleaning and manipulation.
- Seaborn for plotting.
- NLTK for Natural Language Processing (Stop words, Vectorization, Lemmatization etc...)
- Scikit-learn for modeling.

Communication

In addition to the slides and the visuals included in the presentation, we will submit our code and proposal.