10-315 Project Proposal

Project Goal - Classification of Tweets about natural disasters

Possible Dataset - Dataset on tweets about natural disasters

Algorithm - Naive Bayes

Dataset Description and Link:

* <https://www.kaggle.com/datasets/vstepanenko/disaster-tweets>

Briefly describe your project goals and why it is worth while doing.\*

Our goal for this project is to create an accurate classifier to distinguish between real and fake disaster tweets. This is motivated by the tendencies of netizens to latch on to “newsworthy” and provocative content without much investigation. An accurate classifier would help social media companies avoid promoting harmful misinformation and promote useful messages.

Sample Data:

* <https://www.kaggle.com/datasets/thedevastator/historic-volcanic-activity?select=historic_eruptive_activity_tongariro.csv>
* <https://www.kaggle.com/datasets/martincontreras/volcanic-eruptions-dataset-all-to-2020>
* Smithsonian: <https://www.kaggle.com/datasets/jessemostipak/volcano-eruptions>
* <https://www.kaggle.com/datasets/vstepanenko/disaster-tweets>

Links for Code Notebooks (already created)

* Effective EDA and BERT in NLP
  + <https://www.kaggle.com/code/masatoshikato/effective-eda-bert-in-nlp-bert-baseline>
* Disaster Tweet Classification with EDA and CNN
  + <https://www.kaggle.com/code/pinsdev/desaster-tweet-classification-with-eda-and-cnn/data>
* Article about classifying disaster tweets:
  + <https://medium.com/real-or-not-nlp-with-disaster-tweets/real-or-not-nlp-with-disaster-tweets-a-data-science-capstone-project-fafa6c35c16f>