Real or Not? Natural Language Processing with Disaster Tweets

Spring 2020 - INFSCI 2440 - Artificial Intelligence
University of Pittsburgh

Neha Chanu (knc36),
Dhaval Sonani (dvs15),
Tianlin Zhao 'Anthony' (tiz52)

TABLE OF CONTENTS

01 PROBLEM & GOAL OF PROJECT

- **Problem statement** is that with the enormity of tweets being sent on Twitter at any point in time, it is hard manually decipher which tweet is about an emergency situation.
- Because of this, more agencies are interested in programmatically monitoring social media platforms in order to inform stakeholders and take steps towards emergency situation relief.
- Agencies such as disaster relief organizations and news agencies are interested to identify emergencies as soon as possible by monitoring social media
- Thus, the **users of the system** would be disaster relief organizations, news agencies, first responders such as paramedics, firefighters and police personnel
- **Goal of this project** to analyze tweets to **predict which tweets are about real disasters** and which ones are not.
- This project is framed as a **binary classification** supervised machine learning and natural language processing task.

02 SOLUTION : Overview of System Architecture

System architecture:

- Two versions of the system were created - Version A and Version D.
- While both versions differ when it comes to the specifics, the underlying structure is the same:

Figure A: Underlying Structure and Flow of Both Versions

03 SOLUTION : In Depth Look

Version A

- Data Cleaning & Feature Engineering:
  - ➢ Removed URLs, emojis, html tags, numbers, punctuation, stop words
  - ➢ Lemmatization
  - ➢ Combine two columns, 'keyword' and 'tweet text' into one column
  - ➢ GoogleNews-vectors-negative300 for word vectorization
- Machine Learning Models Used:
  - ➢ Adaboost
  - ➢ Ridge
  - ➢ K Nearest Neighbors (K-nn)
  - ➢ Support Vector Machine (SVM) with linear kernel
- Hyperparameter tuning:
  - ➢ Manual tuning

Version D

- Data cleaning & Feature Engineering:
  - ➢ Removed URLs, emojis, html tags, punctuation, stop words
  - ➢ Stemming
  - ➢ Just used 'tweet text' column
  - ➢ Used countVectorizer for bag of words
- Machine Learning Models Used:
  - ➢ Logistic Regression
  - ➢ Decision Tree Classifier
  - ➢ Gradient Boosting Model
  - ➢ K Nearest Neighbors (K-nn)
  - ➢ Support Vector Machine (SVM)
- Hyperparameter tuning:
  - ➢ Using model selection's Gridsearch Cross Validation

04 RESULTS : Model Performance & Evaluation

Best Models

	Model	Test Accuracy	Test F1 Score
Version A	Adaboost	0.74	0.69
	Ridge	0.78	0.72
	K-nn	0.77	0.71
	SVM with linear kernel	0.77	0.72
Version D	Logistic Regression	0.78	0.71
	Decision Tree Classifier	0.74	0.67
	Gradient Boosting Model	0.74	0.59
	K-nn	0.72	0.57
	SVM	0.61	0.12

05 PROJECT LIMITATIONS & SUMMARY

Limitations:

- Generalizability:
  Since the dataset is relatively small (7613 tweets) and the data set was curated, the performance of models in either versions on real world data set may not lead to as high performance results.

- Balance of Datasets:
  This kaggle dataset was almost balanced among tweets that were about a disaster and those that were not. The proportion in real world data sets is likely to differ (figure B).

- Variability within tweets:
  It is likely that tweets in real world data sets have more variability than those in this kaggle dataset (figure C, D).

- Computing Resources:
  Our laptops didn't have enough computation power to run more sophisticated models such as Google Bert. More sophisticated models may have a better performance but at the cost of interpretability.

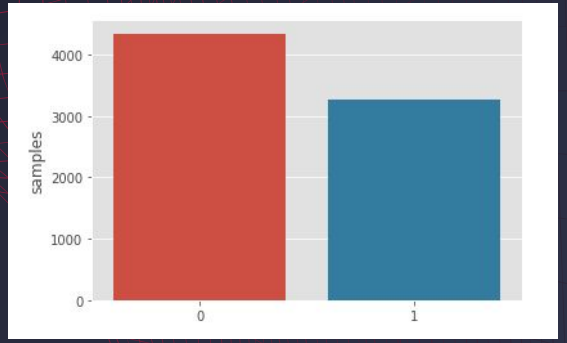


Figure B: Dataset is roughly balanced

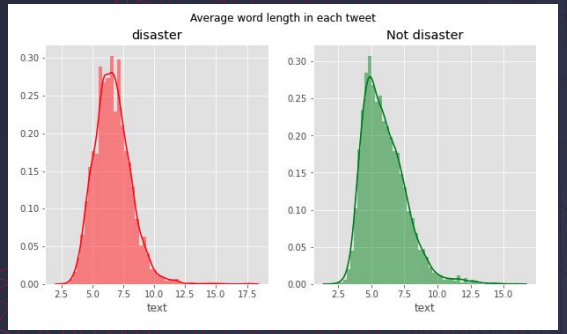


Figure C: Variability within tweets - roughly same avg word length

05 PROJECT LIMITATIONS & SUMMARY

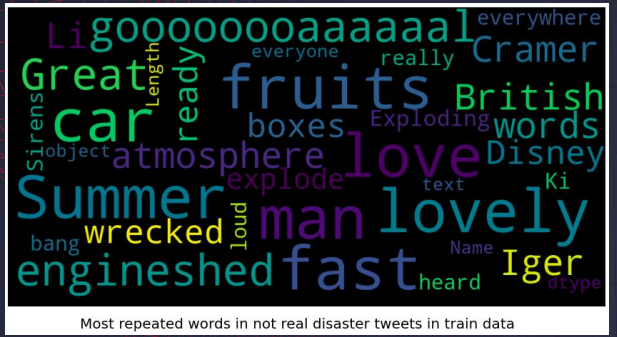
Limitations:

- Generalizability:
  Since the dataset is relatively small (7613 tweets) and the data set was curated, the performance of models in either versions on real world data set may not lead to as high performance results.

- Balance of Datasets:
  This kaggle dataset was almost balanced among tweets that were about a disaster and those that were not. The proportion in real world data sets is likely to differ (figure B).

- Variability within tweets:
  It is likely that tweets in real world data sets have more variability than those in this kaggle dataset (figure C, D).

- Computing Resources:
  Our laptops didn't have enough computation power to run more sophisticated models such as Google Bert. More sophisticated models may have a better performance but at the cost of interpretability.



Most repeated words in not real disaster tweets in train data

Figure D: Frequency of words in tweets NOT about a real disaster.

Most repeated words in real disaster tweets in train data

Figure D: Frequency of words in tweets about a real disaster.

05 PROJECT LIMITATIONS & SUMMARY

Summary:

- Our project was based on a kaggle competition using curated twitter data.
- The **problem statement** is that with the enormity of tweets being sent at any point in time, it is hard manually decipher which tweet is about an emergency situation.
- **Goal of this project** to analyze tweets to **predict which tweets are about real disasters** and which ones are not.
- This project is framed as a **binary classification** supervised machine learning and natural language processing task.

06 SYSTEM DEMO

Order of presenters:

- Version D presented by Dhaval Sonani
- Version A presented by Tianlin Zhao