SJSU SAN JOSÉ STATE UNIVERSITY

Individual Project Report

Natural Language Processing With Disaster Tweets

Abhiteja Mandava, TEAM 10

1 Introduction

Social Network Services(SNS) have become not only an important source of emergency information during disasters but also a medium for expressing immediate responses of warning, evacuation or rescue. Because smartphones are so common, people can use them to broadcast an emergency in real time. As a result, more organizations (such as disaster relief organizations and news companies) are interested in programmatically monitoring tweets, however it's not always clear whether a person's statements are announcing a calamity. By analyzing context, we can utilize this study to track, monitor and predict disasters from the real time data and this study would help make prediction models.

1.1 Problem Description

The problem can be viewed as a binary classification problem and this project's goal is to figure out how to tell which tweets are about "genuine disasters" and which aren't. This project will involve experimentation on various machine learning models that will predict which tweets are about "actual disasters" and which aren't.

2 My Contribution

To begin with, I was involved in collaborative research on project scouting. Once the project was finalized, I was part of the planning and segregation of tasks. Within the scope of my knowledge, we have taken a comprehensive approach to solving the problem associated with the project. We wanted to explore the impact of data sets constructed using different word vectorization methods (Countvectorizer, TF-IDF, Word2Vec, and Word2Vec with Principal Component Analysis) on multiple models (Logistic regression, SVM, Decision Tree, Random Forest, XGBoost, LSTM Glove, Glove, LSTM with Word2Vec, etc.). We then divided tasks which included exploratory analysis, data pre-processing, feature selection, feature extraction, model training, etc.

2.1 Data Pre-processing

I have contributed to data pre-processing by performing data cleaning activities like removing special characters/texts, unwanted URL tags, punctuation etc associated with the data set.

git commit: 449bf47c10ca40e37cb475f10420966231a09e40
Files: 01RandomForestXGboostAbhiteja.py

Data cleansing by removing missing values.

git commit: 073a360c603d1603fbdd534c21d267470c44ce78
Files: 01RandomForestXGboostAbhiteja.py

Exploratory Analysis by plotting the size of both positive(disastrous) and negative(non-disastrous) tweets.

git commit:	57cba0592134273a7187862717cf61f9c7dd4396
Files:	01RandomForestXGboostAbhiteja.py
Files:	01RandomForestXGboostAbhiteja.ipynb

Observing Stop words.

git commit:	c8c60e70af34fcd5ac5a4e96ef228997b1a58318
Files:	01RandomForestXGboostAbhiteja.py
Files:	01RandomForestXGboostAbhiteja.ipynb

2.2 Feature Extraction

I have contributed to feature extraction of the tweets which involved word embedding concepts like CountVectorizer, Tf-Idf, Word2vec.

Feature extraction.

git commit:	e199160c8e8a7bfb8903641037d21e6267d53d00
git commit:	fec1202efd6a48f2683a4f19d62a5c8c13b90796
git commit:	81d86622cc0f910b0826b74c938aad988685a9ac
Files:	01RandomForestXGboostAbhiteja.py
Files:	01RandomForestXGboostAbhiteja.ipynb

2.3 Model : Random Forest

I have contributed to training the Random Forest model for different feature vector sets - CountVectorizer, Tf-Idf, Word2vec.

Random Forest.

git commit:	95d915faa961baa80a3d8b9f25ec06bbdbf31dce
git commit:	28deb7a19cc3e9113b0f1710bcff90476f28964f
Files:	01RandomForestXGboostAbhiteja.py
Files:	01RandomForestXGboostAbhiteja.ipynb

2.4 Model : XGBoost

I have contributed to training the XGBoost model for different feature vector sets - CountVectorizer, Tf-Idf, Word2vec.

XGBoost.

git commit:	a4140b6f22914dc596c7590fd6a3a4e3be10fa38
git commit:	ba1a61b97d34a3cacf236efe040b9d6fb4b2570f
git commit:	81d86622cc0f910b0826b74c938aad988685a9ac
Files:	01RandomForestXGboostAbhiteja.py
Files:	01RandomForestXGboostAbhiteja.ipynb

2.5 Project Report/Paper.md

I have contributed to writing the paper/report. I gave inputs for all the sections of the report

Report.

git commit:	5c5a01c0f294974bef7d1cbd3f0fafec2a53c209
git commit:	3963dedfc5e1ac9f3541b9cc0447daf2cf2966fa
git commit:	997345189537823c3e2aa48c07fa821428a0eb8b
git commit:	19b9e5b58c6f7e20a5ee18d6f85dcc2b807d813f
git commit:	64049ef8fd827607209a53e08af738657a468782
git commit:	efa85dbaa3e06254783a433d4539b91958dc74d6
git commit:	ab0a0f061f6fcd60402501793c0918d835e9b589
git commit:	1c2180cac0b6d1d8dd6d4b8b91b6f60fc6178b73
Files:	paper.md

3 Teammates Contribution

3.1 Yoonjung Choi

- Yoonjung participated in project selection, segregation of tasks.
- Yoonjung contributed to all sections in data pre-processing and feature extraction.
- Yoonjung trained the following models on all feature vector sets: Logistic Regression, Support vector Machine, Decision tree, Ensemble.
- Yoonjung solely integrated all sections of the code.
- Yoonjung contributed to report work.

3.2 Sakruthi Avirineni

- Sakruthi participated in project selection, segregation of tasks.
- Sakruthi contributed to some sections in data pre-processing and feature extraction like observing and removing stop words.
- Sakruthi trained the LSTM model on Word2Vec feature vector set.
- Sakruthi solely developed analysis and code of BERT classifier.
- Sakruthi contributed to report work.

3.3 Ishaan Bhalla

- Ishaan participated in project selection, segregation of tasks.
- Ishaan contributed to data pre-processing required for LSTM.
- Ishaan trained the LSTM model on different feature vector sets.
- Ishaan solely developed analysis and code of LSTM classifier.
- Ishaan contributed to report work.