Disaster Tweet Classification using Natural language Processing

Varsha Reddy Mandadi
Utah State University, CS-5665- Class project

Problem:

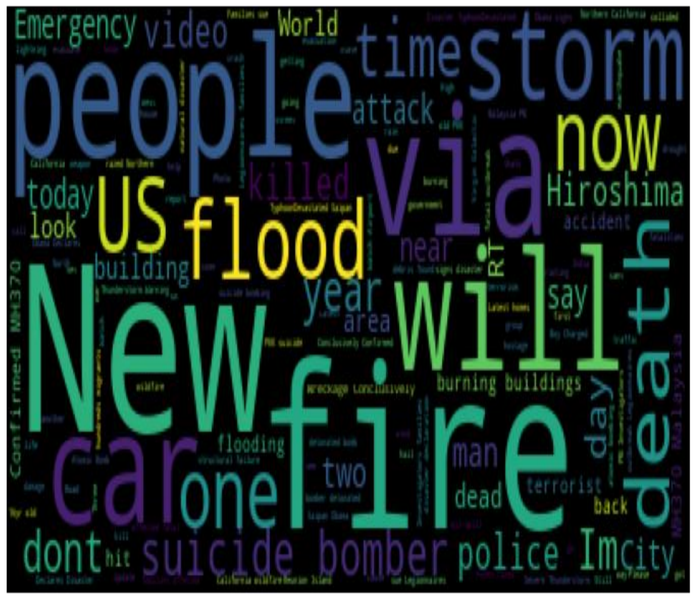
To identify whether a person's words are actually announcing a disaster or not

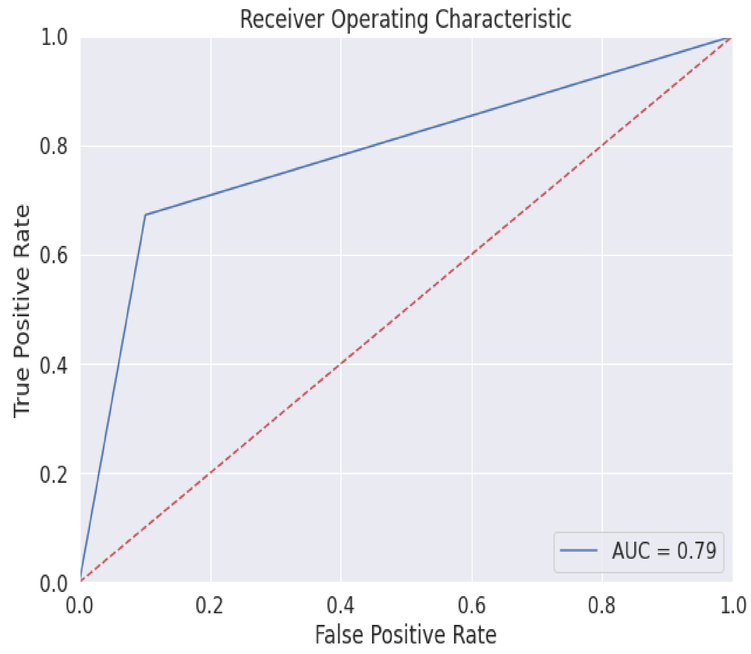
Data:

Sample tweets:
- I can see a fire in the woods...
- There's an emergency evacuation happening now
- What a wonderful day!

Analysis and graphs

Word cloud from disaster tweets





Methods /Approach

Two different approaches used in this project are:

Bag-of-words Method:

One hot encode text data and then apply traditional machine learning models (used Logistic Regression and XGBoost)

Word Embedding Method:

Generate word embedding from the training dataset

Using pretrained word embeddings

Apply Logistic Regression & Multi-Layer Perceptron on word vectors to classify

Results

Logistic Regression **f1_score: 0.75**
Logistic Regression **Accuracy: 0.81**
Logistic Regression Precision: 0.84
Logistic Regression **Recall: 0.69**

Using Logistic Regression, I was able to identify 69% of disaster tweets with 81% accuracy

Conclusions

This is a good starting point using vanilla text classification methods.

Implementing further text cleaning techniques like spelling correction and employing RNN may improve accuracy of the process.

Source code

Git hub link: https://github.com/var2019/Data-science-project--Varsha

References

https://mlexplained.com/2018/02/08/a-comprehensive-tutorial-to-torchtext/
https://gist.github.com/slowkow/7a7f61f495e3dbb7e3d767f97bd7304b
https://www.dataquest.io/blog/tutorial-text-classification-in-python-using-spacy/
https://edumunozsala.github.io/BlogEms/jupyter/nlp/classification/embeddings/python/2020/08/15/Intro_NLP_WordEmbeddings_Classification.html