ANALYSIS OF DISASTER TWEETS' AUTHENTICITY

PROJECT PROPOSAL - GROUP 39

Srilekha Gudipati, Sasank Marabattula, Sukruthi Modem, Srihitha Reddy Kaalam
Department of Computer Science, North Carolina State University, Raleigh, NC 27695
sngudipa, smaraba, smodem, skaalam

1 Data Set

We are using the dataset created by the company figure-eight and originally shared on their 'Data For Everyone' website (https://appen.com/pre-labeled-datasets/). The dataset consists 10,900 hand classified tweets with 11 columns.

2 Project Idea

Twitter is a major communication channel and the ubiquitousness of smartphones enables people to announce an emergency they're observing in real-time. Because of this, more agencies are interested in tools that monitor Twitter. This is of gravity only if there is no ambiguity in the tools' results.

With this particular disaster-based tweets dataset, it's not always clear whether a person's words are actually announcing a disaster, so we want to build a machine learning model that can predict whether a tweet about some disaster is real or not. Our approach is to use some of the prominent fake-news detection algorithms to tackle this problem. The goal is to perform comparative analysis of various machine learning classifiers and find the model that fits best.

3 Software to Write

We will be writing code for data cleaning, data pre-processing, Exploratory Data Analysis, feature engineering, training and prediction using Support Vector Machine, Logistic Regression, Xgboost and Neural Network models. Then, we compare the above models to determine which one is better.

4 Relevant Papers

References

[1] Oluwaseun Ajao, Deepayan Bhowmik, Shahrzad Zargari. IEEE Conference on Fake News Identification on Twitter with Hybrid CNN and RNN Models, 2018.

[2] Julio C. S. Reis, Andre Correia, Fabrıcio Murai, Adriano Veloso, and Fabrıcio Benevenuto. IEEE Publication on Supervised Learning for Fake News Detection, 2019

[3] Zaitul Iradah Mahid, Selvakumar Manickam and Shankar Karuppayah. IEEE Conference on Fake News on Social Media: Brief Review on Detection Techniques, 2018.

5 Division of Work

Exploratory data analysis, data cleaning, pre-processing and principal component analysis will be done by all team members. Srilekha and Sasank will work on modelling the data initially with a Support Vector Machine and Neural Network model later on, while Srihitha and Sukruthi will work on modelling the data initially with a Logistic Regression one and Xgboost model later.

6 Midterm Milestone

By midterm checkpoint, our primary target is to perform data selection and pre-processing steps. As our aim is to build various machine learning models, we plan on building at least one of them if feasible. This helps in establishing a benchmark for other models for comparison.