Real or Not? NLP with Disaster Tweets

ELEC-E5550 - Statistical Natural Language Processing

Group 1: Anand Umashankar 765837, Swati Choudhary 795920

Email : firstname.lastname@aalto.fi

21 January 2019

1 Introduction

The aim of the project is to use natural language processing techniques to detect if the tweets are about a real disaster or not. The dataset is obtained from Kaggle. The dataset has the following attributes : tweet text, keywords and location. The data is then subjected to pre-processing techniques like word2vec, lemmatization, removal of stop words, etc using Python NLTK library. We plan to use two approaches to predict the class.

1. Once the data is preprocessed, we use supervised machine learning classifier like Support Vector Machine(SVM)(This is an initial plan, we might end up using a different classifier).

2. The second method is to use a Recurrent Neural Network(RNN).

The results of the two techniques are then evaluated to see which technique gives more accurate prediction for the problem in consideration.

2 Literature Survey

3 Dataset

4 Preprocessing

5 Experiments

5.1 Embeddings

5.1.1 BERT

5.1.2 Glove

5.2 Classifiers

5.2.1 NN

5.2.2 RNN

6 Reults

7 Conclusions & Future Work

Done : Bert embeddings to NN
    To Do: Bert embeddings to RNN Add twitter glove embeddings Use the glove embeddings to NN Use the glove embeddings to RNN