Natural Language Processing
For
Disasters Tweets

Report by:

Mohamed Ali

Registration number: 210843

Disclamer:

All the results and predictions were based on the real world dataset provided from the train.csv file that contain thousands of tweets pupliched by the following link: https://www.kaggle.com/c/nlp-getting-started/overvie. Also, it contains some inappropriate tweets.

Similar tweets analysis challenges, and the methods used to process them from open-source websites, inspired the implementation of the code used in this assignment.

Motivation

Online social networking has transformed interpersonal communication in recent years. Newer study on social media language analysis is increasingly emphasizing the latter's impact on our daily lives, both personally and professionally. One of the most important ways of analyzing social media content is Natural Language Processing (NLP). Developing sophisticated methods and algorithms to extract useful information from a vast volume of data coming from different sources and languages in varied forms or in free form is a scientific challenge.

With this in mind, applying the knowledge gained in this course to such daily activity seemed tempting, and was exciting enough to endure the challenges faced in this assignment.

Methodology

The task consisted in doing a preprocess of the tweets and then training a machine learning algorithm on them to predict whether they are related to disasters or not. The dataset was quite challenging as it had a lot of tweets 7613, and they were filled with several types of irrelevances.

In the first stage of text processing,the corpus of the tweets was created, following with the cleaning process by removing from the tweets:

1. Stopwords
2. Punctuation
3. URLs
4. HTMLs

Then, the tweets were encoded to make them readable for a machine learning model. This is called vectorization, which is a jargon term for a traditional method of turning raw data (i.e. text) into vectors of real numbers, which is the format supported by machine learning models. This approach has been around since the dawn of computing, has proven to be effective in a variety of disciplines, and is currently being applied in NLP.

The vectorisation used had a value of the fixed random state value = 30

Finally, a machine learning model was trained on the tweets. Given the big size of the data, random forest classifier was the machine learning algorithm used and a confusion matrix shows its fitness.

Results:

First, I would like to share some insights into the data:
The average word length in Disaster tweets is 6.5.
The average word length in Not disaster tweets tweets is 4.98.
Whether a Disaster or Not tweet, the majority of the dataset word count of the tweets between the range 120:140 words per tweet.
The most frequent stopword in both types of tweets is : “the”
The most frequent punctuation in both types of tweets is : “-”

With the help of word clouds:
The most frequent word in the Disaster tweets is: Shelter
The most frequent word in the Not disaster tweets is: Love

Finally, the trained model yielded an accuracy of 80%. The confusion matrix can be seen below.

	Predicted 0 (Not disasters)	Predicted 1(Disasters)
Actual 0 (Not disasters)	801	59
Actual 1 (Disasters)	246	417