Natural Language Processing with Disaster Tweets

Problem Statement

Natural Language Processing (NLP) is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language. Data can come in many forms: time stamps, sensor readings, images, categorical labels, and so much more. But text is still some of the most valuable data out there. Although the text data contains a lot of information, it is highly unstructured, and that is especially hard when we are trying to build an intelligent system which interprets and understands free flowing natural language just like humans. We need to be able to process and transform noisy, unstructured textual data into some structured, vectorized formats which can be understood by any machine learning algorithm.
Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they are observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies).

Objectives

Within this context explained above, this project has the goal to build a machine learning model that predicts which Tweets are about real disasters and which one's aren't, with access to a dataset of 10,000 tweets that were hand classified. However, it's not always clear whether a person's words are actually announcing a disaster, this is why the need for NLP arises, which is the ability of a computer to understand human text and process it to give meaningful output.
Within this context explained above, this project has the goal to build a machine learning model that predicts which Tweets are about real disasters and which ones aren't, with access to a dataset of 10,000 tweets that were hand classified (NLP with Disaster Tweets).
We are to predict whether a given tweet is about a real disaster or not. If so, predict a 1. If not, predict a 0.

Group 9: ***Fatoumata*** ***Fanta*** ***Angela***

Natural Language Processing with Disaster Tweets

Hypotheses

Classification of disaster tweets can help in decision making related to emergency response. With an accurately predicting model, Twitter as a social media platform can be a source of useful information which can be used to inform emergency response agencies of a possible disaster.

Data exploration and Methodology

1. Data description

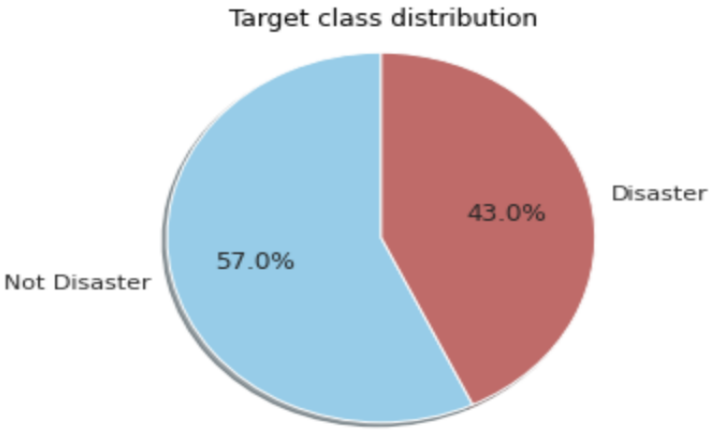
★ File

We have three csv files available with about 10.000 tweets

- train.csv - the training set with 7613 samples
- test.csv - the test set with 3263 samples
- sample_submission.csv - a sample submission file in the correct format with 3263 samples

★ Columns

- id - a unique identifier for each tweet
- text - the text of the tweet
- location - the location the tweet was sent from (may be blank)
- keyword - a particular keyword from the tweet (may be blank)
- target - in train.csv only, this denotes whether a tweet is about a real disaster (1) or not (0)



2. Methodology

★ Cleaning Data

Text data is well known for being highly unstructured. Because of that, one of the most important steps in building NLP projects is cleaning the data.We write some functions to clean our data. Before doing that we concatenate our test and train data.

- remove_URL: this function remove all the url in the data
- remove_html: remove all the html extension in the text
- remove_punct: remove all the punctuation in the text
- remove_StopAndStem: remove Stopwords and Stemming
- remove _emoji: remove all the emoji in the text
- remove_UC : removing Useless Characters

★ Models

BERT(Bidirectional Encoder Representations from Transformers): is a pre-trained deep learning model introduced by Google AI Research which has been trained on Wikipedia and BooksCorpus. Unlike word2vec or Glove, it is contextual. Word bank has different representations in bank deposit and river bank in BERT. Additionally, this model is bidirectional, rather from left to right or from right to left, it uses both contexts to understand a text**.**

Logistic regression, despite its name, is a classification model rather than a regression model. Logistic regression is a simple and more efficient method for binary and linear classification problems. It is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes. It is an extensively employed algorithm for classification in industry. The logistic regression model, like the Adaline and perceptron, is a statistical method for binary classification that can be generalized to multiclass classification.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

Results

We used one third of the dataset for testing and the rest for training, and we came up with the following results:
Having trained three models and also made use of Bert, SVM, and Logistic regression, we had different results for each of them. For the one trained on SVM, we obtained an accuracy of 0.79. On the one trained with Bert, we obtained an accuracy of 0.81, and for the one trained with logistic regression, we had an accuracy of 0.79.

Natural Language Processing with Disaster Tweets

Conclusion

We present Bert as a model better than Logistic regression and in classification. From the results obtained, Bert outperformed the rest of the models, as it had the highest accuracy of 0.81. We can also conclude that tweet classification can be useful in enabling a swift response to emerging disasters.

Related papers

Some of the papers related to this project may be accessed from the link below:

https://www.researchgate.net/publication/311990101_Identifying_and_Categorizing_Disaster-Related_Tweets

https://arxiv.org/abs/2202.00795

https://paperswithcode.com/paper/cross-lingual-disaster-related-multi-label