Mbargane Aboubacry SENE

msene@aimsammi.org

04 September 2022

Title : Natural Language Processing with Disaster Tweets

- Problem statement

  Twitter has become an important communication channel in times of emergency. We're challenged to build a machine learning model that predicts which Tweets are about real disasters and which one's aren't. The dataset for the competition contains text that may be considered profane, vulgar, or offensive. For this reason we are going to describe in a few lines the procedure we have adopted, in perfect accordance with our objective.

- Objective

In simple terms our objective is the following, for each ID in the test set, to predict 1 if the tweet is describing a real disaster, and 0 otherwise, which is classification work. We intend to do this under the following assumptions, below.

- Hypotheses and Data exploration Methodology

  Let us specify that we use the BERT model and the following version: 'bert-base-cased'. In our work, we use the count vectorizer with Logistic Regression and Naive Bayes, TF-IDF with logistic regression and Naive Bayes. In the case of Logistic Regression and Naive Bayes, we used cross-validation for more accurate validation scores. We also decided to use k-fold cross-validation on the dataset and used accuracy as when evaluating the model. We first perform training-validation splits on the training dataset with 0.2 percent for the validation of the Model.

Results

First we implement Logistic Regression and Naive Bayes using a count vectorizer. The count vectorizer with Logistic regression gives a validation accuracy of 0.73026 and Naive Bayes have an average 0.48965. For the two models with Tfidf, Logistic regression yields an accuracy of 73.31228 percent, and count vectorizer and Naive Bayes have an accuracy of 0.47109 percent. Finally, we train the BERT with version '`bert-base-cased` '; with 3 epochs and 50 for batch_size. This model outperforms the two other models with a validation accuracy of 0.8286. You will find below the support of our documentation.

Related papers

- XAVIER Cedric BOUSBIB Ruben. "Sentiment Analysis of a Tweet With Naive Bayes". In: github.io, https://rubenbsb.github.io/pdfs/nlp-project-mva.pdf.
- https://stackoverflow.com/questions/tagged/bert-language-model
- https://deepnote.com/@abid/Disaster-tweet-classification-64278d77-b455-4fcb-b98a-076ff504a9ee
- https://huggingface.co/tftransformers/bert-base-cased