Kaggle Tweets Classification

Tweets Binary Classification Using ML-DL LaTeX

Nikos Nteits

July 28, 2025

Agenda

Dataset

Preprocessing

Processing

Results

Final Remarks

Dataset

Kaggle Tweets Dataset

Each sample in the train and test set has the following information:

- 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).

Kaggle Tweets Dataset

Relatively Balanced Dataset:

- target $=0$: 4342
- target $=0$: 3271

Preprocessing

Preprocessing Steps:

- Special characters Removal
- Text in square brackets Removal
- Non-word characters Removal
- URLs Removal
- HTML tags Removal
- Words containing numbers Removal
- Stopwords Removal
- Stemming

Processing

Processing:

- **Multiple Vectorization techniques were implemented**
- **Multiple models were tried**

Vectorizers:

- Binary Count Vectorizer
- Count Vectorizer
- Tfidf Vectorizer
- Tfidf Vectorizer 1-2 grams
- Word2Vec 300 (mean)
- Glove Twitter 200d
- USE-4
- USE-large 5
- Bertweet base (mean)

Models:

- Logreg
- Lasso
- Ridge
- SVM
- Multinomial Naive Bayes
- Decision Tree

Models:

Combinations of Vectorizers and Models shown above were ran . At last a Pretrained Roberta Base Model was used with a classification head on top. The model was fine tuned for 5 epochs with 0,1 dropout rate and learning rate of 0.00002.

Results

Results:

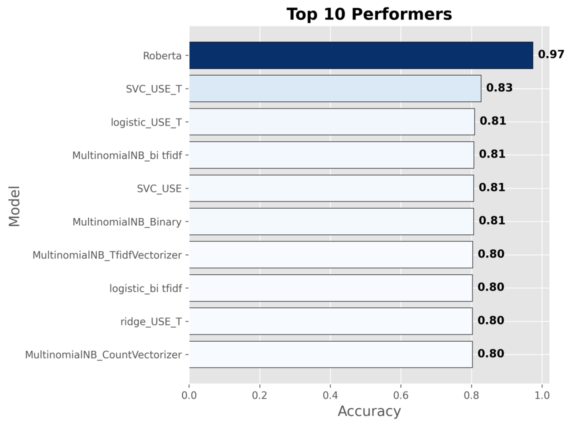

Figure 1

Final Remarks

Final Remarks:

Roberta significantly outperforms every handcrafted Vectorizer-Model combination ,something foreseeable as the model has been trained on Twitter data hence transfer is natural.As for the rest, universal sentence encoder seems to have an edge over the rest as models with use as vectorizer slightly outperform their coresponding siblings.