IMRSV — does NLP and Speech analysis

How to classify Sentences

toxic
threat
insult
neutra
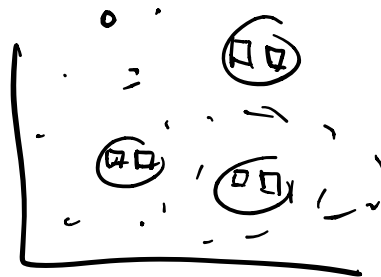
Counts
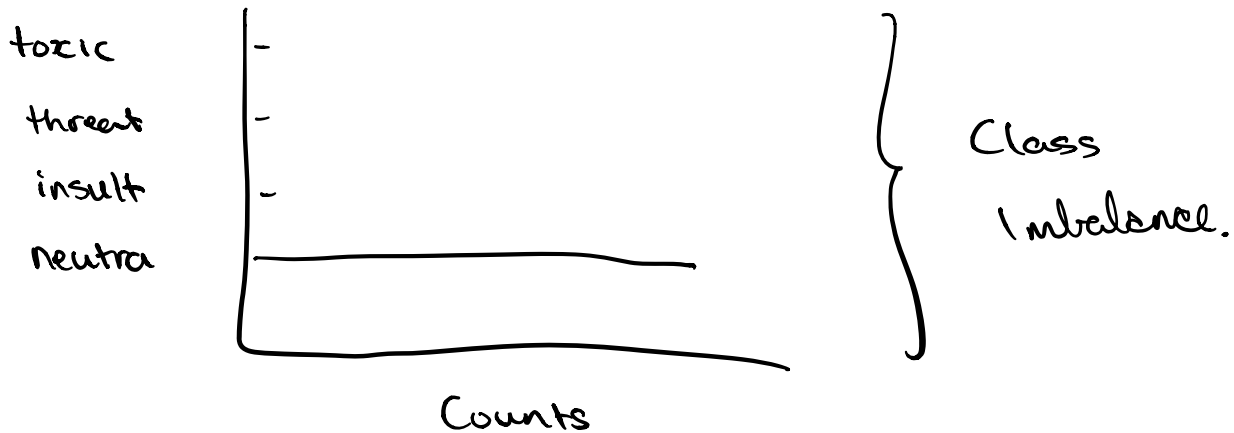
} Class
  Imbalance.

Problems:

Overlapping Classes

Small Disjuncts

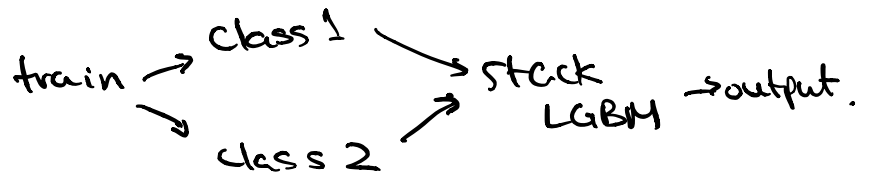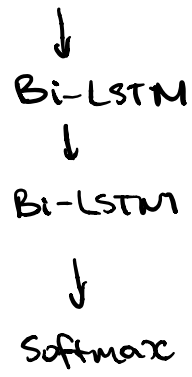True Positive rate = $TP/(TP+FN)$

- -
- -

○ ROC - AUC.

Solutions
  ○ Data vs Model

— DATA —
    • augmentation using translations ($en \rightarrow fr \rightarrow en$)
    ○ random undersampling (discard overpop
    ○ random oversampling (magnify small sample)

— MODEL —

Embeddings

$\downarrow$

Bi-LSTM

$\downarrow$

Bi-LSTM

$\downarrow$

Softmax

train $\rightarrow$ Class 1 $\searrow$ stack → output.
$\searrow$ Class 2 $\nearrow$ LGBM

How to find data use Out of Fold (OOF) predictions.