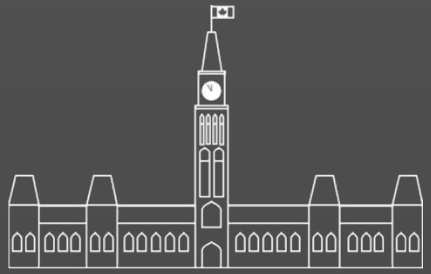
IMRSV
DATA LABS

WHO WE ARE

IMRSV
DATA LABS

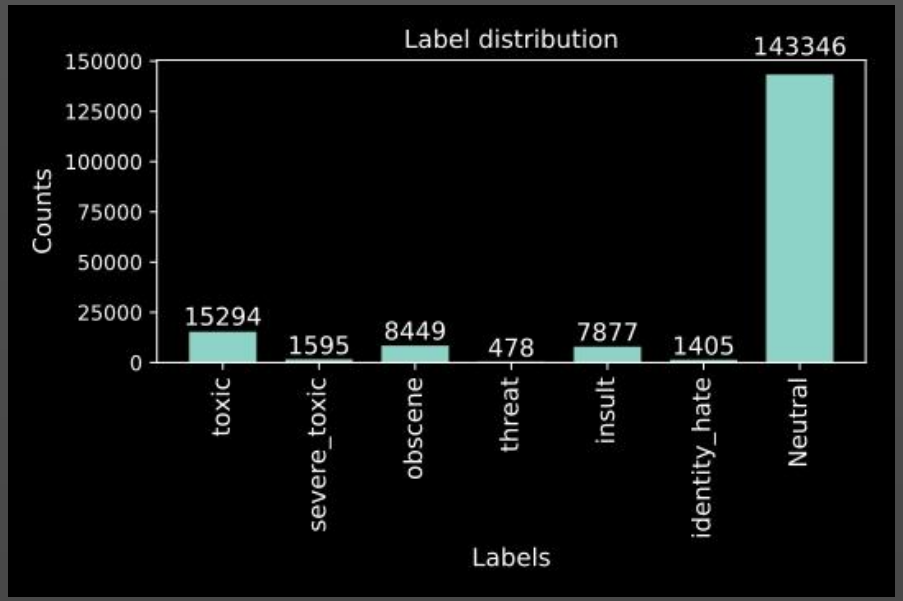
Venture backed ML research lab

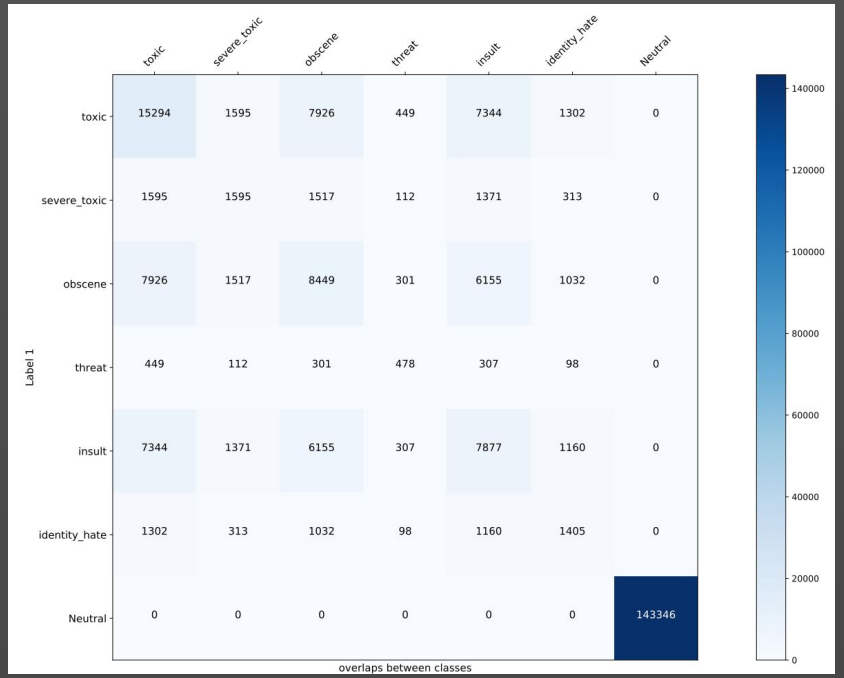


Located in Ottawa

Handling class imbalance in NLP

How unbalanced can a dataset be?

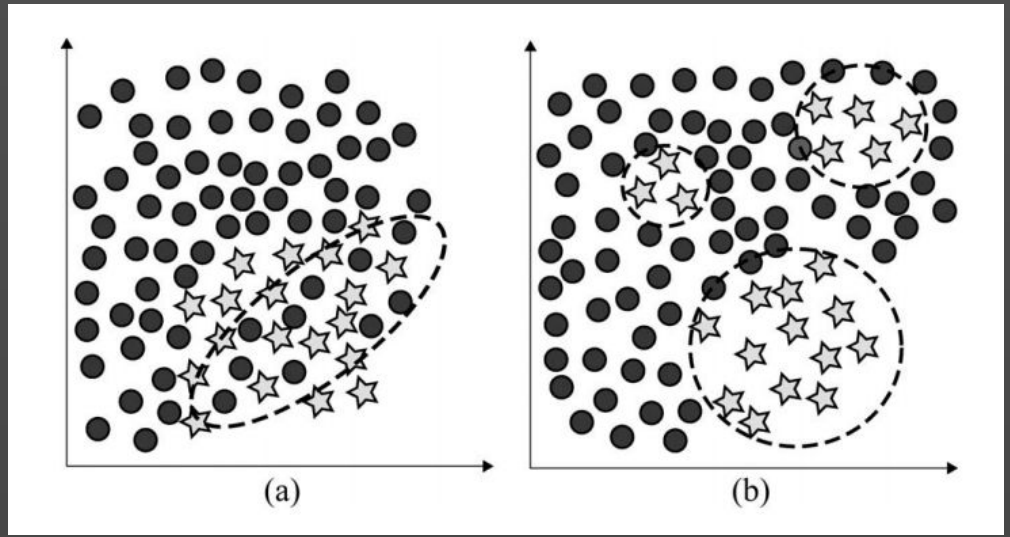




A review of standard text classification practices for multi-label toxicity identification of online content - Isuru Gunasekara, Isar Nejadgholi

The problem of Imbalanced Data Sets

- Small sample size
- Overlapping or class separability (a)
- Small disjuncts (b)



Source : https://sci2s.ugr.es/keel/pdf/algorithm/articulo/2011-IEEE%20TSMC%20partC-%20GalarFdezBarrenecheaBustinceHerrera.pdf

Cost functions

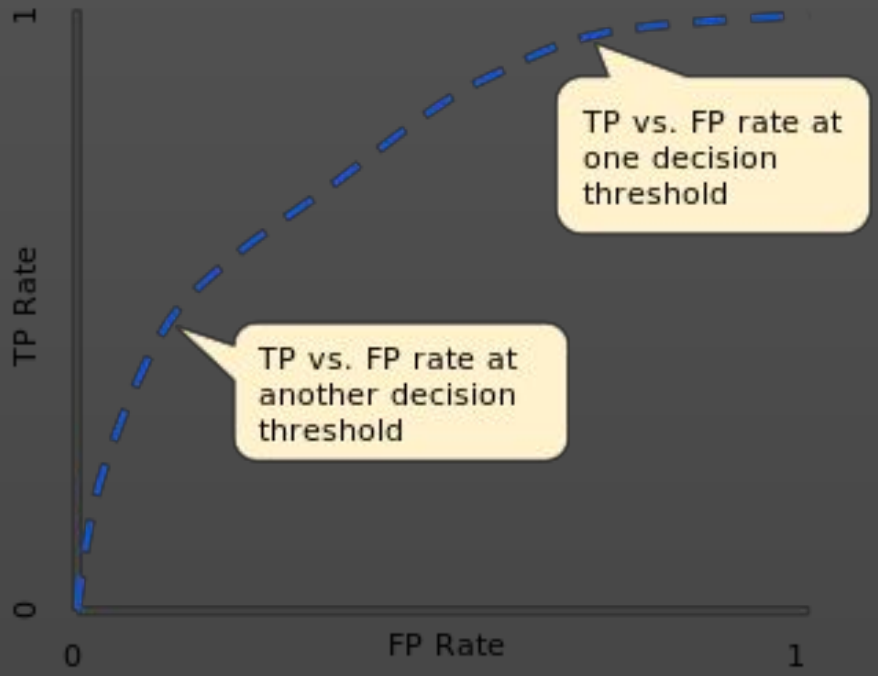
- How to select the right cost function for an unbalanced dataset?
- Simply using accuracy would not give any meaningful indication of performance
  - Acc = (TP+TN)/(TP+FN+FP+TN)
  - In a binary classification, if 90% of the dataset is ‘true’, any classifier that predicts every sample as ‘true’ would get a 90% accuracy

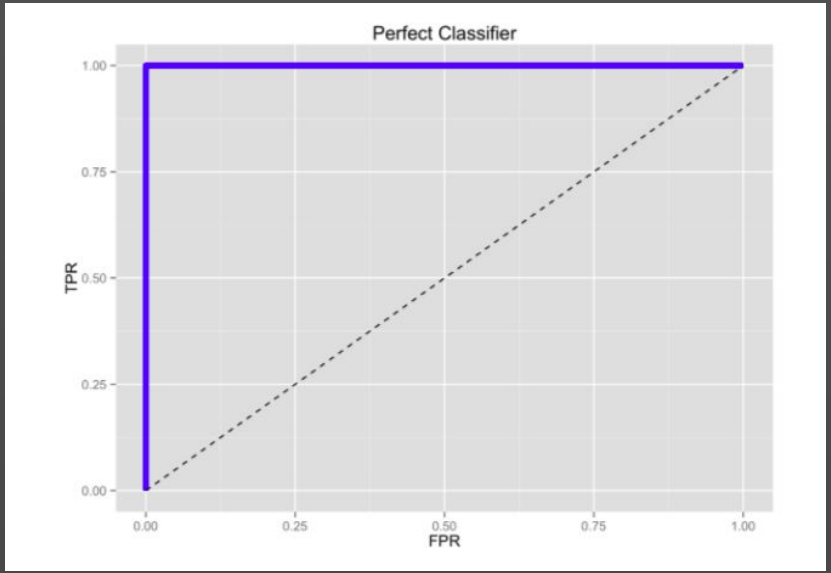
What could we do?

- True positive rate = TP/(TP+FN) = Percentage of positive instances correctly classified
- True negative rate = TN/(FP+TN) = Percentage of negative instances correctly classified
- False positive rate = FP/(FP+TN) = Percentage of negative instances misclassified
- False negative rate = FN/(TP+FN) = Percentage of positive instances misclassified

But none of these measures alone are not adequate by itself 

ROC AUC





Sources: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
https://fderyckel.github.io/2017-03-15-Logistic_Regression_Part_I/

Data level approaches for handling imbalance

- Data augmentation using translations
- Random undersampling
- Random oversampling
- Synthetic minority oversampling technique (SMOTE) [1]
- Modified synthetic minority oversampling technique (MSMOTE) [1]
- Selective preprocessing of imbalanced data (SPIDER) [1]

Translations for generating data

- En -> Other language -> En
  - En -> French -> En
  - En -> Spanish -> En
  - En -> German -> En

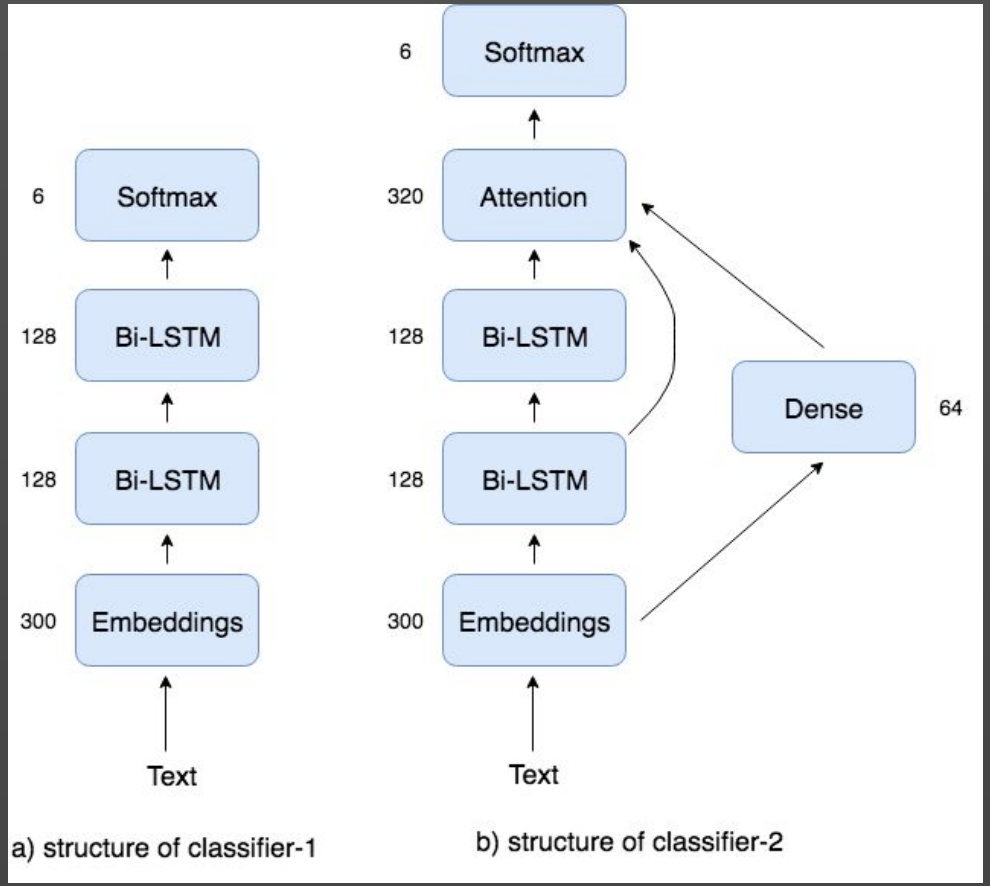
Random undersampling

- Random elimination of majority class examples
- But can discard potentially useful data

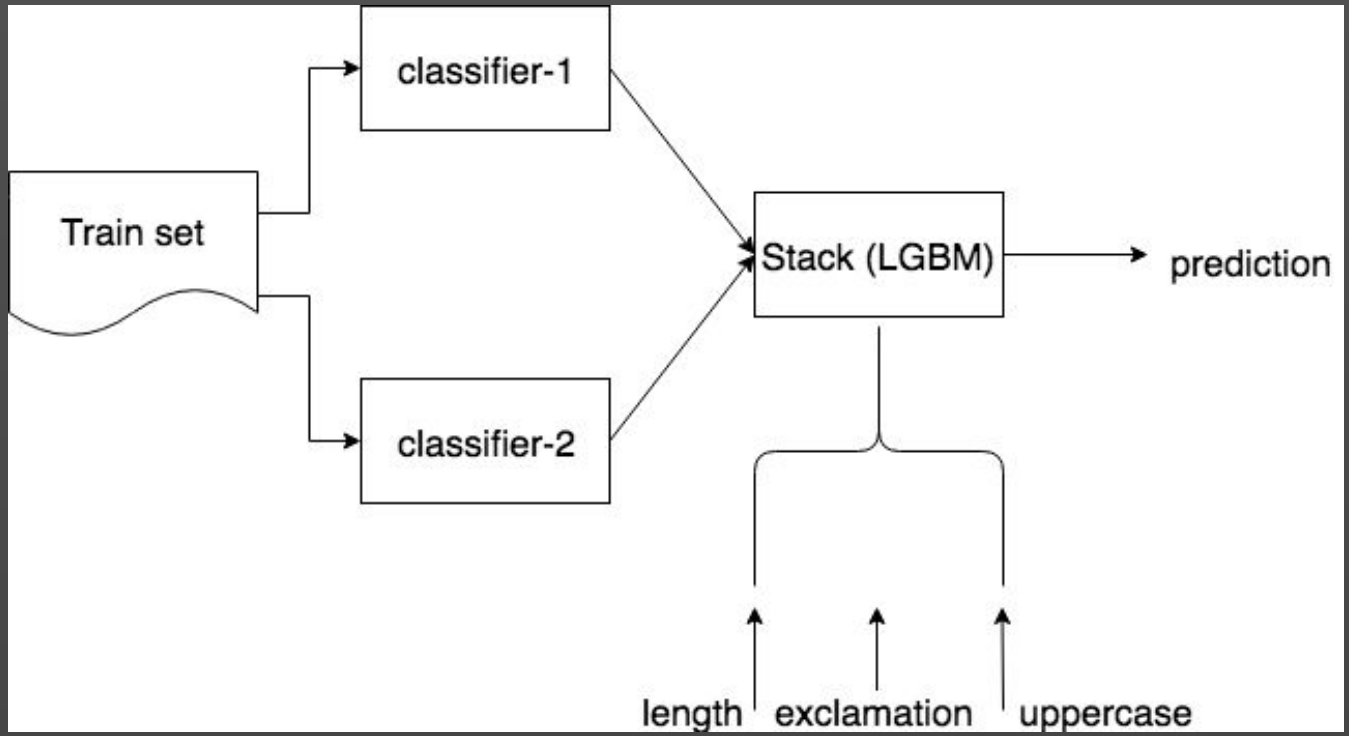
Random oversampling

- Randomly replicate minority class instances
- Generally outperforms undersampling
- But can cause overfitting

Model based data augmentation



Model based data augmentation



How to find data to create a stacking model?

- Out Of Fold (OOF) predictions!
  - Dataset is divided into 10 folds with 90/10 split
  - Each model is trained 10 times for these folds
  - Each model does inference on the 10% OOF data
  - These 10 sets of 10% chunks are then combined to create a full dataset of soft labels for stacking

How to create more data from a stack?

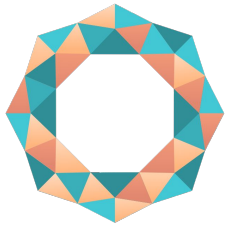
- Run unlabelled data through the two classifiers
- Run the soft labels through the stacker
- Train the individual models separately using the labels (minority classes) obtained from the stacker

Increase in ROC AUC Score

Classifier	**Training**	**ROC AUC**
CLS-1 (BI-LSTM)	Supervised	0.9842
CLS-2 (BI-LSTM+Attention)	Supervised	0.9844
LGBM	Supervised	0.9847
CLS-1 (BI-LSTM)	Semi-supervised	0.9860
CLS-2 (BI-LSTM+Attention)	Semi-supervised	0.9862

References

1. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches - Mikel Galar et al.
   https://sci2s.ugr.es/keel/pdf/algorithm/articulo/2011-IEEE%20TSMC%20partC-%20GalarFdezBarrenecheaBustinceHerrera.pdf
2. How to handle Imbalanced Classification Problems in machine learning?
   https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
3. A review of standard text classification practices for multi-label toxicity identification of online content - Isuru Gunasekara, Isar Nejadgholi - ALW2 in EMNLP 2018 https://drive.google.com/file/d/1Ea9QuE1g5oBfBg7ik-UM9wIjnSvzaAkL/view



IMRSV
DATA LABS

QUESTIONS?