Understanding Data Requirements for Toxic Comment Classification

Roshan Konda, Ethan Channell, Pratik Nallamotu

Problem Statement

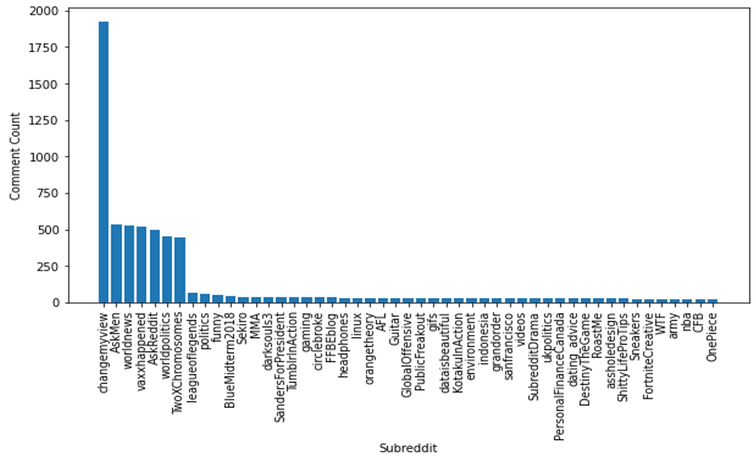
Toxic comments are defined as any type of **negative** or **offensive speech** that would cause a participant to leave a discussion.

Classifying these comments can be very **difficult**, **time consuming**, and **psychologically taxing**.

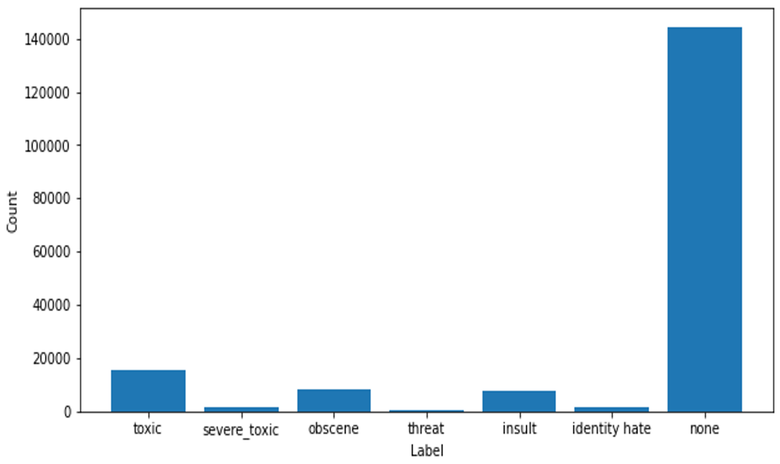
Our project aims to apply techniques to **reduce the load on human annotators** and still achieve good performance using **weak supervision** and **active learning** techniques.

Datasets

Ruddit dataset - posts for different subreddits are rated from -1 (maximally supportive) +1 (maximally offensive). **Discretized** into three categories, **Toxic, Neutral, and Positive.**



Jigsaw Dataset - comments collected from wikipedia and annotated with 6 categories.



Baseline Methods

Logistic Regression, SVM, Random Forest, LSTM, BERT

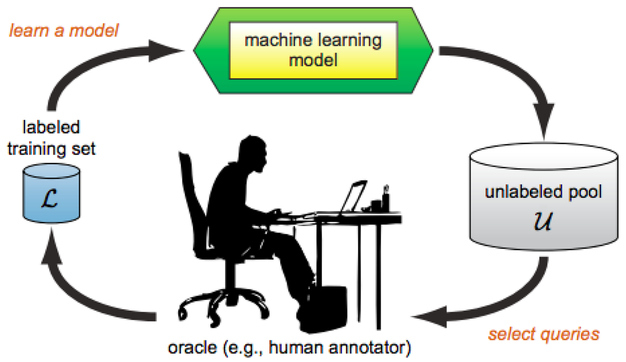
Method	Dataset	
	Ruddit	Wikipedia
Log Reg. (Count Vectorizer)	0.769	0.867
Log Reg. (TFIDF)	0.782	0.867
SVM (Count Vectorizer)	0.789	0.833
SVM (TFIDF)	0.795	0.873
Random Forest (Count Vectorizer)	0.795	0.855
Random Forest (TFIDF)	0.807	0.856
LSTM	0.800	0.906
BERT	**0.824**	**0.940**

BERT resulted in the **best accuracy** for both of the datasets.

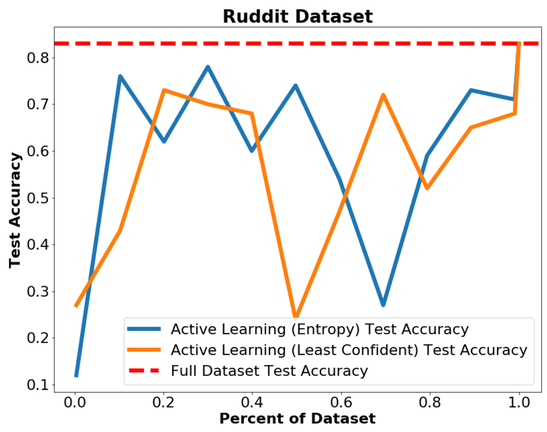
Active Learning

Least Confidence - $x^*_{LC} = argmax_x(1\text{-}P_\theta(\hat{y}|x))$

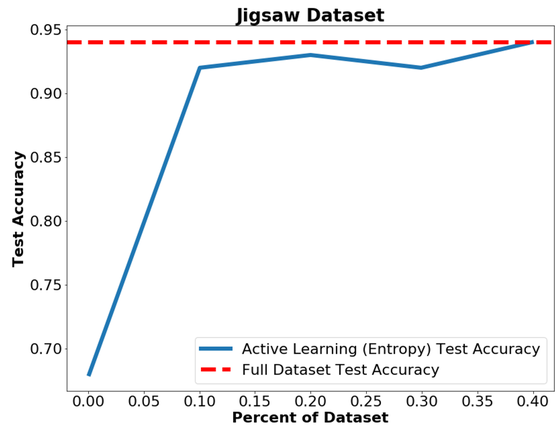
Prediction Entropy - $x^*_h = argmax_x(-\sum P_\theta(y_i|x)\log(P_\theta(y_i|x))$



Credit: https://odsc.medium.com/active-learning-your-models-new-personal-trainer-a89722c0db5a



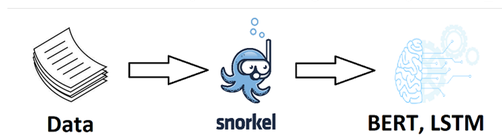
Active Learning Cont.



Weak Supervision

Snorkel was used to develop **labeling functions**.

Labeling functions include finding **bad words**, **keywords associated with toxicity**, and **subjectivity and polarity**.



Method	Dataset	
	Ruddit	Wikipedia
LSTM	0.682	0.510
BERT	0.744	0.520

Conclusion

Compared to our baseline methods, the use of **weakly supervised labeled data** led to **decreased performance**. Toxic language is **nuanced** and sometimes **not explicit**.

For **Jigsaw dataset** got comparable results as full dataset using **40%** of data and **Prediction Entropy** query strategy.

For **Ruddit dataset active learning was not able to reduce data required**. This may be because we discretized the offensivensive ness score.