Toxic comment sentiment classification on Wikipedia comments

Ity Bahadur,
Naima Ahmed Fahmi and
Sheikh Nabil Mohammad

UCF

Motivation

- Topic selection:
    - Why sentiment analysis?
- Data source:
    - Kaggle dataset

Dataset and Tools used

- Toxic comment classification of Wikipedia comments from Kaggle (Kaggle Dataset)
- 159571 samples
- Training Data: Provided by Kaggle (train.csv file)
- Test Data: Provided by Kaggle (test.csv and test_labels.csv)
- NLTK and Scikit Learn

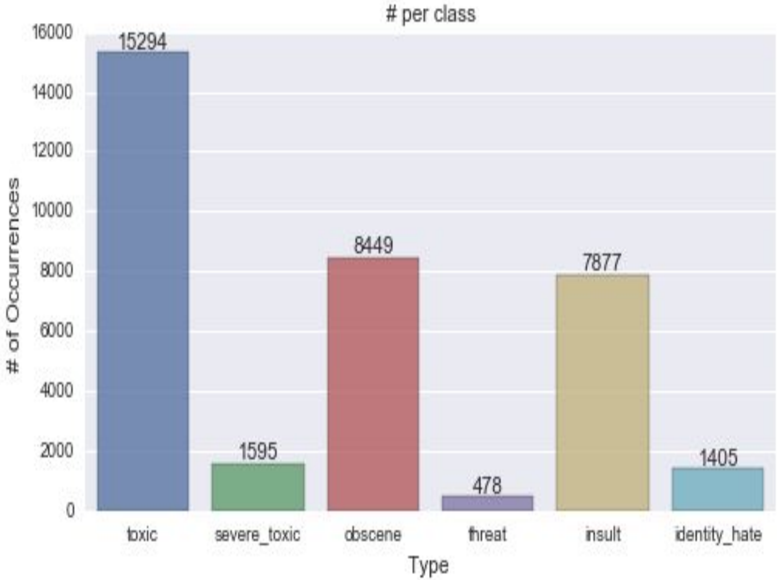
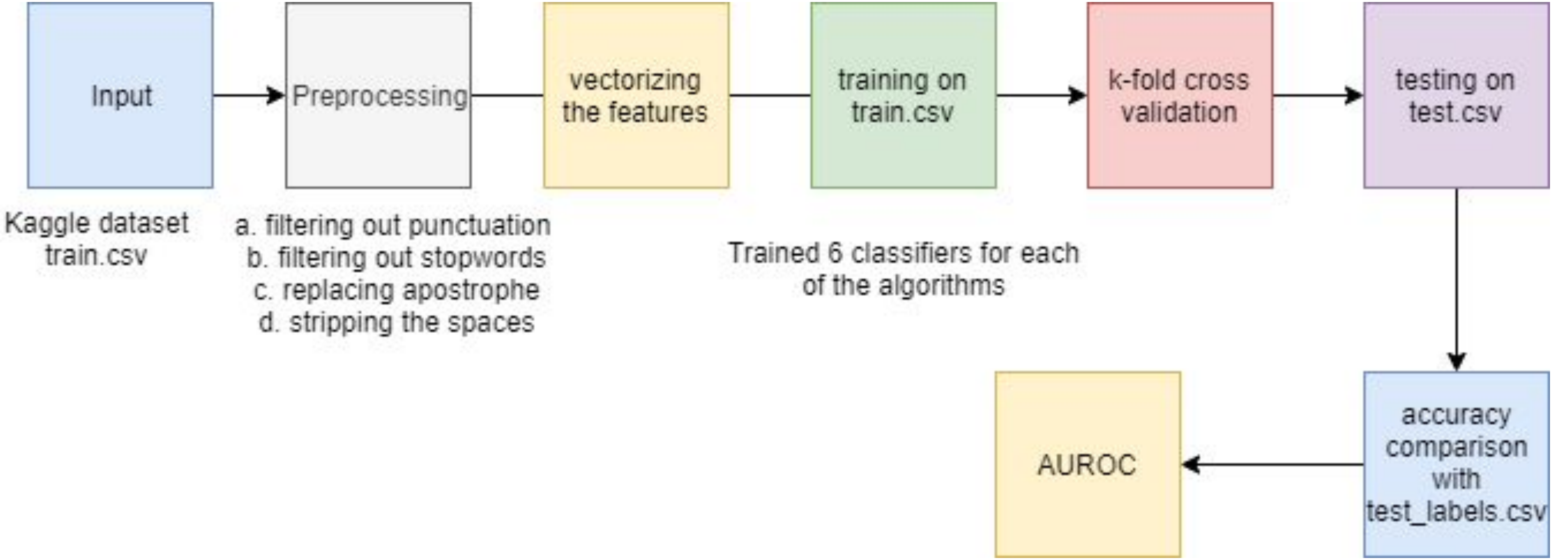


Fig: distribution of labels in comments

Methodology



Input

Kaggle dataset
train.csv

Preprocessing

a. filtering out punctuation
b. filtering out stopwords
c. replacing apostrophe
d. stripping the spaces

vectorizing the features

training on train.csv

Trained 6 classifiers for each of the algorithms

k-fold cross validation

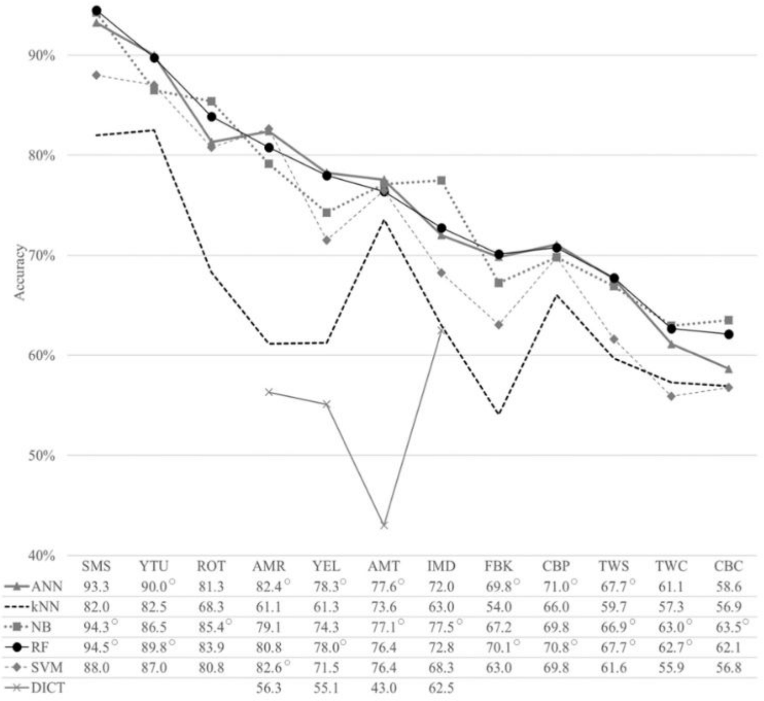
testing on test.csv

accuracy comparison with test_labels.csv

AUROC

Algorithms implemented

- Multinomial Naïve Bayes
- Linear SVM
- Logistic Regression
- Random Forest
- Decision Tree



	SMS	YTU	ROT	AMR	YEL	AMT	IMD	FBK	CBP	TWS	TWC	CBC
ANN	93.3	90.0°	81.3	82.4°	78.3°	77.6°	72.0	69.8°	71.0°	67.7°	61.1	58.6
kNN	82.0	82.5	68.3	61.1	61.3	73.6	63.0	54.0	66.0	59.7	57.3	56.9
NB	94.3°	86.5	85.4°	79.1	74.3	77.1°	77.5°	67.2	69.8	66.9°	63.0°	63.5°
RF	94.5°	89.8°	83.9	80.8	78.0°	76.4	72.8	70.1°	70.8°	67.7°	62.7°	62.1
SVM	88.0	87.0	80.8	82.6°	71.5	76.4	68.3	63.0	69.8	61.6	55.9	56.8
DICT				56.3	55.1	43.0	62.5					

Fig. 1. Accuracies of automated text classification in reflecting human intuition across 12 social media types. Note: ° indicate insignificant differences between the best methods ($p > .05$). DICT is the average of five lexicon-based methods, i.e., LIWC, NRC, AFINN, BING, and VADER (see Appendix B for details).

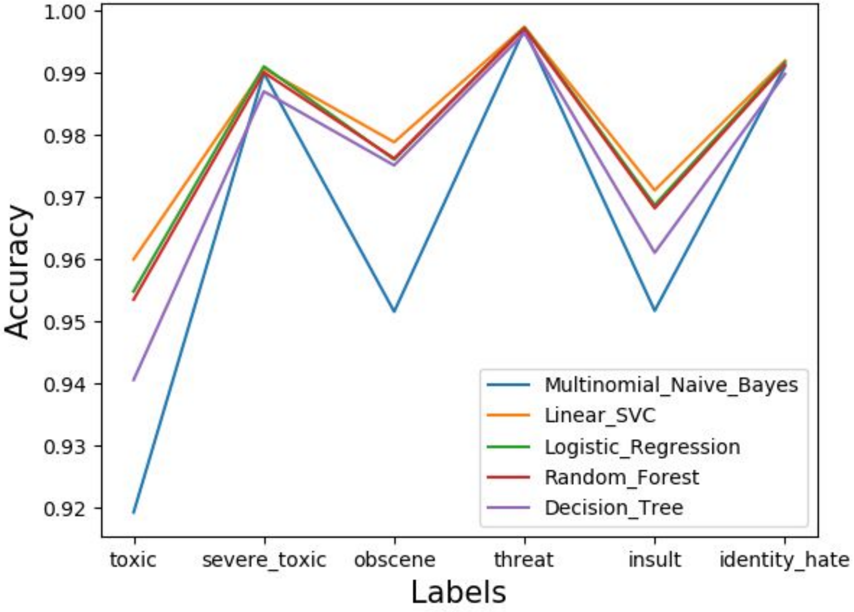
Results from Survey paper

Performance Comparison (Accuracy)

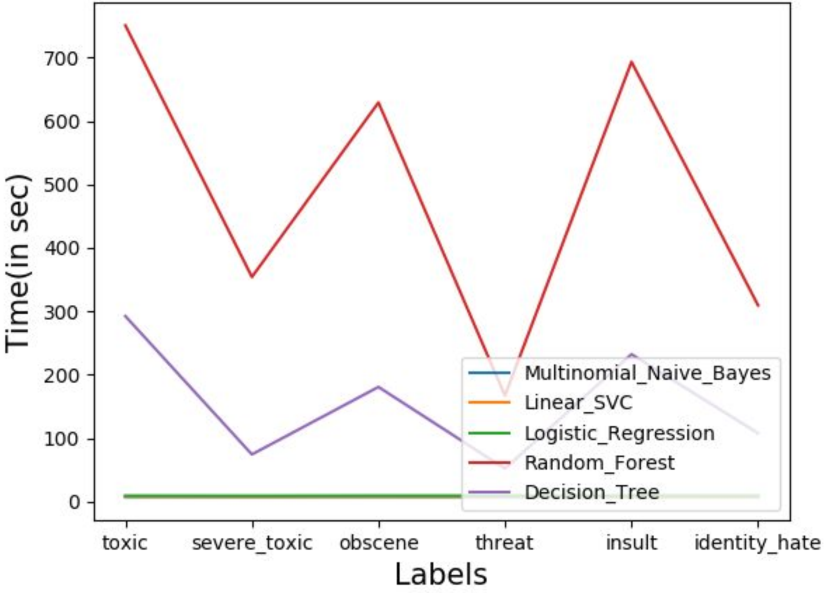
Labels	**MultiNomial NB**	**Linear SVC**	**Logistic Regression**	**Random Forest**	**Decision Tree**
Toxic	0.919	0.960	0.955	0.953	0.940
Severe Toxic	0.990	0.991	0.991	0.990	0.987
obscene	0.952	0.979	0.976	0.977	0.975
Threat	0.997	0.997	0.997	0.997	0.997
Insult	0.952	0.971	0.969	0.968	0.962
Identity_hate	0.991	0.992	0.992	0.991	0.990
Average	**0.967**	**0.982**	**0.980**	**0.979**	**0.975**

Performance Comparison

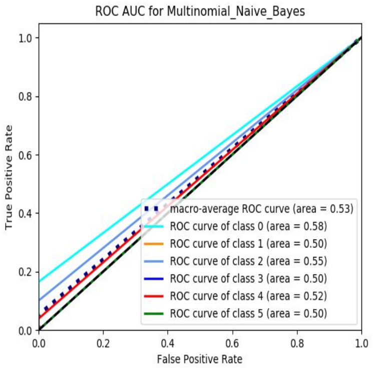
Accuracy



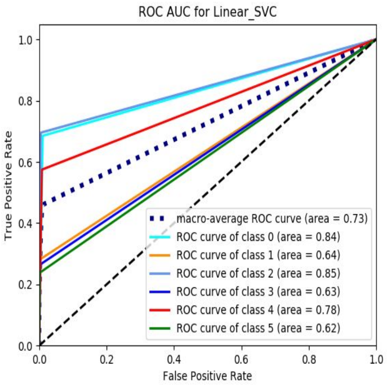
Time



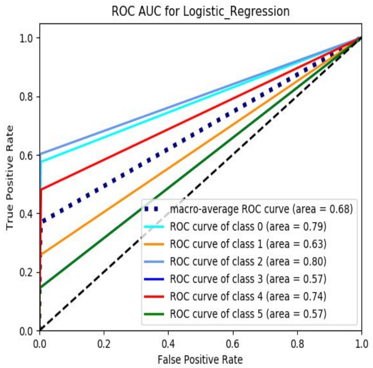
Performance Comparison (ROC AUC)



ROC AUC for Multinomial_Naive_Bayes

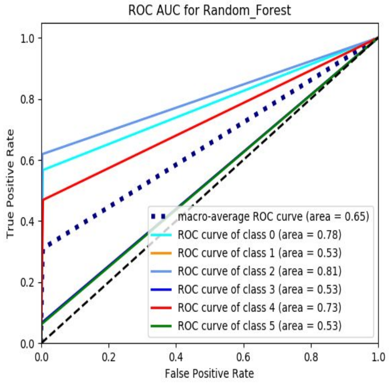


ROC AUC for Linear_SVC

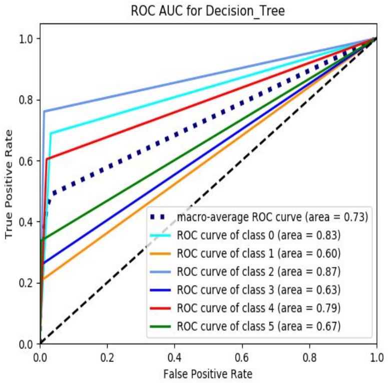


ROC AUC for Logistic_Regression

0 - Toxic
1 - Severe Toxic
2 - Obscene
3 - Threat
4 - Insult
5 - Identity Hate



ROC AUC for Random_Forest



ROC AUC for Decision_Tree

Discussion

- All classifiers have good AUC scores for classifying ‘toxic’ and ‘obscene’ texts.
- Worst time complexity: Random forest and Decision Tree

Future work:

- Use of n-gram/bag of words for feature extraction

Thank You!

Questions?

References

1. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/overview

2. https://www.sciencedirect.com/science/article/pii/S0167811618300545