DISCo: Detecting Insults in Social Commentary

Jayadev Bhaskaran*, Amita Kamath**, Suvadip Paul**
*ICME **Dept. of Computer Science, Stanford University
{jayadev, kamatha, suvadip}@stanford.edu

CS 229
Fall 2017

MOTIVATION

- **Goal:** Classify comments in online discussion platforms as "personal attacks", i.e. abusive comments
- Does contextual information about the user help?

PROBLEM DEFINITION

- **Dataset:** Wiki-detox, "personal attacks" corpus
- Each comment is classified, irrespective of severity, by at least 10 independent "workers" (majority vote)
- **Dataset size:** Around 100,000 annotated comments
- **Evaluation**: F1 score, ROC AUC
- **Additional considerations:** Condition on user priors
- **Train-Dev-Test:** 60-20-20 (random shuffling)

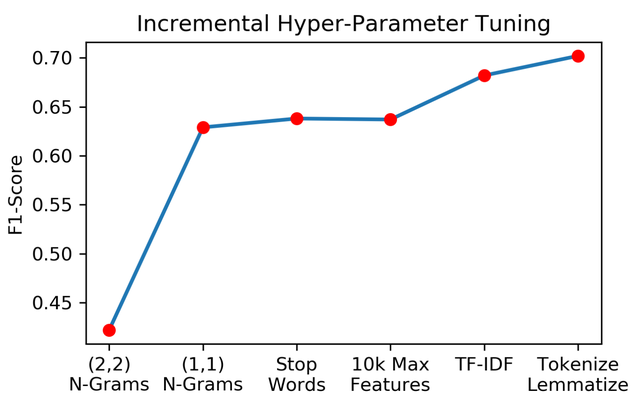
BASELINES

Data Pre-processing

- Cleaning data (punctuation removal, lemmatization)
- Choice of word embeddings (n-grams, GloVe, word2vec)
- Use of stop-words/limiting max. features
- TFIDF weighting

Models

1. **Logistic Regression** (baseline)
2. **Multinomial Naïve Bayes**
3. **SVM** – linear kernel
- Hyper-parameter tuning: regularization constant, kernel type (SVM)



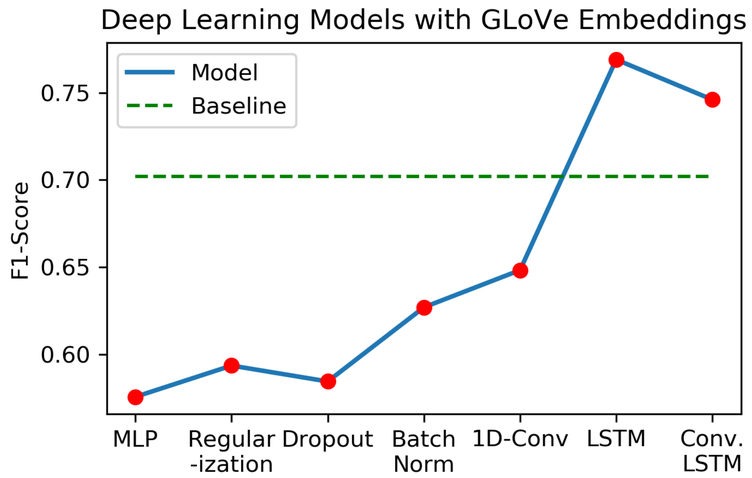
DEEP LEARNING MODELS

Models

1. **MLP:** (200x200x50)
2. **ConvNets:** 4 Conv (128x5x1), MaxPool, 3 Conv (128x5x1), AvgPool
3. **LSTM:** 1 LSTM (128)
4. **ConvLSTM:** 1 Conv (128x5x1), MaxPool, 1 LSTM (128)

Hyper-parameter tuning: Regularization constant, learning rate, dropout, batch normalization, early stopping

Architecture tuning: Increasing number and width of layers



RESULTS

Model	F1 Score	ROC AUC
Logistic Regression	0.702	0.959
Multinomial Naïve Bayes	0.598	0.933
SVM (linear kernel)	0.734	N/A
MLP (with TFIDF ngrams)	0.747	0.952
MLP (GLoVe)	0.635	0.911
ConvNets (GLoVe)	0.648	0.917
LSTM (GLoVe)	0.769	0.962
ConvLSTM (GLoVe)	0.746	0.957

ANALYSIS AND OBSERVATIONS

- Preprocessing: 1-grams outperformed other n-grams (key insult in a comment usually concentrated at one word)
- TFIDF weighting/stop words improve performance (reduce weights given to common words)
- SVM: Choice of kernels – linear kernel yielded highest performance
- MLP with TFIDF weighted 1-grams outperformed all 3 conventional methods
- Deep Learning methods with GLoVe embeddings: LSTM/ConvLSTM gave better results than 1-gram based MLPs – **performance does improve using sequential information obtained through GLoVe embeddings**
- We find that insults/trolls and abusive commenters could be any user, i.e. anyone could become abusive (consistent with literature)
- However we find **anonymous users are more likely to abuse**

	Logistic Regression			MLP with 1-grams		
User Anonymity Data	**AUC**	**Recall**	**F1 Score**	**AUC**	**Recall**	**F1 Score**
Not Incorporated	0.907	0.32	0.488	0.870	0.52	0.616
Incorporated	0.913	0.38	0.544	0.876	0.54	0.632

FUTURE WORK

- Investigate further if using knowledge about user-specific metadata improves model performance
- Determine if some article topics are more likely to incite abusive comments than others (Topic Detection)
- Temporal characteristics – is there an observable trend in the quantity/nature of abusive comments over time?
- Explore the effect of incorporating worker demographics (age, gender, education) on classification

REFERENCES

[1] E. Wulczyn, N. Thain, and L. Dixon. "Ex machina: Personal attacks seen at scale," Proceedings of the International Conference on World Wide Web (WWW), 2017
[2] J. Pennington, R. Socher, and C. D. Manning. "GloVe: Global vectors for word representation," Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), 2014