CERIAS

The Center for Education and Research in Information Assurance and Security

Authorship Attribution of Predators in Chat Conversations

Kanishka Misra, Hemanth Devarapalli, Julia Taylor Rayz

{kmisra, hdevarap, jtaylor1}@purdue.edu

PURDUE RESEARCH FOUNDATION

Abstract

Authorship Attribution (AA) of written content presents several advantages within the digital forensics domain. While AA has been successful when applied to long documents, recent works have shown improved performance of neural AA models on short texts such as tweets and online conversations (Schwartz et al., 2013; Ruder et al., 2016; Shrestha et al., 2017). Concurrently, the rise of social media as well as a plethora of chat messaging platforms have made it easier for teenagers to be vulnerable to online predators. In this work, we present a new model to attribute authors to messages from a corpus consisting of chat conversations, some of which involve online predators, and perform subsequent analysis of neural representations of messages. Our results show comparable performance to prior work for Authorship Attribution and highlight differences between predatory and non-predatory message styles.

Results Summary

Table 1: Micro Avg. f1 scores

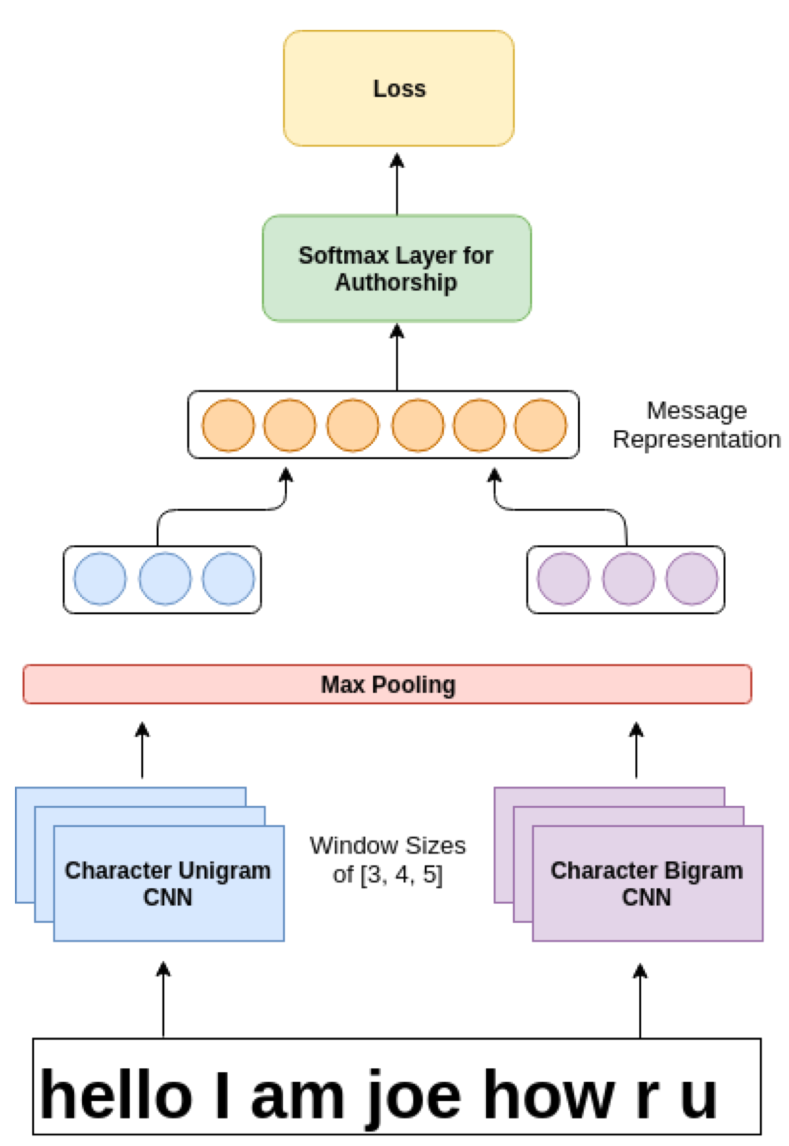
Model	10 Authors	50 Authors
Ruder et al., 2016	0.525	0.3524
Shrestha et al., 2017	**0.588**	0.4474
Ours (AA-CNN)	0.557	0.4382
Ours (AA-CNN-PC)	0.549	**0.4484**

Table 2: Differences between Pred. and Non Pred. messages in both Models

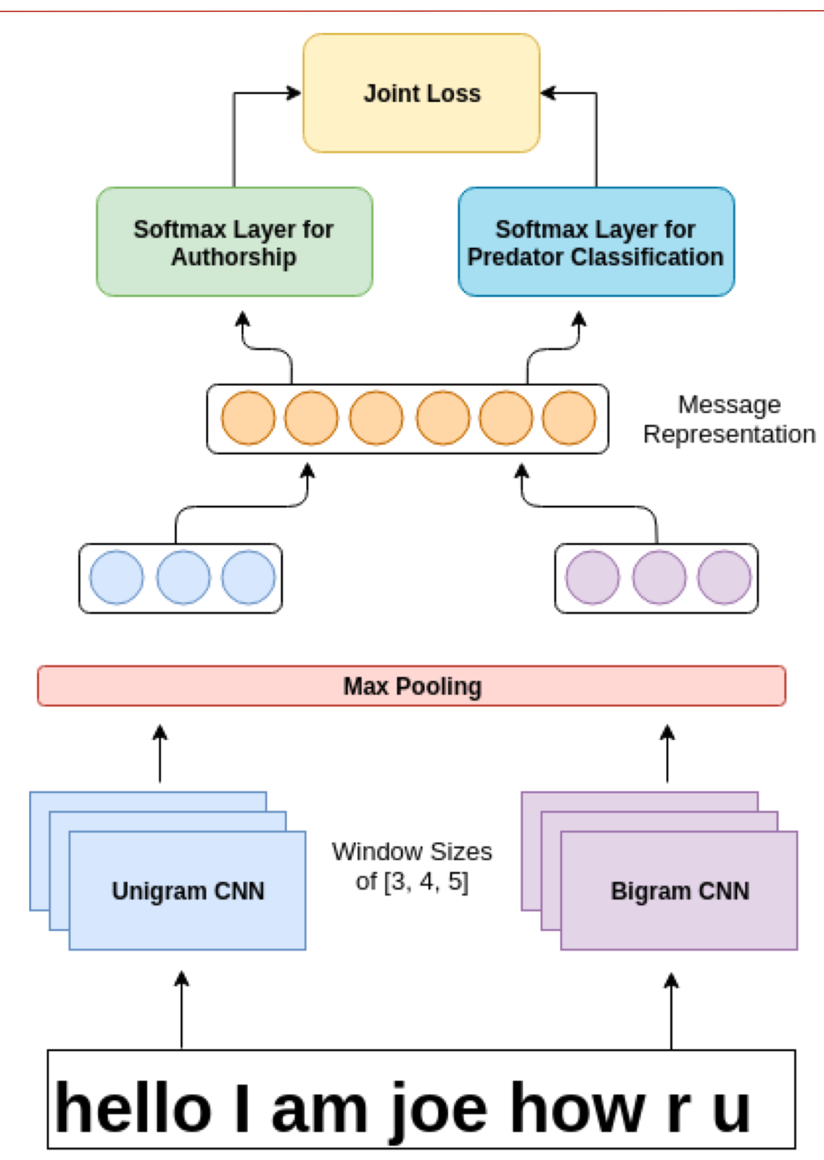
Model	ΔMAS
AA-CNN	0.021 ($t = 1048.3$, $p = 2.2\text{x}10^{-16}$)
AA-CNN-PC	0.025 ($t = 1285.8$, $p = 2.2\text{x}10^{-16}$)

- **Both model give comparable performance relative to Baselines** (Improvement in 50 Author set & Second best in the 10 author set.
- Models were further probed by investigating the differences between Encoded messages from Predators as well as Non Predators.
- Difference measured by the change in Mean Avg. Similarity (ΔMAS) of pred. messages to other pred. messages versus pred. messages to non pred. messages.
- $$\Delta MAS = \frac{1}{N_i}\frac{1}{N_j}\sum_{i}^{N_i}\sum_{j}^{N_j} cos(V_i^{predator}, V_j^{predator})1_{i\neq j} - \frac{1}{N_i}\frac{1}{N_j}\sum_{i}^{N_i}\sum_{j}^{N_j} cos(V_i^{predator}, V_j^{non-predator})$$
- ΔMAS showed that **the simple AA-CNN model implicitly learned the difference between predator and non-predator style.**

Model Architecture



Architecture of AA-CNN



Architecture of AA-CNN-PC

Methodology

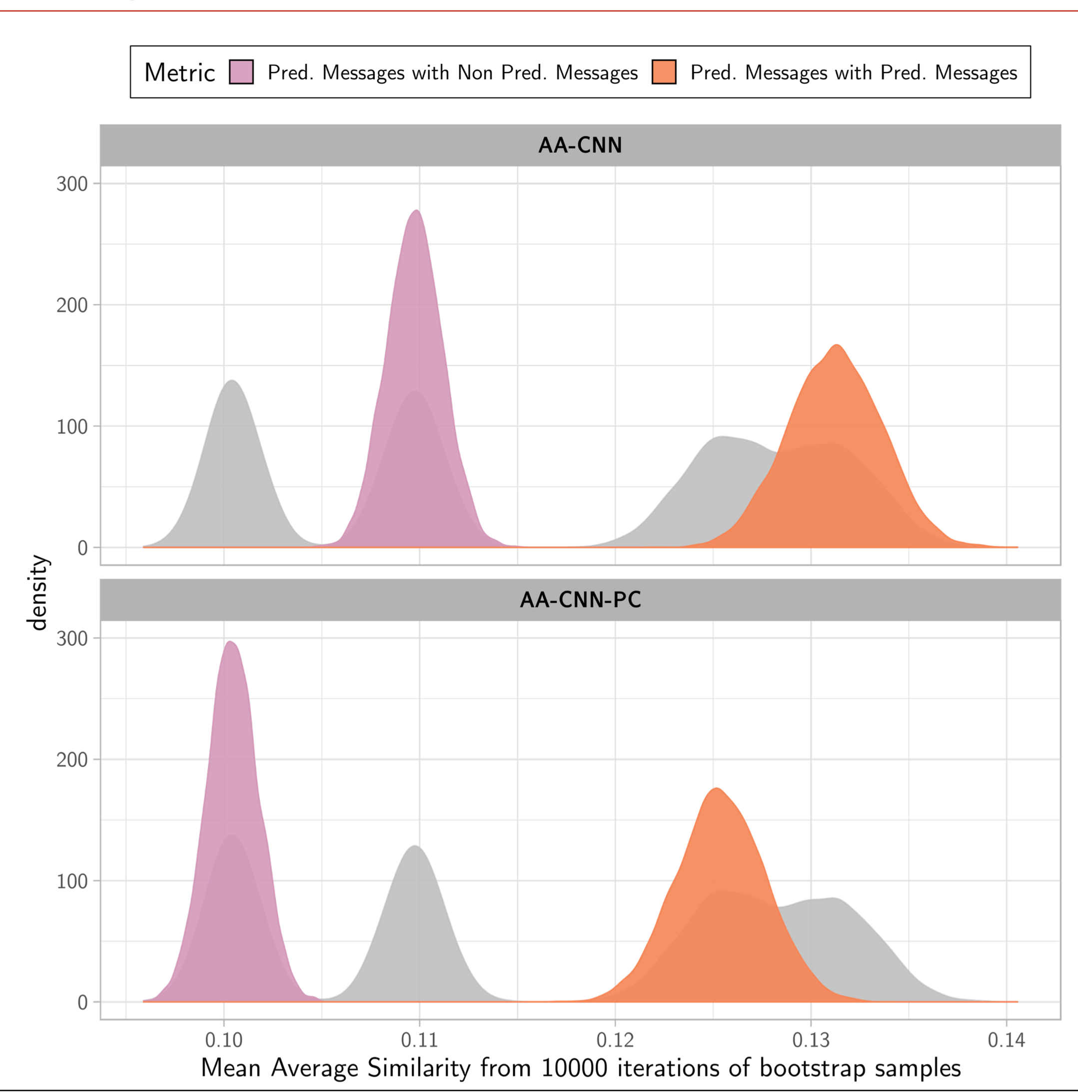
Corpus:

- Consists of messages from predators and non predators.
- Collected from PAN 2012, and Perverted Justice.
- A set of 50 authors is randomly selected with a train-dev-test split of 400-100-100.

Model:

- Utilizes both unigram and bigram signals using a Convolution Neural Network.
- AA-CNN-PC consists of an auxiliary layer to classify predators.
- Hyperparameters:
  - Feature Map size = 100
  - Kernel size = [3, 4, 5]
  - Character n-gram embedding dimension = 100
  - Dropout after Embedding layer = 0.5
- Trained for 50 epochs with a minibatch size of 32, using the Adam gradient method.
- Loss function = Negative Log Likelihood for both models.
- Loss function for the AA-CNN-PC: $L_{final} = L_{AA} + L_{PC}$

Analysis of Encoded Message Style



MAS for predators with predators and non predators (resampled)

t-SNE projections of messages in AA-CNN

References

- Ruder, S., Ghaffari, P., & Breslin, J. G. (2016). Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. arXiv preprint arXiv:1609.06686.
- Schwartz, R., Tsur, O., Rappoport, A., & Koppel, M. (2013). Authorship attribution of micro-messages. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1880-1891).
- Shrestha, P., Sierra, S., Gonzalez, F., Montes, M., Rosso, P., & Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (Vol. 2, pp. 669-674).

Acknowledgements

This project would not be possible without the financial backing of the Dept. of CIT and Purdue Research Foundation. Many thanks to Tatiana Ringenberg and the rest of the AKRaNLU lab for their feedback.

PURDUE UNIVERSITY Discovery Park

CERIAS