Authorship Analysis of Online Predatory Conversations using Character Level Convolution Neural Networks

Conference Paper · October 2019

DOI: 10.1109/SMC.2019.8914323

CITATIONS
0

READS
77

4 authors:

Kanishka Misra
Purdue University
6 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Hemanth Devarapalli
Purdue University
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Tatiana Ringenberg
Purdue University
9 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Julia M Rayz
Purdue University
100 PUBLICATIONS **680** CITATIONS
SEE PROFILE

2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)
Bari, Italy. October 6-9, 2019

Authorship Analysis of Online Predatory Conversations using Character Level Convolution Neural Networks

Kanishka Misra, Hemanth Devarapalli, Tatiana R. Ringenberg, and Julia Taylor Rayz
Department of Computer and Information Technology, Purdue University, USA

***Abstract*— Authorship Attribution (AA) of written content presents several advantages within the digital forensics domain. While AA has been traditionally applied to long documents, recent works have shown improved performance of neural AA models on short texts such as tweets and online conversations 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 an authorship attribution model that trains on a corpus of online conversations involving predators, and perform subsequent analysis of the message representations. Our results show comparable performance relative to prior work for Authorship Attribution and highlight differences between predatory and non-predatory message styles.**

I. Introduction

Authorship Attribution (AA) is a task of assigning an author to a text whose author is unknown. An author is selected from a set of candidate authors, given their known texts as samples to learn their style [26, 10]. This task has been successful in attributing authors of text samples at the document level, such as blog posts [15], fan-fiction stories [11], written pieces of literary text [16], books [17], etc. More recently, with the advent of social media and the age of short texts, such as text messages and tweets, research has ventured into the authorship analysis of texts at the sentence or phrase level [23, 25].

Apart from having direct applications within the realm of digital forensics and plagiarism detection, authorship analysis models have also been found to extract useful information regarding the authors themselves, such as age or gender [13]. By utilizing this feature of authorship analysis models, we examine their application in the domain of online chat conversations, specifically those involving online predators, and present an analysis of messages that have been posted by predators and non-predators as encoded by the models.

In this paper, we present a modified application of Authorship Attribution models, specifically those that are trained to learn the writing styles of predatory users from their messages as found in online chat conversations. To that end, we exploit recently proposed Neural Network models that have been shown to perform well in Authorship Attribution of short texts [21, 25] and analyze the difference in the encoded representations of predatory and non-predatory messages.

The research setting of this paper consists of several conversations, partly of predatory nature, where a predator talks to a decoy who is pretending to be a minor. In prior works, tasks focusing on a similar setting have aimed at (1) analyzing the linguistic properties of predators, decoys and real victims [8, 4]; (2) predicting whether a given user is a predator or a non-predator[9, 2, 6, 5]; (3) Identifying the specific lines that may contain predatory content [9]; and (4) differentiating between offenders that wish to physically meet their victims vs. keep the interaction to an online setting [20]. The work presented here deviates from predicting the user type, or word-choice based analysis of predatory users, and focuses on analyzing the difference between the messages as encoded by authorship attribution models.

Briefly, recent work in the authorship attribution of short texts have shown convolution neural networks (CNN) trained on character level units to outperform the state of the art methods [21, 25]. Building on previous work, in this paper, we propose modified variants of the CNN models: (1) AA-CNN, that utilizes unigram and bigram vector representations of characters to attribute each individual message to its author; and (1) `AA-CNN`, that utilizes unigram and bigram vector representations of characters to attribute each individual message to its author; and (2) `AA-CNN-PC`, that jointly learns the objective of (1) but also contains an auxiliary layer to learn whether the user is a predator or not, using the same weights that encode the message style.

Specifically, we are trying to address the following research questions:

1) Whether the newly proposed models, that learn from multiple character n-gram signals produces comparable performance compared to its predecessors.
2) Whether the encoded representations of the messages from a predator differ from those by non-predators, i.e., whether the `AA-CNN` model implicitly learns predatory style along with author style, or does it have to be jointly trained with explicit predatory signal (`AA-CNN-PC`).

The contributions of this work are summarized as follows:

- We present two models for authorship attribution of short texts, one that encodes style of individual messages using a combination of features inspired from previous work; and one that jointly learns style as well as the type of the author (predatory vs non-predatory).
- We present an analysis to test whether style based CNN trained on messages encode the differences between the various types of authors in the conversations (predatory vs non-predatory).

The rest of the paper is organized as follows: Section II introduces prior work relevant to the topic. Section III describes our corpus, Section IV describes the model as well

978-1-7281-4569-3/19/$31.00 ©2019 IEEE

as the baselines established in prior. Then, in Section V, we present the results of our models compared to the baselines and test whether our model implicitly learns to differentiate between predators and non-predators (RQ2), we compare the message representations from `AA-CNN` and `AA-CNN-PC` by measuring the difference between encoded messages from predators and non-predators and assess these differences using statistical tests.

II. BACKGROUND

A. Authorship Analysis

A typical approach that has produced promising results for Authorship Attribution is the use of stylistic features such as character n-grams and function words [26, 1], or syntactic features such as part of speech tags [14], along with a machine learning algorithm, such as a Support Vector Machines classifier. While several variations of features as well as machine learning classifiers have been successful, this success has mostly been in long documents. Performance of traditional Authorship Attribution methods for short texts has been shown to decline as the text length grows smaller [7].

B. Authorship Attribution of Short Texts

To circumvent this problem, several solutions have been proposed. A simple workaround for making a text artificially longer is treating several short texts from the same author as one long one. Merging short texts, such as tweets, into longer documents [3, 19] results in higher performance gains. Schwartz et al. [23] introduce the concept of k-signatures, which builds the unique profile of the author by taking all the features that appear in the authors texts at least k% of the time while not appearing in any other authors texts during training. They achieve state of the art results in their extensive experiments on tweets, a popular category of short texts.Inches et al. [9] present a measurement of a users unique profile by measuring the Kullback-Leibler Divergence between each users vocabulary and the vocabulary of their interlocutors in the chat. The metric is used to differentiate between users that have conversations with more than one other user.

While these methods achieve favorable results in the authorship attribution of short text documents, they still require manual feature selection (as in the case of k-signatures and user profiles), or concatenation, all of which are done by taking each user's entire known texts into account.

To perform authorship attribution on shot texts by automatically and efficiently learning feature representations, recent work has relied on using Neural Network models. Specifically, Sari et al. [22] represent a document by a continuous bag of character ngrams and learn dense representations for the character ngrams which are then given as input into a linear classifier with a softmax layer to achieve improved performance in Authorship Attribution for two out of five different corpora. Ruder et al [21] replicate the multi-channel convolution neural network [12] and experimented on a variety of web-scale data, including tweets and achieve state of the art performance at the time by using a character-level non-static channel (0.87 F1 score on tweets with 50 authors). Shrestha et al. [25] propose two very large character level CNN models (large in terms of filter size) that extract unigram and bigram level information, respectively, from tweets and outperform word level CNNs, SVM based models, and LSTMs in terms of authorship attribution (0.76 F1 score on tweets with 50 authors, separate corpus from Ruder et al). We use Ruder et al. [21] and Shrestha et al [25] as our baselines. All these models compared their results with traditional methods such as SVMs over word and character n-grams and reported improvements over them.

C. Conversations Involving a Predator

Within the domain of chat conversations that involve predator, extensive research has been done in order to analyze important linguistic differences between predators, teenagers, and decoys [8, 4]. Machine learning models have been trained on conversations involving a predator and a decoy (pretending to be a minor) to detect whether a given user is a predator or not. This was the major focus of the PAN-2012 competition [9], where conversations with predators and non-predators were combined to form a large corpus of 2 million messages and participating teams were asked to detect the predatory users. While this was successfully solved by using a binary classifier, the second task, that of detecting specific lines containing explicit predatory content led to poor results, overall. The best performing model for the second task used all the predicted predatory lines, resulting a massive loss of Precision (0.09) but a high recall score (.89). We hypothesize that this is because the traditional models were unable to extract useful signals at the line level due to the sparsity of their feature matrix. Finally, Ringenberg et al. [20], and Seigfried-Spellar et al. [24] differentiated between contact and fantasy offenders using a chat conversations from Perverted Justice, an online repository of chat conversations between predators and volunteer decoys.

The present work deviates from the rest of the studies conducted on predatory conversations, since we attempt to encode predator and non-predatory style in terms of an authorship attribution model. Specifically, we probe these models in their ability to not only differentiate between authors, but also between the type of authors, i.e., whether the author is a predator or not in terms of their style.

III. CORPUS

The corpus used in this work is built by combining conversations from Perverted Justice (PJ), which hosts 623 chat logs of several volunteer decoys and their conversations with online predators, and the PAN-2012 Sexual Predator Identification corpus [9], which consists of messages involving a predator (crawled from PJ), and several conversations from IRC chat logs. We combine all predator and decoy authors from PJ with all the non-predatory conversations from PAN, with restrictions that the message length in consideration should be between 3 and 200 words, as well as that an author should have at least 600 messages across all

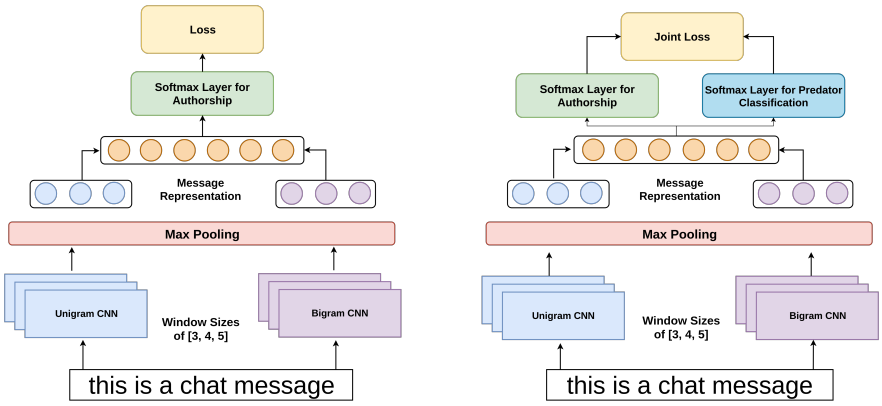

Fig. 1: Proposed Neural Network Architectures, AA−CNN (left) and AA−CNN−PC (right)

the conversations that they take part in. This leaves us with 345 total authors. We randomly sample sets of 10 authors and 50 authors to remain consistent with the analysis in [21, 25]. All html, email or URL strings were discarded from the messages.

IV. METHODOLOGY

We propose two models, AA−CNN and AA−CNN−PC, both of which are Convolution Neural Networks (CNN) that accept character unigram and bigram input features to encode each message's representation. Then, depending on the task (Authorship Attribution or Predator Classification), the encoded vector is passed on to fully connected layers to produce the output. Fig 1. illustrates the architectures of both these models.

A. Continuous Character n-gram Embeddings

Character based features have been shown to be extremely robust in Authorship Attribution models [26]. Traditionally, these have been used as a discrete bag-of-feature representations, while more recent work has found continuous representations of characters and character n-grams, to have performance gains in Authorship Attribution [22, 25]. We choose to use character unigram and character bigrams for our model, inpired by a combination of Ruder et al. [21] and Shrestha et al. [25]. The convolution over unigrams will produce short length windows (as used by Shrestha et al. [25]), and over bigrams will use longer windows (as used by Ruder et al. [21]) to encode the full message.

B. Character Level Convolution Neural Network

The input to the CNN is a text that is padded to a fixed character length (the maximum length across the corpus), n, represented by $x_{i:n}$, where $x_i \in \mathbb{R}^{1\times k}$ is the k dimensional vector of the i^{th} character-ngram in the text (either a unigram, or a bigram).

The convolution layer slides over the input text with different window sizes, $h = [h_1, h_2, ..h_l]$, and applies filters weights, $w \in \mathbb{R}^{1\times h_i.k}$ for each window size h_i. Each filter generates a new feature c_i using the following operation:

$$c_i = g(w.x_{i:i+h-1} + b) \quad (1)$$

Where $g(.)$ is a non linearity, in our case being a ReLU operation, and $b \in \mathbb{R}$ is the bias term present in each of the layers. After a convolution over all the windows, the features are concatenated to form a feature map for each filter: $c_{filter} = [c_1; c_2; c_3, ; ..c_{n-h_i+1}]$

Using the max-over-time pooling function, all m feature maps are reduced and concatenated to a vector $p \in \mathbb{R}^{1\times m}$. The pooled vectors for all l windows are concatenated to get the final output of the convolutional layers: $s \in \mathbb{R}^{1\times lm}$. Since we use two such convolution layers - one each for unigram and bigram, we concatenate the final output of both layers to get a fixed size vector $v \in \mathbb{R}^{1\times 2lm}$. This fixed size vector is now given as input to a softmax layer for classifying the author of the input text. The model is trained to minimize the cross-entropy loss.

C. Auxiliary Layer for Predator Classification

To extract style based information about predatory users, while also maintaining performance of the Authorship Attribution model, we adjust the AA−CNN model and add an auxiliary layer to serve as a classifier for detecting predators, resulting in the AA−CNN−PC model. Since the auxiliary layer is jointly trained with the main Authorship Attribution layer, the final loss for the AA−CNN−PC model is defined as:

$$L_{final} = L_{AA} + L_{PC} \quad (2)$$

where L_{AA} is the loss for the authorship attribution layer (same as in AA−CNN) and L_{PC} is the loss for the auxiliary Predator Classification layer (binary cross entropy loss). While training, we want the model to attribute the input message to its author while also explicitly learning about whether the author is a predator or not, as opposed to AA−CNN, where if the model does learn about the author being a predator or not, it will be implicit.

D. Model Training

For our models, we set the embedding dimension $k = 100$, three windows $h = [3, 4, 5]$, and $m = 100$ filters. This yields a final vector $v \in \mathbb{R}^{1\times 600}$. The max length, n is inferred from corpus and is found to be 200. We apply a dropout of probability 0.5 after the embedding layer, and train over

mini-batches of size 32 for 50 epochs with Adam as the optimizer with a learning rate of 0.001.

E. Baseline Models

For comparison, we select two network architectures which reported state of the art results in Authorship Attribution. The papers were selected based on how the text was processed, which can be the whole document at once, or sentence at a time. Our proposed network architecture works sentence at a time, and hence we chose papers which process text similarly. The selected papers work with short texts, tweets in particular.

Ruder et al.[21] proposed character level CNNs for Authorship Attribution that achieved the best results on tweets, emails and reddit comments with 10 and 50 authors. They use a network that processes single characters from the input that are randomly initialized with a 300-dimensional embedding layer. They use window sizes of 6,7, and 8 and a feature map size of 100. They apply a dropout of 0.5 and train their network for 15 epochs, with a Stochastic Gradient Descent with the Adadelta update rule and a learning rate of 0.001.

Shrestha et al. [25] proposed a variant of Ruder's [21] architecture which achieved state of the art on tweets with 50 authors. The proposed architecture work with bigrams instead of single character level embeddings. In their case, the embedding size is 300, window sizes are 3, 4, and 5, and the feature map size is 500. They use a dropout of 0.25 probability and train their network for 100 epochs using the Adam optimizer with a learning rate of 0.0001.

V. EXPERIMENTS

We experiment on two sets of data: (1) A set of 10 randomly selected authors, with 5 being predators and 5 being non-predators; and (2) A set of 50 randomly selected authors, with 25 being predators and 25 being non-predators. We split our corpus into a set of 400, 100, and 100 sample messages for each author as the training, development, and testing set. Finally, we train our models and propose the following two experiments.

A. Performance of AA-CNN and AA-CNN-PC models

The performance of our two models is compared to the baselines - described in the previous section. We train the baselines as stated in the papers that introduced them and present results on the 10 and 50 author sets in Table 1. We use the micro-averaged F_1 score as our comparison metric. Some differences between differences between each of the models in terms of the architectures (the feature embedding size, as well as the feature map size) are highlighted in Table 1 as well.

1) Results: The results from Table 1 indicate comparable performance of our models as compared to the baselines. For the 10 authors set, both our models resulted in a slightly lower F1 score than one of the baselines and a higher F1 score than another. For the 50 authors set, our AA-CNN-PC model outperforms both baselines. Interestingly, the AA-CNN shows improved performance for 10 authors compared to the AA-CNN-PC model, while for the 50-author set, the opposite happens. Both models show that the difference between the performance of 10 and 50 authors is smaller than that of the baselines.

TABLE I: MICRO-AVERAGED F_1 SCORES OF AA-CNN AND AA-CNN-PC COMPARED TO THE BASELINES FOR 10 AND 50 AUTHOR SETS

Model	Model Architecture		10 Authors	50 Authors
Ruder et al., 2016	Emb. size	300	0.5250	0.3524
	Feature maps	100		
Shrestha et al., 2017	Emb. size	300	**0.5880**	0.4474
	Feature maps	500		
Ours (AA-CNN)	Emb. size	100×2	0.5570	0.4382
	Feature maps	100		
Ours (AA-CNN-PC)	Emb. size	100×2	0.5490	**0.4484**
	Feature maps	100		

2) Discussion: The results from comparisons to our baselines indicate similar performance of our models to the state of the art in Authorship Attribution of short texts. While we are unable to outperform one of the baselines in the 10-author set, we do outperform them in the 50-author set, but by a very low margin. However, our model architecture is much simpler than that of Shrestha et al. [25] we use a third of the embedding size they used and a fifth of the feature map size as compared to them. These parameters were set so that we can easily add an auxiliary layer in AA-CNN-PC without altering the batch-size and the number of epochs in order to train at the same speed, without any memory issues.

Within our models, we find AA-CNN to beat the jointly trained AA-CNN-PC model for the 10-author set, but the opposite happens in the 50-author set experiments. We hypothesize that the AA-CNN-PC model benefits from observing differences between predators and non-predators in the 50-author set, since it trains on 5 times more examples than in the case of the 10-author set, where the difference is less pronounced. Due to the lack of sufficient examples, the AA-CNN-PC might be overfitting in the predator classification task, and in turn experiencing a slight increase in the Authorship Attribution loss in the 10-author set. This effect would be less when it gets more examples to train on. We leave this exploration of joint-training loss as something to be covered in future work.

B. Probing message representations for differences between Predators and non-Predators

Both our proposed models learn to encode chat messages using their character unigram and bigram embeddings and use this representation to attribute each message to its author. While the AA-CNN only serves as an authorship attribution

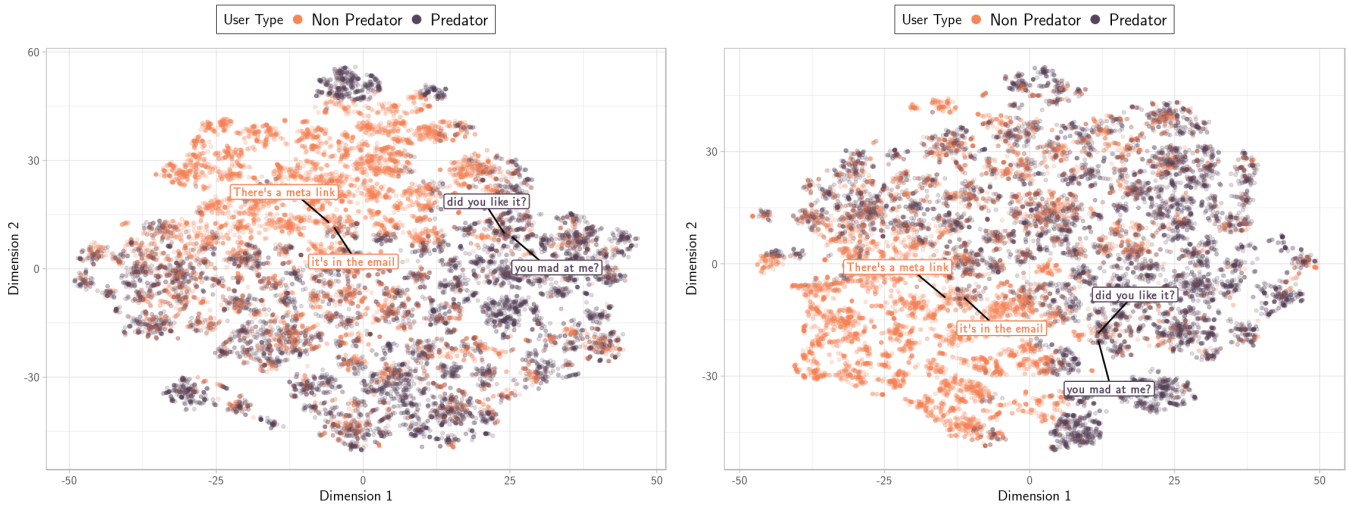

Fig. 2: t-SNE visualization of message representations in the training set, extracted AA-CNN model (left) and AA-CNN-PC (right). Four example messages have been shown, two each from the predator and non-predator groups.

model, the AA−CNN−PC model jointly trains to encode author style as well as the type of author (predator vs. non-predator) using its auxiliary binary classification layer. In this section, we analyze the differences in the encoded message representations in both the models. Specifically, we are interested in probing the message representations in order to learn whether the AA−CNN model implicitly extracts signals for differences between predators and non-predators, as compared to the AA−CNN−PC model (which we expect, learns to do this explicitly, due to the additional layer). In order to compare message representations in both models, we first extract the final representations from the CNN layers of the trained models and separate them into two groups (predator vs non-predator) using the ground truth labels. Then, in order to measure the differences between the representations of the groups, we propose a metric, the Mean Average Similarity (MAS), which in our case, will measure the similarity of predatory messages to either other predatory messages, or to non-predatory messages. Mathematically, for message vectors v_i^a and v_j^b , belonging to group A (of length N_i) and B (of length N_j) respectively, the MAS is given as:

$$MAS(v_i^a, v_j^b) = \frac{1}{N_i}\frac{1}{N_j}\sum_{i}^{N_i}\sum_{j}^{N_j} cos(v_i^a, v_j^b) \quad (3)$$

The difference between the two groups in our case is further measured as the change in the MAS (ΔMAS) between the message representations of predators and that of non-predators. It is computed as:

$$\Delta MAS = MAS(v_i^{predator}, v_j^{predator})1_{i\neq j} - MAS(v_i^{predator}, v_j^{non-predator}) \quad (4)$$

This metric measures the difference between the similarity of predatory messages to each other compared to non-predatory messages. For 10000 iterations, we randomly sample 1000 predatory and 1000 non-predatory messages and compute the ΔMAS for each iteration. Then, we conduct a one-sided t-test to measure the significance of MAS for both the AA-CNN, as well as AA−CNN−PC models. The results are summarized in Table II.

TABLE II: CHANGE IN THE MEAN AVERAGE SIMILARITY (PREDATOR - NON PREDATOR) IN AA−CNN AS WELL AS AA−CNN−PC FOR 2000 RESAMPLES OVER 10000 ITERATIONS.

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

1) Results: The results shown in Table II indicate a statistically significant difference between the similarity of predator messages with other predatory messages, as compared to with non-predatory messages. This difference is greater in AA- CNN-PC as compared to the AA−CNN model.

2) Visualizing the encoded representations of messages: To visually highlight the differences between predatory and the non-predatory messages, we computed a t-SNE [18] visualization with default parameters for both the AA-CNN and the AA−CNN−PC model. The plots are shown in Fig. 2.

3) Discussion: From the results of this experiments, we find that both models learn the differences between messages from a predator and those from a non-predator. This is seen both in our statistical comparisons, as well as in the t-SNE projections (Fig. 2). While the AA−CNN model is never explicitly provided an information about whether the user of the message is a predator or not (as is in the case of AA-CNN-PC), it still encodes the differences, although not as much as in the model that jointly learns these differences. The AA−CNN−PC jointly adapts to encoding styles of the kind of author while performing the task of authorship attribution, and thus it is not surprising that the message

from predators were more closer together in this model as compared to the simple AA−CNN model.

VI. Conclusion

In this paper, we presented two models, one that is trained to do authorship attribution on a corpus of chat messages from predators and non-predators, and one that performs authorship attribution and jointly learns whether the author of the message is a predator or a non-predator. Our models were simpler yet comparable in terms of performance with our baselines. We also presented an analysis of the messages as encoded by our CNN models and found that the Authorship Attribution model was able to implicitly encode information about the kind of author without any explicit signal, and differentiated between predators and non-predators, while this difference was more pronounced in the model that explicitly received signal about the author type.

Predators present a risk to minors online and one way of reducing this risk is to develop tools to detect predatory behavior from evidence, usually found in the form of chat conversations. While the models in this work did show certain aspects of encoding predatory behavior purely from style (characters), in the future, we would like to work on combining this with more contextual models, such as sequence encoders. Using the context of a chat messages, we hope to build better models of detecting risk over the internet to aid in making it a better space for everyone.

Acknowledgements

This research was partially supported by Purdue Research Foundation.

References

[1] S. Argamon et al. “Stylistic text classification using functional lexical features”. In: *Journal of the American Society for Information Science and Technology* 58.6 (2007), pp. 802–822.

[2] D. Bogdanova, P. Rosso, and T. Solorio. “On the impact of sentiment and emotion based features in detecting online sexual predators”. In: *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*. Association for Computational Linguistics. 2012, pp. 110–118.

[3] S. R. Boutwell. “Authorship attribution of short messages using multimodal features”. PhD thesis. Monterey, California. Naval Postgraduate School, 2011.

[4] M. M. Chiu, K. C. Seigfried-Spellar, and T. R. Ringenberg. “Exploring detection of contact vs. fantasy online sexual offenders in chats with minors: Statistical discourse analysis of self-disclosure and emotion words”. In: *Child abuse & neglect* 81 (2018), pp. 128–138.

[5] M. Ebrahimi, C. Y. Suen, and O. Ormandjieva. “Detecting predatory conversations in social media by deep convolutional neural networks”. In: *Digital Investigation* 18 (2016), pp. 33–49.

[6] M. Ebrahimi et al. “Recognizing predatory chat documents using semi-supervised anomaly detection”. In: *Electronic Imaging* 2016.17 (2016), pp. 1–9.

[7] M. Eder. “Does size matter? Authorship attribution, small samples, big problem”. In: *Digital Scholarship in the Humanities* 30.2 (Nov. 2013), pp. 167–182. ISSN: 2055-7671.

[8] K. Guice. *Predators, decoys, and teens: A corpus analysis of online language*. Hofstra University, 2016.

[9] G. Inches and F. Crestani. “Overview of the International Sexual Predator Identification Competition at PAN-2012.” In: *CLEF (Online working notes/labs/workshop)*. Vol. 30. 2012.

[10] P. Juola et al. “Authorship attribution”. In: *Foundations and Trends® in Information Retrieval* 1.3 (2008), pp. 233–334.

[11] M. Kestemont et al. “Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection”. In: *Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al.* 2018, pp. 1–25.

[12] Y. Kim. “Convolutional neural networks for sentence classification”. In: *arXiv preprint arXiv:1408.5882* (2014).

[13] M. Koppel, S. Argamon, and A. R. Shimoni. “Automatically categorizing written texts by author gender”. In: *Literary and linguistic computing* 17.4 (2002), pp. 401–412.

[14] M. Koppel and J. Schler. “Exploiting stylistic idiosyncrasies for authorship attribution”. In: *Proceedings of IJCAI’03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Vol. 69. 2003, pp. 72–80.

[15] M. Koppel, J. Schler, and S. Argamon. “Authorship attribution in the wild”. In: *Language Resources and Evaluation* 45.1 (2011), pp. 83–94.

[16] M. Koppel, J. Schler, and E. Bonchek-Dokow. “Measuring differentiability: Unmasking pseudonymous authors”. In: *Journal of Machine Learning Research* 8.Jun (2007), pp. 1261–1276.

[17] M. Koppel et al. “Unsupervised decomposition of a document into authorial components”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics. 2011, pp. 1356–1364.

[18] L. v. d. Maaten and G. Hinton. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[19] G. K. Mikros and K. Perifanos. “Authorship attribution in greek tweets using author’s multilevel n-gram profiles”. In: *2013 AAAI Spring Symposium Series*. 2013.

[20] Tatiana Ringenberg et al. “Exploring Automatic Identification of Fantasy-Driven and Contact-Driven Sexual Solicitors”. In: *2019 Third IEEE International Conference on Robotic Computing (IRC)*. IEEE. 2019, pp. 532–537.

[21] S. Ruder, P. Ghaffari, and J. G. Breslin. “Character-level and multi-channel convolutional neural networks for large-scale authorship attribution”. In: *arXiv preprint arXiv:1609.06686* (2016).

[22] Y. Sari, A. Vlachos, and M. Stevenson. “Continuous n-gram representations for authorship attribution”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Vol. 2. 2017, pp. 267–273.

[23] R. Schwartz et al. “Authorship attribution of micromessages”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1880–1891.

[24] K. C. Seigfried-Spellar et al. “Chat Analysis Triage Tool: Differentiating Contact-Driven vs. Fantasy-Driven Child Sex Offenders”. In: *Forensic science international* (2019).

[25] P. Shrestha et al. “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. 2017, pp. 669–674.

[26] E. Stamatatos. “A survey of modern authorship attribution methods”. In: *Journal of the American Society for information Science and Technology* 60.3 (2009), pp. 538–556.