Authorship Detection in Dark Web Marketplaces using LSTM and RNN Neural Networks

Jacqueline Garrahan
Department of Mathematics
Northeastern University
Boston, MA 02115
garrahan.j@husky.neu.edu

Abstract

This project aims to apply LSTM and RNN neural networks to authorship classification in dark web forum posts. By using passages represented by GloVe vectors, the neural networks are trained on semantic data inputs. The findings are suggestive of LSTM's utility in authorship classification tasks, and cast reasonable doubt on anonymity in forums with verbal communications.

1 Introduction

The advent of the Silk Road in 2011 marked a movement to develop extensive infrastructure for dark web marketplaces [1]. These platforms, characterized by deep libertarian commitments to free-market principles, have become a haven for drug traffickers and psychonauts. Marketplace sites are hosted on Tor's Onion cloud, and may only be accessed through the Tor browser, which enforces encrypted relays on network traffic [2]. Because of Tor's systematic controls around network routing, anonymity remains a challenge for law enforcement. Officials are often reliant upon user networking mistakes or warrants for server access to apprehending suspects. High profile cases include the arrest Silk Road creator and administrator, Ross Ulbricht, who was sentenced to life in prison without parole in 2015 after authorities tracked his VPN server access logs [3].

This project aims to explore anonymity in the Silk Road forums by exploiting the inherent uniqueness of communication. In the forums, users are able to interact using registered accounts. Though usernames may be considered unique user identifiers due to registration, users may maintain many accounts for the purpose of obfuscation. Predictions about authorship may contribute to the deanonymization of users under this framework. Using a limited number of short posts, we show that it is possible to make predictions about authorship.

Prior work by Yao and Liu, explored authorship detection using recurrent neural networks for the classification of passages from literary works by distinct authors [4]. This project adapts their testing framework to accommodate the casual conversation style of forum communications by using pre-trained Twitter GloVe data [5]. In addition to exploring RNN, we also explored the performance of a LSTM layer on the classification task. The resulting neural network is composed of one hidden layer, a dropout filter, and a LSTM layer. We explore the model using cross validation on dropout proportion, optimization, and hidden layer size and our resulting LSTM model yields 0.985 testing and 0.488 validation accuracy measures.

2 Data

Our data was taken from a web crawl of the Silk Road marketplace conducted in October of 2013 by independent researcher Gwern Branwen [6]. We chose a subset of forum data consisting of 2199 pages representing 4200 distinct usernames. The html was parsed to extract unique post content (quotations removed, etc.) and all username unique content was concatenated. Words not represented in the Twitter GloVe vocabulary set and stop words were removed from the concatenated text body. The text was then subdivided into 18 word passages and each word in the passage was replaced by its GloVe vector representation. Each passage was ultimately represented by an 18×200 matrix. We reduced our data space by imposing a threshold minimum of 250 passages for author inclusion. Our resulting set consisted of 10 unique authors with a total of 4283 passages.

3 GloVe vectors

GloVe vectors are semantic vector space models, which represent words with real numeric values [5]. The GloVe model is unique in semantic vectorization schemes because it utilizes global matrix factorization alongside local context. This is accomplished by training a least squares model on a vocabulary co-occurance matrix X. The conditional probability $P_{i,j} = P(j|i) = \frac{X_{i,j}}{X_i}$ is used to construct vectors of low dimensionality via factorization.

For this model, we chose to use pre-trained Twitter GloVe data to capture the semantic properties of passages. The Twitter GloVe data is most appropriate for this application because of the heavy colloquial vocabulary used in the forum setting. Because prior research has shown that larger GloVe vectors are able to capture more semantic information, we chose to use the largest Twitter vector, with size 200 [5]. From our passages, we created a vocabulary set consisting of all represented words. An embedding matrix was created for our model using the Twitter GloVe vector representations of the vocabulary space. The resulting matrix consisted of 1,695,940 non-trainable parameters.

4 Model

The neural network was implemented with a single hidden layer, a dropout layer, and a RNN or LSTM layer. The dropout layer was included before the RNN or LSTM layer in order to prevent overfitting of reinforced parameters. Our matrix representations of each passage, x, with GloVe vectors $x_i, i \in \{1...18\}$were fed into the neural network. The result is output, y, with nonzero y_i corresponding to author classification i.

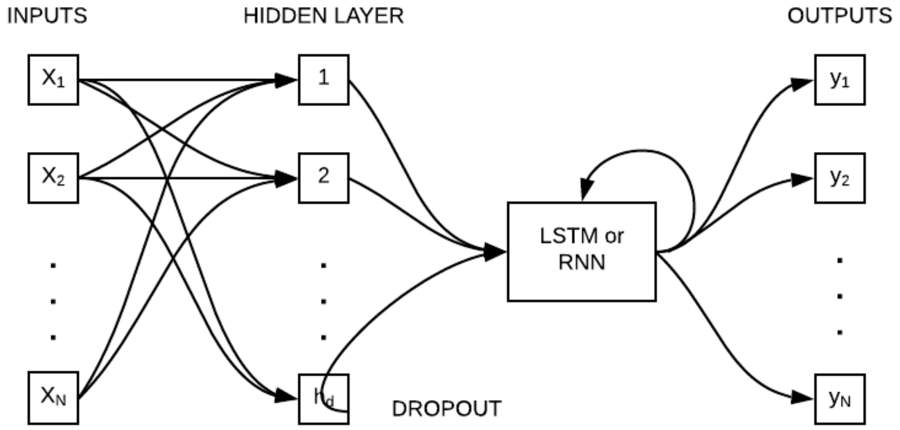


Figure 1: Model framework.

5 Cross validation

We used 5-fold cross validation to select the appropriate parameters for the model. For the purpose of this exploration, we chose to reduce the cross validation space by fixing activation and loss parameters.

We used softmax activation and categorical cross-entropy for the loss function. Because this was an unique authorship classification task, the categorical cross-entropy appropriately allowed for one-hot categorical classification. Our author labels were converted to length 10 categorical vectors.
We first conducted cross validation on our LSTM model, testing hidden layer dimension, optimization, and dropout. Next, we tested the RNN model on hidden layer dimension and dropout, using the best performing optimization from the LSTM model due to time constraints. Each condition was trained over 100 epochs.

5.1 Hidden layer size

For hidden dimension, we explored hidden dimensions of $h_d = 50$ and $h_d = 100$. The resulting layers consisted of 10,050 and 20,100 parameters, respectively.

5.2 Optimizer

We considered three different optimizers for the model: Adam, Stochastic Gradient descent, and Adagrad. [7]

Adam:

Adam (Adaptive Moment Estimation) uses adaptive learning for each parameter by using past gradients and squared gradients:

$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

With bias-corrected moment estimates:

$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

Giving the update rule:

$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t$

Stochastic Gradient Descent:

Stochastic gradient descent updates parameters for all training examples x_i with labels y_i, using update rule:

$\theta = \theta - \eta \nabla_\theta J(\theta; x_i; y_i)$

Adagrad:

Adagrad adapts the learning rate to parameters (θ_i by using large updates for infrequent parameters and small updates for frequent parameters. The gradient at time t is given by:

$g_{t,i} = \nabla_{\theta_t} J(\theta_{t,i})$

The parameter update uses diagonal matrix G_t to update each parameter at time t. Entry $G_{ti,i}$ gives the sum of the squares of the prior gradients of θ_i :
$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} g_{t,i}$

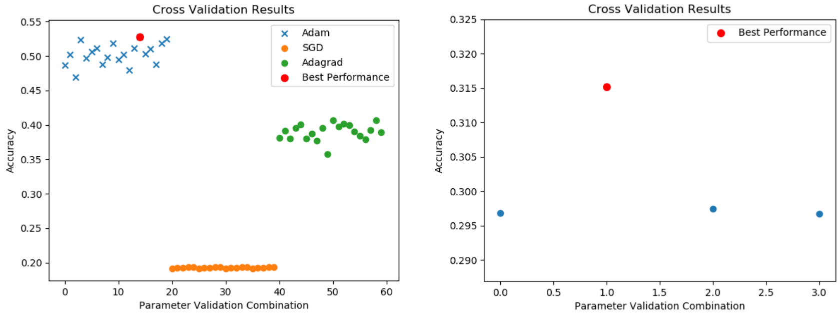
5.3 Dropout

Prior work has shown that neural nets with a large number of parameters may be prone to overfitting and dropout may help to mitigate excessive co-adaption [8]. For this reason, we chose to include a dropout cell in the neural network. We tested two dropout proportions, 0.1 and 0.2. [8]

6 Results

6.1 Cross validation

Our cross validation on the LSTM model results showed that the Adam optimizer performed significantly better than Stochastic Gradient Descent or Adagrad. The three optimization cases formed distinct clusters within the cross validation space. A dropout parameter of 0.1 and hidden layer size of 100 yielded the most accurate results.
For our RNN model,



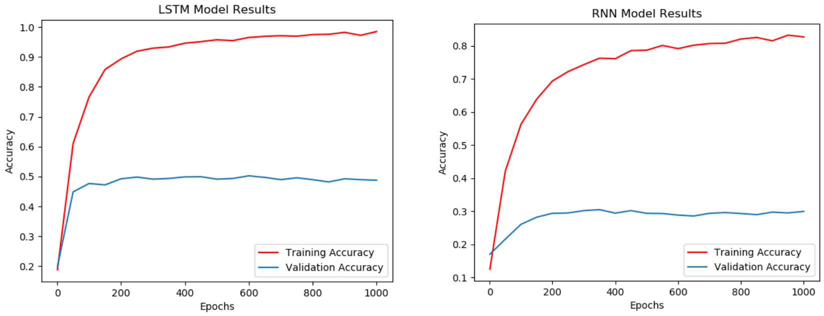
(a) LSTM cross validation results

(b) RNN cross validation results

Figure 2: 5-fold cross validation results.

6.2 Model

Using the best fit parameters from the cross validation trials, we ran the models for 1000 epochs with a 0.6/0.4 test/validation split. The LSTM model out-performed the RNN model. The LSTM model yielded 0.985 testing and 0.488 validation accuracy measures. The RNN model yielded 0.8264 testing and 0.2993 validation accuracy measures.



(a) LSTM

(b) RNN

Figure 3: Model results.

7 Conclusions

Our LSTM model was able to effectively characterize dark web market semantic data in order to generate authorship predictions, to roughly 0.5 accuracy. This LSTM implementation performed significantly better than the the RNN framework on the same classification task. This is an interesting contribution because prior work in neural networks for authorship detection has not focused on LSTM implementations.

Future areas of interest may include deeper evaluation of GloVe vector constructions. Twitter GloVe data has been used for classification tasks on Twitter; however, this is the first known initiative to apply the work to dark web markets [9]. Creating a GloVe data set trained on forum content may lead to better results. Another drawback to our use of GloVe semantic vectorization approach is the equal weighting of co-occurance weights, which gives equal stock to rare and frequent co-occurance pairs and may lead to noise [5]. Sufficient analysis of the forum data set vocabulary is needed to account for rare events.

Finally, because of feasibility limitations on the cross validation space, we were forced to limit our parameter exploration. This limitation led to edge cases yielding the best performance, which suggests that significant refinement of these parameters is possible.

Despite these future research needs, the current work successfully demonstrates the ability of semantic exchanges over anonymous forums to lend to identification and suggests that LSTM may perform well on authorship classification tasks. These findings will be of interest to both authorities and forum participants, to whom anonymity is of paramount concern.

References

[1] Van Hout, Marie Claire, and Tim Bingham. "'Silk Road', the Virtual Drug Marketplace: A Single Case Study of User Experiences." International Journal of Drug Policy 24, no. 5 (2013): 385-91. Accessed August 20, 2018. doi:10.1016/j.drugpo.2013.01.005.

[2] "Tor Overview." Tor. Accessed August 20, 2018. https://www.torproject.org/about/overview.html.en.

[3] Hume, Tim. "How the FBI Caught Ross Ulbricht, Alleged Creator of Silk Road." CNN. October 05, 2013. Accessed August 20, 2018. https://edition.cnn.com/2013/10/04/world/americas/silk-road-ross-ulbricht/index.html.

[4] Yao, Leon, and Derrick Liu. "Wallace: Author Detection via Recurrent Neural Networks."

[5] Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

[6] Gwern, Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Sohhlz, Delyan Kratunov, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. "Dark Net Market archives, 2011-2015". July 12, 2015. Web. Accessed August 19, 2018. https://www.gwern.net/DNM-archives https://www.gwern.net/DNM-archives.

[7] Ruder, Sebastian. "An overview of gradient descent optimization algorithms". September 15, 2016. Accessed August 20, 2018. arXiv:1609.04747.

[8] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." The Journal of Machine Learning Research 15, no. 1 (2014): 1929-1958.

[9] Wang, Chen-Kai, Onkar Singh, Zhao-Li Tang, and Hong-Jie Dai. "Using a Recurrent Neural Network Model for Classification of Tweets Conveyed Influenza-related Information." In Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017), pp. 33-38. 2017.