Shakespeare and Satoshi - De-anonymizing Writing Using BiLSTMs with Attention

Varun Ramesh
vramesh2@stanford.edu

Jean-Luc Watson
jlwatson@stanford.edu

Abstract

We apply bidirectional LSTMs with attention to the problem of attributing anonymous and pseudonymous works. We focus on two instances of the problem - the attribution of anonymous plays believed to have been written by Shakespeare, and the identification of Satoshi Nakamoto, the creator of Bitcoin. In the case of Shakespeare, we manage a test accuracy of 91.95% on a hold-out dataset, but found our results on disputed works to be mixed, with some successes. In the case of Satoshi, we manage a test accuracy of 85.94%, but have no conclusive result on Satoshi Nakamoto's identity. Our findings cast doubt on the successes touted by previous papers and projects - authorship classification on labeled data does not inherently generalize to unlabeled data.

1 Introduction

1.1 Stylometry

Authorship of a work may be unknown for several reasons. Authors publish works anonymously or pseudonymously to avoid retribution or unwanted attention. In historical cases, records may be incomplete, destroyed, or conflicting; thus the true authorship remains in dispute. Furthermore, authorship may be intentionally misrepresented to increase sales or attention towards a published work, thus misinforming future observers who lack the social context of the time.

Stylometry, the study of an author's writing style, can be used to determine the original author of a disputed work. Today, stylometry typically relies on the use of computers and modern statistical techniques to compare disputed works with a corpus of well-known attributed works. Older techniques include naive Bayes classification and principal component analysis, but recent stylometric efforts have focused on neural networks.

Finally, as stylometric models develop, writers will be able to better anonymize their work. They can analyze their own anonymous writing as compared to their non-anonymous publications and adjust their writing accordingly - such models become a tool for incremental anonymization.

1.2 Shakespeare

Many plays published during the 17th century have been, sometimes for lack of a better known option, attributed to William Shakespeare. The well-known First Folio, consisting of 36 plays, is considered his only undisputed collection of works, while only a few plays outside of these are generally accepted to be authentically Shakespeare. For the other plays, however, considerable uncertainty exists, especially works authored and published anonymously during his lifetime, which have later been attributed to Shakespeare by scholars employing rudimentary stylometric techniques.

1.3 Satoshi Nakamoto

Satoshi Nakamoto is the creator of Bitcoin, a peer-to-peer digital cryptocurrency. Nakamoto sent emails and registered domains using anonymous services, so it is generally accepted that the moniker is a pseudonym. There is considerable interest in determining who Nakamoto is, and many publications have written high-profile articles about him/her [3].

2 Background/Related Work

The first application of neural networks to stylometry was Tweedie et. al [7], where a multi-layer perception was used to detect authorship of the anonymously published federalist papers. The results match with scholarly consensus. However, the authors used a simple feed-forward network, with the rates of special "function" words as the input. This requires special preprocessing and tuning in order to select the right "function" words.

Since then, RNNs have been applied to authorship attribution tasks [2]. Lin et. al uses bidirectional LSTMs with a mechanism known as self-attention to create feature vectors for several tasks, including author profiling [5] [1].

Prior CS224d projects have compared vanilla RNNs, forward LSTMs, Naive Bayes classifiers, and linear SVMs [6], [8] [9]. Another CS224n project compared GRUs, LSTMs, and Siamese Networks at both the article and sentence level [6].

3 Approach

3.1 Data

3.1.1 Shakespeare Dataset

Our Shakespeare dataset consists of 18 contemporary plays, including 9 works authored by William Shakespeare, 2 by Thomas Dekker, 1 manuscript written by Ben Jonson, 3 works by Thomas Middleton, and another 3 by Christopher Marlowe, each consisting of a full-text file found freely on the Internet. We semi-automatically sanitized the incoming text to remove any stage directions, act/scene titles, and most importantly, character names that might unduly influence the classifier. The raw downloads and the processed files can both be found in our code repository. After parsing, the dataset consists of over 46 thousand lines across all works.

We then segment the works into 75-word sequences before passing them to the classifier. This yields a total of 4,973 input examples, 2,873 of which are derived from Shakespeare's works and the other 2,101 from all of the other playwrights listed above. The number of examples can be increased given smaller sequences - we originally attempted 20-word sequences which yielded greater than 17 thousand examples, but at a detriment to classifier performance. The test and validation splits for the dataset are each 10%.

In addition to the 18 labeled works, we collected five "disputed" works on which to evaluate our model, listed below with the circumstances of their uncertain origin:

- *The Puritan*: originally attributed to William Shakespeare due to manuscripts listing "W.S." as the author, this play is now regarded to be the work of Thomas Middleton.
- *Thomas Lord Cromwell*: once again, originally attributed to Shakespeare after his death, but the style of writing is significantly different from any of his other works.
- *The Birth of Merlin*: verified to have been written after Shakespeare's death in 1616, the work is a good imitation of his style.
- *Pericles, Prince of Tyre*: interestingly, the first 2 acts are thought to have been written by other authors, but historians believe that Shakespeare was heavily involved in the development of the final 3 acts. In our experimentation (see Section 4), we evaluate each of the acts independently.

[1] Author profiling is a related task, where the demographics of a writer are determined through analysis of their written text.

- *Vortigern and Rowena*: Hailed as a lost Shakespearean work, this play was soon debunked as a hoax originating in Ireland. *Vortigern* is one of the most well-known of the Shakespeare apocrypha.

3.1.2 Satoshi Dataset

For the Satoshi dataset, we scraped emails, blog posts, essays, and forum profiles for Gavin Andresen, Roger Ver, Hal Finney, Jed McCaleb, Nick Szabo, Craig Steven Wright, and Wei Dai. These figures were selected because they have all been suspected of being Satoshi Nakamoto and have all written a significant amount of text in informal contexts similar to that of Satoshi Nakamoto's writing. Notably absent is Dorian Nakamoto, who was left out because we could only find 81 writings by him through his Reddit AMA.

Scrapers were developed for the BitcoinTalk forums, Reddit, blog feeds, comment feeds, Hacker News, the Cypherpunk Mailing list Archives, and various essays. The writings of Satoshi Nakamoto were downloaded from `http://nakamotoinstitute.org/`. Headers and signatures that would obviously identify sources were cleaned out. The final dataset consists of 10,493 documents. We filter out documents under 10 words, and split documents over 200 words into multiple sequences. This gives us 13,855 data examples. The test split and validation split are each 10% of the total data set.

3.2 Network Architecture

Our network architecture, shown in Figure 1, first consists of an embedding layer which converts integer input sequences into sequences of 128-dimensional vectors. We then use these vector sequences as inputs to both a forward and a backward LSTM, both of which have a 128-dimensional hidden state. Each of these LSTMs then feeds their sequences into an attention layer, described in section 3.3. The attention-augmented results for the forward and backward LSTMs are then concatenated, resulting in a 512-dimensional vector. This vector is then sent through a dense layer of 50-units and finally ends up in a dense layer with the class scores. In the case of Satoshi, the final dense layer has 7 units and uses a softmax activation function. The Shakespeare model has only one unit with a sigmoid activation function. Dropout for all layers is set to 0.5.

For the Shakespeare dataset we use binary cross-entropy loss, while the Satoshi dataset uses categorical cross-entropy. Both are trained using an Adam optimizer.

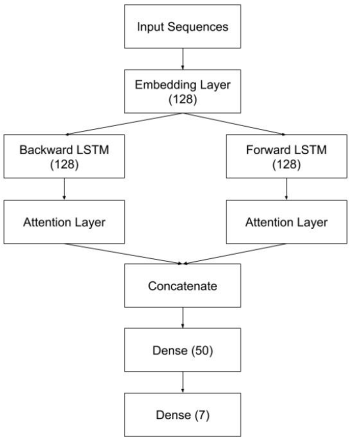


Figure 1: The architecture for our authorship classification model.

3.3 Attention

We implemented dot-product and multiplicative attention. For dot-product attention, we take the final hidden state of the LSTM and dot-product it with all of the previous hidden states. We then run a softmax function on those outputs. Finally, the previous hidden states are combined using the softmax outputs as weights, and this combined attention vector is concatenated to the original LSTM's final hidden state. [2]

Multiplicative attention follows the same process, but multiplies the prior hidden states by a weight matrix before calculating the dot-product with the final hidden state.

3.4 Implementation

We implemented our model in Python with Keras [4], using the Tensorflow backend [1]. We implemented attention mechanisms ourselves by creating a custom Keras layer. Input sequences are prepadded so that they are all the same length.

4 Results

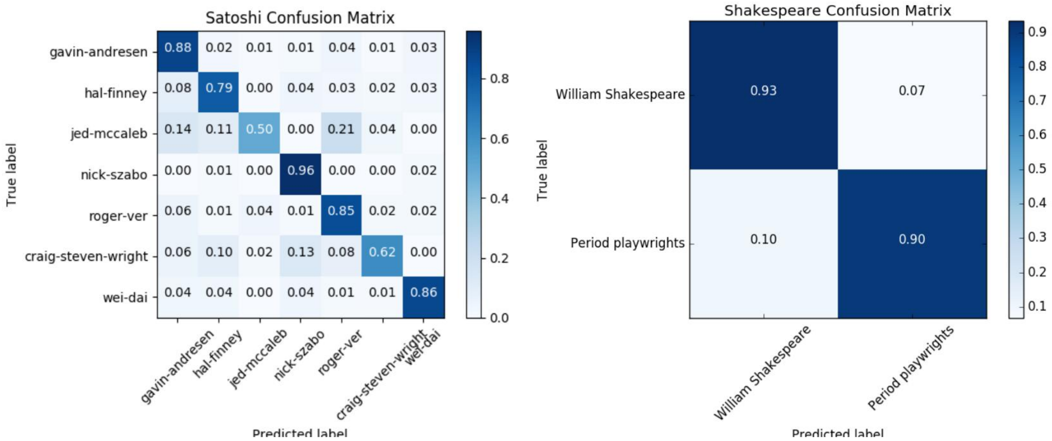


Figure 2: The normalized confusion matrices of our final models, as run on hold-out test data.

4.1 Shakespeare Results

Our Shakespeare model reported a test accuracy of 91.95%. From the confusion matrix, we can see that 7% of sequences from Shakespeare plays are misinterpreted as being from other authors, while 10% of sequences from other plays are misinterpreted as being by Shakespeare.

Finally, we tested our classifier on a number of works with disputed origins, as described in Section 3.1.1. We segmented each work into seventy-five word phrases, much like our training and validation data, then used the classifier to determine whether or not the phrase was likely to have been written by Shakespeare. We then used majority voting to classify the overall work. The results are shown in Figure 3 below. The overall results are vaguely promising. For many of the works that Shakespeare did not take part in writing (according to the modern academic consensus) – *Vortigern and Rowena*, *Thomas Lord Cromwell*, *The Birth of Merlin*, and *Pericles Act I* – more passages than not are identified by our model as belonging to some other author. Likewise, for *Pericles Act III* and *Pericles Act IV*, which Shakespeare is thought to have made contributions to, more passages are attributed to him than not. However, there is still some variability (e.g. *Pericles Act V* is also though to have been written by William Shakespeare, but is not indicated as such by the model) in the results, and even successful classifications are separated by a small number of examples from the mis-classifications. Thus our performance on the disputed works is lower than we had anticipated given our validation and test results on the original dataset. We discuss the implications of this further in Section 7.

[2] For backwards LSTMs, we perform the same process, since backward LSTMs also return sequences in reverse order.

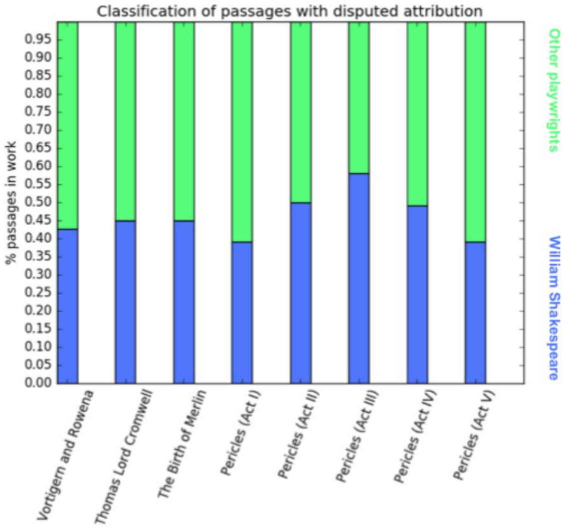


Figure 3: Results from running our LSTM model over period works with unknown or "disputed" authorship. For each work, the blue portion of the bar represents the proportion of 75-word phrases within the work that were deemed to match Shakespeare's style, and the green portion represents phrases resembling other contemporary playwrights.

4.2 Satoshi Results

On the Satoshi dataset, our model had a test accuracy of 85.94%. The confusion matrix is shown in Figure 2, which shows that we can distinguish between the posts of several people with greater than 80% accuracy. However, posts by Jed McCaleb and Craig Steven Wright are often confused for posts by others. This is likely due to the fact that they have the fewest training examples (266 and 466 respectively).

We then ran our classifier on the works of Satoshi Nakamoto, which gave us the chart shown in Figure 4. This suggests that Gavin Andresen's writing is similar to that of Satoshi Nakamoto. However, this is to be expected for two reasons - Gavin Andresen is the lead developer of Bitcoin, and he likely talks about the same topics that Satoshi talked about. Furthermore, Andresen also has the most amount of labeled examples in our dataset (3,474 examples), so it's likely that the classifications are biased towards him. Thus we cannot make any conclusions about the identity of Satoshi Nakamoto.

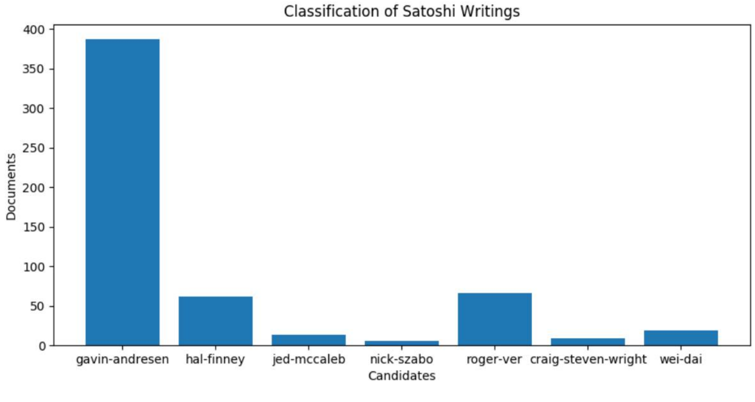
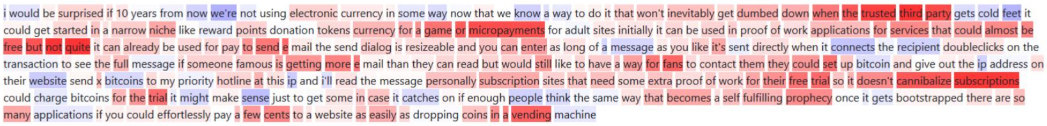


Figure 4: The final classification results on words by Satoshi Nakamoto.

We also generated activation maps for the hidden states of our LSTMs. One such activation map is show in Figure 5. Most of the activation maps are hard to interpret, but this one seems to be activated by software buzzwords, including “trusted third party,” “micropayments,” and “subscriptions.”



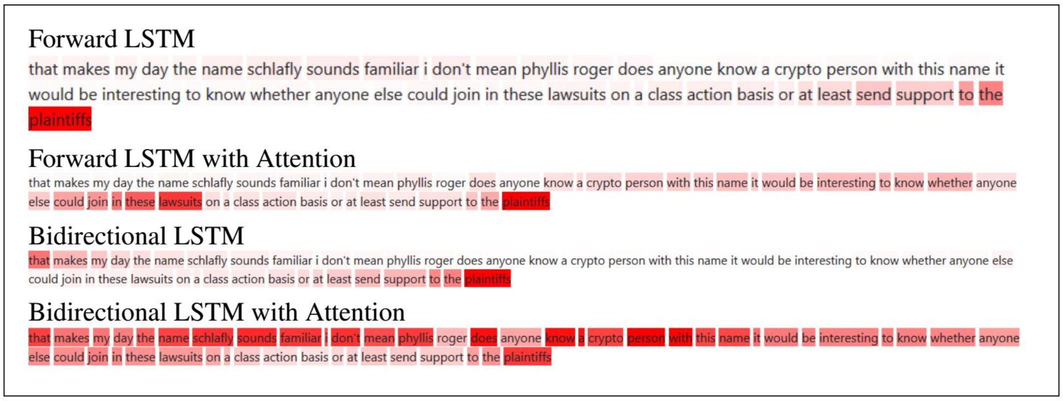
i would be surprised if 10 years from now we're not using electronic currency in some way now that we know a way to do it that won't inevitably get dumbed down when the trusted third party gets cold feet it
could get started in a narrow niche like reward points donation tokens currency for a game or micropayments for adult sites initially it can be used in proof of work applications for services that could almost be
free but not quite it can already be used for pay to send e mail the send dialog is resizeable and you can enter as long of a message as you like it's sent directly when it connects the recipient doubleclicks on the
transaction to see the full message if someone famous is getting more e mail than they can read but would still like to have a way for fans to contact them they could set up bitcoin and give out the ip address on
their website send x bitcoins to my priority hotline at this ip and i'll read the message personally subscription sites that need some extra proof of work for their free trial so it doesn't cannibalize subscriptions
could charge bitcoins for the trial it might make sense just to get some in case it catches on if enough people think the same way that becomes a self fulfilling prophecy once it gets bootstrapped there are so
many applications if you could effortlessly pay a few cents to a website as easily as dropping coins in a vending machine

Figure 5: Hidden state activation map for a sequence. Red corresponds with positive activation, while blue corresponds with negative activation. This hidden state seems to correspond with software buzzwords.

5 Experiments

5.1 BiLSTM and Attention

For our recurrent layer, we tried both forward LSTMs and BiLSTMs, as well as LSTMs with attention and without attention. To visualize the impact of these changes, we created saliency maps, which show the magnitude of the derivative of the predicted class score with respect to the input sequence. These saliency maps are shown in Figure 6.



Forward LSTM

that makes my day the name schlafly sounds familiar i don't mean phyllis roger does anyone know a crypto person with this name it
would be interesting to know whether anyone else could join in these lawsuits on a class action basis or at least send support to the
plaintiffs

Forward LSTM with Attention

that makes my day the name schlafly sounds familiar i don't mean phyllis roger does anyone know a crypto person with this name it would be interesting to know whether anyone
else could join in these lawsuits on a class action basis or at least send support to the plaintiffs

Bidirectional LSTM

that makes my day the name schlafly sounds familiar i don't mean phyllis roger does anyone know a crypto person with this name it would be interesting to know whether anyone else
could join in these lawsuits on a class action basis or at least send support to the plaintiffs

Bidirectional LSTM with Attention

that makes my day the name schlafly sounds familiar i don't mean phyllis roger does anyone know a crypto person with this name it would be interesting to know whether anyone
else could join in these lawsuits on a class action basis or at least send support to the plaintiffs

Figure 6: These saliency maps show the magnitude of the gradients of each word, with respect to the correct class score.

Due to the vanishing gradient and bottleneck problems, forward LSTMs are highly influenced by the last few words in a sentence. This is undesirable, since we effectively ignore most of the sentence. To fix this, we added simple dot-product attention, as described in section 3.3. This allows more words to have an influence on the final classification, however it is still biased towards words near the end of a sequence. Another method we tied was to run LSTMs bidirectionally. Without attention, BiLSTMs are influenced by words near the beginning and end of the sentence, but ignore words in the middle. However, when we combine BiLSTMs with attention, we see that words across the entire sequence can influence the class score, which is desirable.

As shown in Table 1, these changes also reflect an increase in the accuracy of the network as tested on the validation set. We can see that bidirectionality and attention both increase the accuracy. Using both together increased the accuracy for the Satoshi model, but not for the Shakespeare model.

5.2 LSTM vs. GRU

We also tested replacing the LSTM cell with a GRU cell, and found that the GRU was slightly better for both datasets. However, this difference is not enough to demonstrate that GRUs are a better choice for either dataset. The results are shown in Table 2.

Architecture	Satoshi Val Accuracy (%)	Shakespeare Val Accuracy (%)
Forward LSTM	81.13	86.52
Forward LSTM w/ Attention	85.13	92.76
BiLSTM	82.92	90.14
BiLSTM w/Attention	86.49	91.15

Table 1: Validation set accuracy for different configurations of LSTMs and Attention.

Cell Type	Satoshi Validation Accuracy (%)	Shakespeare Val Accuracy (%)
LSTM	86.49	91.15
GRU	87.06	91.35

Table 2: Validation set accuracy for different cell types. Cells are run bidirectionally with an attention layer.

5.3 Attention Mechanisms

Simple dot-product attention can be undesirable, as it forces the RNN hidden state to be both meaningful for classification and meaningful for attention, which may not be satisfiable at the same time. Thus, we decided to try multiplicative attention. However, as shown in Table 3, this did not produce an increase in accuracy for the Satoshi model, but it increased accuracy slightly for the Shakespeare model.

Attention Mechanism	Satoshi Validation Accuracy (%)	Shakespeare Val Accuracy (%)
Dot-Product	87.06	91.15
Multiplicative	85.99	92.76

Table 3: Validation set accuracy for different attention mechanisms.

5.4 Sequence Length vs. Accuracy

We wanted to know if the classifier was better at classifying longer sequences as compared to shorter sequences. We decided to test this by binning the test set by sequence length and checking the accuracy for each bin. This is only applicable to the Satoshi dataset, where documents are of differing lengths. By contrast, the Shakespeare dataset consists of a large sequence split into many sequences of uniform length. Results are shown in Figure 7. It turns out, the classifier is pretty good even for smaller sequences, but accuracy does improve slightly for longer sequences.

6 Future Work

A core difficulty of working with old plays from the late 16th and early 17th century is that they vary widely, which makes generalizing a model much more difficult. Further emphasis should be placed on data processing, especially with an eye towards developing strategies based on the type of written work itself. Specifically, we focused on the textual content of the written lines in isolation, but a lot of information remains in the structure of the play and the rhyme and verse.

For the Satoshi dataset, we could start by removing jargon and topic-specific words from the inputs. The idea here is to force the network to learn the way people say things, not the specific things that they say. Also specific to the Satoshi dataset, we could expand our model to include more information than just parsed words. Currently our model ignores punctuation and spacing. One way to fix this is to simply use a char-RNN. Another method is to parse punctuation and unusual spacing as tokens.

In terms of the neural architecture, one opportunity for future work is to try more attention mechanisms. We tried dot-product and multiplicative attention, but we could also try additive attention or

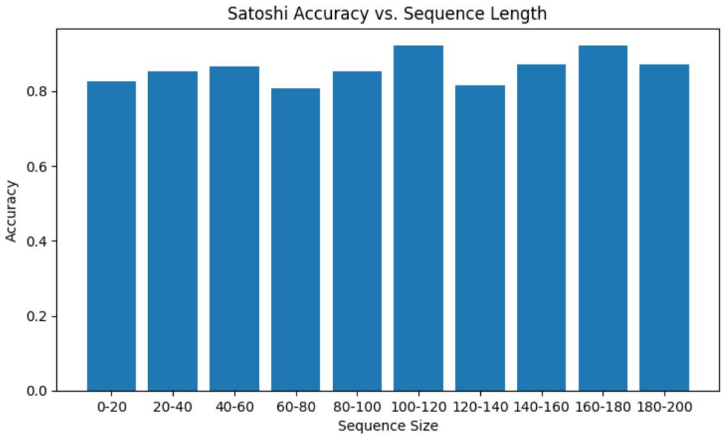


Figure 7: This chart shows how accuracy varies with sequence length for the Satoshi model.

self-attention. Another possibility is to experiment with stacking more LSTM/GRU layers on top of each other.

7 Conclusion

Although we found that LSTMs were effective at correctly re-labeling the examples in our hold-out dataset, we found that this did not translate to meaningful classifications of disputed works. The Satoshi model reports that Nakamoto's writing is similar to Gavin Andresen's, which makes sense but doesn't provide definitive identification. Furthermore, the Shakespeare model showed high accuracy in recovering labels, but produced mediocre results for disputed works. Our results cast doubt on the utility of previous works that use neural networks for authorship classification. Our accuracy is comparable to previous publications, but at the same time has not led to success in classifying disputed works. More work needs to be done to combat overfitting, as well as distribution mismatch between disputed and undisputed works.

8 References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] D. Bagnall. Author identification using multi-headed recurrent neural networks. *CoRR*, abs/1506.04891, 2015.

[3] S. Bearman. Bitcoin's creator may be worth $6 billion but people still don't know who it is. https://www.cnbc.com/2017/10/27/bitcoins-origin-story-remains-shrouded-in-mystery-heres-why-it-matters.html, 2017.

[4] F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[5] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[6] C. Qian, T. He, and R. Zhang. Deep learning based authorship identification. https://web.stanford.edu/class/cs224n/reports/2760185.pdf.

[7] F. J. Tweedie, S. Singh, and D. I. Holmes. Neural network applications in stylometry: The "federalist papers". *Computers and the Humanities*, 30(1):1–10, 1996.

[8] L. Yao and D. Liu. Wallace: Author detection via recurrent neural networks. https://cs224d.stanford.edu/reports/YaoLeon.pdf.

[9] L. Zhou and H. Wang. News authorship identification with deep learning. https://cs224d.stanford.edu/reports/ZhouWang.pdf.