Continuous *N*-gram Representations for Authorship Attribution

Yunita Sari, Andreas Vlachos and **Mark Stevenson**
Department of Computer Science, University of Sheffield, UK
{y.sari, a.vlachos, mark.stevenson}@sheffield.ac.uk

Abstract

This paper presents work on using continuous representations for authorship attribution. In contrast to previous work, which uses discrete feature representations, our model learns continuous representations for *n*-gram features via a neural network jointly with the classification layer. Experimental results demonstrate that the proposed model outperforms the state-of-the-art on two datasets, while producing comparable results on the remaining two.

1 Introduction

Authorship attribution is the task of identifying the author of a text. This field has attracted attention due to its relevance to a wide range of applications including forensic investigation (e.g. identifying the author of anonymous documents or phishing emails) (Chaski, 2005; Grant, 2007; Lambers and Veenman, 2009; Iqbal et al., 2010; Gollub et al., 2013) and plagiarism detection (Kimler, 2003; Gollub et al., 2013).

From a machine learning perspective, the task can be treated as a form of text classification. Let $D = d_1, d_2, ..., d_n$ be a set of documents and $A = a_1, a_2, ..., a_m$ a fixed set of candidate authors, the task of authorship attribution is to assign an author to each of the documents in D. The challenge in authorship attribution is that identifying the topic preference of each author is not sufficient; it is necessary to also capture their writing style (Stamatatos, 2013). This task is more difficult than determining the topic of a text, which is generally possible by identifying domain-indicative lexical items, since writing style cannot be fully captured by an author's choice of vocabulary.

Previous studies have found that word and character-level *n*-grams are the most effective features for identifying authors (Peng et al., 2003; Stamatatos, 2013; Schwartz et al., 2013). Word *n*-grams can represent local structure of texts and document topic (Coyotl-Morales et al., 2006; Wang and Manning, 2012). On the other hand, character *n*-grams have been shown to be effective for capturing stylistic and morphological information (Koppel et al., 2011; Sapkota et al., 2015).

However, previous work relied on discrete feature representations which suffer from data sparsity and do not consider the semantic relatedness between features. To address this problem we propose the use of continuous *n*-gram representations learned jointly with the classifier as a feed-forward neural network. Continuous *n*-grams representations combine the advantages of *n*-grams features and continuous representations. The proposed method outperforms the prior state-of-the-art approaches on two out of four datasets while producing comparable results for the remaining two.

2 Related Work

An extensive array of authorship attribution work has focused on utilizing content words and character *n*-grams. The topical preference of authors can be inferred by their choice of content words. For example, Seroussi et al. (2013) used the Author-Topic (AT) model (Rosen-Zvi et al., 2004) — an extension of Latent Dirichlet Allocation (Blei et al., 2003) — to obtain author representations. Experiments on several datasets yielded state-of-the-art performance.

Character *n*-grams have been widely used and have the advantage of being able to capture stylistic information. By using only the 2,500 most frequent 3-grams, Plakias and Stamatatos

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 267–273,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

(2008) successfully achieved 80.8% accuracy on the CCAT10 dataset, while Sapkota et al. (2015) reported slightly lower performance using only affix and punctuation 3-grams. Escalante et al. (2011) represent documents using a set of local histograms. This approach achieved an accuracy of 86.4%.

Beside being effective indicators of an author's writing style, both content words and character *n*-grams are also straightforward to extract from documents and are therefore widely used for author attribution. More complex features which require deeper textual analysis are also useful for the problem but have been used less frequently since the complexity of analysis required can hinder performance (Stamatatos, 2009). There have been several attempts to utilize semantic features for author attribution tasks, e.g. (McCarthy et al., 2006; Argamon et al., 2007; Brennan and Greenstadt, 2009; Bogdanova and Lazaridou, 2014). These approaches commonly use WordNet as a source of semantic information about words and phrases. For example, McCarthy et al. (2006) used WordNet to detect causal verbs while Brennan and Greenstadt (2009) used it to extract word synonyms. Our proposed model does not rely on any external linguistic resources, such as WordNet, making it more portable to new languages and domains.

3 Continuous *n*-grams Representations

This work focuses on learning continuous *n*-gram representations for authorship attribution tasks. Continuous representations have been shown to be helpful in a wide range of tasks in natural language processing (Bengio et al., 2003; Mikolov et al., 2013). Unlike the previous authorship attribution work which uses discrete representations, we represent each *n*-gram as a continuous vector and learn these representations in the context of the authorship attribution tasks being considered.

To learn the *n*-gram feature representations jointly with the classifier we adopt the shallow neural network architecture of fastText, which was recently proposed by Joulin et al. (2016). This model is similar to a standard linear classifier, but instead of representing a document with a discrete feature vector, the model represents it with a continuous vector obtained by averaging the continuous vectors for the features present. More formally, fastText predicts the probability distribution over the labels for a document as follows:

$$p(y|x) = softmax(BAx) \tag{1}$$

where x is the frequency vector of features for the document, the weight matrix A is a dictionary containing the embeddings learned for each feature, and B is a weight matrix that is learned to predict the label correctly using the learned representations (essentially averaged feature embeddings).

Since the documents in this model are represented as bags of discrete features, sequence information is lost. To recover some of this information we will consider feature *n*-grams, similar to the way convolutional neural network architectures incorporate word order (Kim, 2014) but with a simpler architecture.

The proposed model ignores long-range dependencies that could conceivably be captured using alternative architectures, such as recurrent neural networks (RNN) (Mikolov et al., 2010; Luong et al., 2013). However, topical and stylistic information is contained in shorter word and character sequences for which the shallow neural network architectures with *n*-gram feature representations are likely to be sufficient, while having the advantage of being much faster to run. This is particularly important for authorship attribution tasks which normally involves documents that are much longer than the single sentences which RNNs typically model.

4 Experiments

4.1 Datasets

We use four datasets in our experiments: Judgment, CCAT10, CCAT50 and IMDb62. These datasets have a different number of authors and document sizes, which allows us to perform experiments and test our approaches in different scenarios. All datasets were made available by the authors of their respective papers. Table 1 shows descriptive statistics for the datasets.

Judgment (Seroussi et al., 2011). The Judgment dataset was collected from judgment writing of three Australian High Court's judges (Dixon, McTiernan, and Rich) on various topics. In this dataset, the number of documents per author is not fixed; there are 902 docs from Dixon, 253 docs from McTiernan and 187 docs from Rich. Following Seroussi et al. (2013), we only use documents with undisputed authorship

	Judgment	CCAT10	CCAT50	IMDb62
# authors	3	10	50	62
# total documents	1,342	1,000	5,000	79,550
avg characters per document	11,957	3,089	3,058	1,401
avg words per document	2367	580	584	288

Table 1: Dataset statistics.

and run experiments with 10-fold cross-validation.

CCAT10 (Stamatatos, 2008). This dataset is a subset of Reuters Corpus Volume 1 (RCV1) (Rose et al., 2002) and consists of newswire stories by 10 authors labelled with the code CCAT (which indicates corporate/industrial news). The corpus was divided into 50 training and 50 test texts per author. In the experiments we follow prior work (Stamatatos, 2013) and measure accuracy using the train/test partition provided.

CCAT50. This corpus is a larger version of CCAT10. In total there are 5,000 documents from 50 authors. Same as CCAT10, for each of the author there are 50 training and 50 test documents.

IMDb62 (Seroussi et al., 2010). The IMDb62 dataset consists of 62,000 movie reviews and 17,550 message board posts from 62 prolific users of the Internet Movie database (IMDb, `www.imdb.com`). Following Seroussi et al. (2013), 10-fold cross-validation was used.

4.2 Model Variations

We perform experiments with three variations of our approach:

- **Continuous word *n*-grams**. In this model we use word uni-grams and bi-grams. We set the 700 most common words as the vocabulary.

- **Continuous character *n*-grams**. Following previous work (Sanderson and Guenter, 2006), we use up to four-grams, as it is found to be the best *n* value for short English text. We follow Zhang et al. (2015) by setting the vocabulary to 70 most common characters including letters, digits, and some punctuation marks.

- **Continuous word and character *n*-grams**. This model combines word and character *n*-grams features.

4.3 Hyperparameters Tuning and Training Details

For all datasets, early stopping was used on the development sets and models trained with the Adam update rule (Kingma and Ba, 2015). Since none of the datasets have a standard development set, we randomly selected 10% of the training data for this purpose. Both word and character embeddings were initialized using Glorot uniform initialization (Glorot and Bengio, 2010). Keras's (Chollet, 2015) implementation of fastText was used for the experiments. The softmax function was used in the output layer without the *hashing trick*, which was sufficient for our experiments given the relatively small sized datasets. Code to reproduce the experiments is available from `https://github.com/yunitata/continuous-n-gram-AA`.

For the Judgment, CCAT10 and CCAT50 datasets an embedding layer with embedding size of 100, dropout rate of 0.75, learning rate of 0.001 and mini-batch size of 5 were used. The model was trained for 150 epochs. The values for the dropout rate and mini-batch size were chosen using a grid search on the CCAT10 devset. Other hyperparameters values (i.e. learning rate and embedding size) are fixed. For IMDb62, we used the same dropout rate. In order to speed up the training process on this dataset, the learning rate, embedding size, mini-batch size and number of epochs were set to 0.01, 50, 32 and 20 respectively.

5 Results and Discussion

Table 2 presents the comparison of the proposed approaches against the previous state-of-the-art methods on the four authorship attribution datasets considered. Overall, our results show the effectiveness of continuous *n*-grams representations

Model	Judgment	CCAT10	CCAT50	IMDb62	Average
SVM with affix+punctuation 3-grams (Sapkota et al., 2015)	-	78.80	69.30	-	-
SVM with 2,500 most frequent 3-grams (Plakias and Stamatatos, 2008)	-	80.80	-	-	-
STM-Asymmetric cross (Plakias and Stamatatos, 2008)	-	78.00	-	-	-
SVM with bag of local histogram (Escalante et al., 2011)	-	**86.40**	-	-	-
Token SVM (Seroussi et al., 2013)	91.15	-	-	92.52	-
Authorship attribution with topic models (Seroussi et al., 2013)	**93.64**	-	-	91.79	-
Continuous *n*-gram words (1,2)	90.31	77.80	70.16	87.87	81.54
Continuous *n*-gram char (2,3,4)	91.29	74.80	**72.60**	**94.80**	83.37
Continuous *n*-gram words (1,2) and char (2,3,4)	91.51	77.20	72.04	94.28	83.51

Table 2: Comparison against previous results.

which outperform the previous best results on the CCAT50 and IMDb62 datasets. In the Judgment dataset, our models obtain comparable results with the previous best. However as can be seen in the table, the accuracy on CCAT10 is substantially worse than the one reported by Escalante et al. (2011)'s result. Our attempt to reproduce their result failed by obtaining only 77% in the accuracy. Another attempt by Potthast et al. (2016) reported slightly worse accuracy of 75.4%.

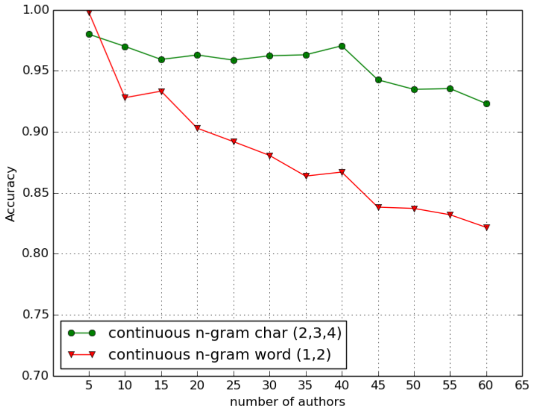

Figure 1: Accuracy on IMDb62 data subset with varying number of authors

5.1 Word vs Character

Table 2 demonstrates that the character models are superior to the word models. In particular, we found that models which employ character level *n*-grams appear to be more suitable for datasets with a large number of authors, i.e. CCAT50 and IMDb62. To explore this further, we ran an additional experiment varying the number of authors on a subset of IMDb62. For each of the authors we use 200 documents, with 10% of the data used as the development set and another 10% as the test set. Figure 1 shows a steep decrease in the accuracy of word models when the number of authors increases. The drop in accuracy of the character *n*-gram model is less pronounced.

Character models also achieve a slightly better result on the Judgment dataset which consists of only three authors. This can be explained by the fact that the documents in this corpus are significantly longer (almost ten and four times longer than those in IMDb62 and CCAT50 respectively (see Table 1). The large numbers of word *n*-grams make it more difficult to learn good parameters for them. Combining word and character *n*-grams only produced a very small improvement on this dataset.

5.2 Domain Influence

The majority previous work on authorship attribution has concluded that content words are more effective for datasets where the authors can be discriminated by the document topic (Peng et al., 2004; Luyckx, 2010). Seroussi et al. (2013) show that the Judgment and IMDb62 datasets fall into this category and approaches based on topic models achieve high accuracy (more than 90%). However, our results demonstrate stylistic information from continuous character *n*-grams can outperform word-based approaches on both datasets. In addition, this results also support the superiority of character *n*-grams that have been reported in the previous work (Peng et al., 2003; Stamatatos, 2013; Schwartz et al., 2013).

5.3 Feature Contributions

An ablation study was performed to further explore the influence of different types of features by removing a single class of *n*-grams. For this experiment the character model was used on the two CCAT datasets. Three feature types are defined including:

1. **Punctuation *N*-gram:** A character *n*-gram which contains punctuations. There are 34 punctuation symbols in total.

2. **Space *N*-gram:** A character *n*-gram that contains at least one whitespace character.

3. **Digit *N*-gram:** A character *n*-gram that contains at least one digit.

In addition, we also assess the influence of the length of the character *n*-grams. Results are presented in the Table 3.

	CCAT10	CCAT50
all features (char model)	74.80	72.60
(–) punctuation *n*-grams	73.80	68.80
(–) space *n*-grams	71.80	70.20
(–) digit *n*-grams	75.60	71.28
(–) bi-grams	76.20	72.08
(–) tri-grams	74.80	71.84
(–) four-grams	74.40	71.16

Table 3: Results of feature ablation experiment.

Table 3 demonstrates that removing punctuation and space *n*-grams leads to performance drops on both of the datasets. On the other hand, leaving out digit *n*-grams and bi-grams improves accuracy on the CCAT10 dataset. Other *n*-gram types do not seem to affect the results much.

6 Conclusion

This paper proposed continuous *n*-gram representations for authorship attribution tasks. Using four authorship attribution datasets, we showed that this model is effective for identifying writing style of the authors. Our experimental results provide evidence that continuous representations are suitable for a stylistic (as opposed to topical) text classification task such as authorship attribution.

Acknowledgments

We thank the anonymous reviewers for their comments. The first author would like to acknowledge Indonesia Endowment Fund for Education (LPDP) for support in the form of a doctoral studentship.

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, April.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, March.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Dasha Bogdanova and Angeliki Lazaridou. 2014. Cross-language authorship attribution. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 2015–2020, Reykjavik, Iceland.

Michael Robert Brennan and Rachel Greenstadt. 2009. Practical attacks against authorship recognition techniques. In *Proceedings of the 21st Conference on Innovative Applications of Artificial Intelligence, IAAI 2009*, Pasadena, California, USA. AAAI.

Carole E. Chaski. 2005. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence*, 4:1–13.

François Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

Rosa María Coyotl-Morales, Luis Villaseñor Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. Authorship attribution using word sequences. In *Proceedings of the 11th Iberoamerican Conference on Progress in Pattern Recognition, Image Analysis and Applications, CIARP 2006*, pages 844–853, Berlin, Heidelberg. Springer-Verlag.

Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL HLT 2011*, pages 288–298, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics, AISTATS 2010*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. Society for Artificial Intelligence and Statistics.

Tim Gollub, Martin Potthast, Anna Beyer, Matthias Busse, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Recent trends in digital text forensics and its evaluation plagiarism detection, author identification and author profiling. In *Proceedings of Conference and Labs of the Evaluation Forum, CLEF 2013*, pages 282–302, Valencia, Spain.

Tim D. Grant. 2007. Quantifying Evidence for Forensic Authorship Analysis. *International Journal of Speech, Language and Law*, 14(1):1–25.

Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi. 2010. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1-2):56–64.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Marco Kimler. 2003. Using style markers for detecting plagiarism in natural language documents. Master's thesis, Department of Computer Science, University of Skövde, Sweden, August.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceeding of the 3rd International Conference for Learning Representations, ICLR 2015*, San Diego, CA, May.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, March.

Maarten Lambers and Cor J. Veenman. 2009. Forensic authorship attribution using compression distances to prototypes. In *Proceeding of the 3rd International Workshop on Computational Forensics, IWCF 2009*, pages 13–24, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceeding of the Conference on Computational Natural Language Learning, CoNLL 2013*, pages 104–113, Sofia, Bulgaria.

Kim Luyckx. 2010. *Scalability Issues in Authorship Attribution*. Ph.D. thesis, CLiPS Computational Linguistics Group, University of Antwerp, Belgium, December.

Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, and Danielle S. McNamara. 2006. Analyzing writing styles with coh-metrix. In *Proceedings of the 19th Annual Florida Artificial Intelligence Research Society International Conference, FLAIRS 2006*, pages 764–770, Melbourne Beach, FL. AAAI Press.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. *Interspeech*, 2:3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS 2013*, pages 3111–3119, USA. Curran Associates Inc.

Fuchun Peng, Dale Schuurmanst, Vlado Kesel, and Shaojun Wan. 2003. Language Independent Authorship Attribution using Character Level Language Models. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics, EACL 2003*, Budapest, Hungary.

Fuchun Peng, Dale Schuurmans, and Shaojun Wang. 2004. Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345, September.

Spyridon Plakias and Efstathios Stamatatos. 2008. Tensor space models for authorship identification. In *Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications, SETN 2008*, pages 239–249, Berlin, Heidelberg. Springer-Verlag.

Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, et al. 2016. Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In *Proceedings of the European Conference on Information Retrieval, ECIR 2016*, volume 9626, pages 393–407, March.

Tony Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus - from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation, LREC 2002*, pages 827–832, Las Palmas, Canary Islands.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI 2004*, pages 487–494, Arlington, Virginia, United States. AUAI Press.

Conrad Sanderson and Simon Guenter. 2006. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006*, pages 482–491, Sydney, Australia, July. Association for Computational Linguistics.

Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. 2015. Not all character n-grams are created equal: A study in authorship attribution. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL HLT 2015*, pages 93–102, Denver, Colorado, May–June. Association for Computational Linguistics.

Roy Schwartz, Oren Tsur, Ari Rappoport, and Moshe Koppel. 2013. Authorship attribution of micro-messages. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1880–1891, Seattle, Washington, USA, October. Association for Computational Linguistics.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2010. Collaborative inference of sentiments from texts. In *Proceedings of 18th International Conference on User Modeling, Adaptation, and Personalization, UMAP 2010*, pages 195–206, Big Island, HI, USA, June. Springer.

Yanir Seroussi, Russell Smyth, and Ingrid Zukerman. 2011. Ghosts from the high court's past: Evidence from computational linguistics for dixon ghosting for mctiernan and rich. *University of New South Wales Law Journal*, 34(3):984–1005.

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. 2013. Authorship Attribution with Topic Models. *Computational Linguistics*, 40(2):269–310.

Efstathios Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing and Management*, 44(2):790 – 799.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, March.

Efstathios Stamatatos. 2013. On the Robustness of Authorship Attribution Based on Character n-gram Features. *Journal of Law and Policy*, 21(2):421–439.

Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, ACL 2012*, pages 90–94, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS 2015*, pages 649–657, Cambridge, MA, USA. MIT Press.