# Using Deep Networks and Transfer Learning to Address Disinformation

Numa Dhamani [1] Paul Azunre [2] Jeffrey L. Gleason [1] Craig Corcoran [1] Garrett Honke [3] Steve Kramer [1] Jonathon Morgan [1]

## Abstract

We apply an ensemble pipeline composed of a character-level convolutional neural network (CNN) and a long short-term memory (LSTM) as a general tool for addressing a range of disinformation problems. We also demonstrate the ability to use this architecture to transfer knowledge from labeled data in one domain to related (supervised and unsupervised) tasks. Character-level neural networks and transfer learning are particularly valuable tools in the disinformation space because of the messy nature of social media, lack of labeled data, and the multi-channel tactics of influence campaigns. We demonstrate their effectiveness in several tasks relevant for detecting disinformation: spam emails, review bombing, political sentiment, and conversation clustering.

## 1. Introduction

Electronic communication is more embedded and essential to human life than ever before. This communication increasingly relies on the distributed, self-publishing model of social media platforms. The increasing ease of distributed communication does not come without drawbacks: disinformation—deceptive information spread deliberately to change behavior, influence public opinion, or obscure the truth—has infiltrated the online information ecosystem, with damaging consequences (Carvalho et al., 2011; Mocanu et al., 2015).

Malicious electronic communication takes many forms. It spans from low-level social engineering attacks (i.e., phishing) to more sophisticated, distributed efforts to disseminate state propaganda (Inkster, 2016; Okoro & Nwafor, 2013; Woolley, 2016). We propose that the language characteristics of known and identified sources of malicious electronic communication can be used as a signal for the detection and mitigation of these efforts across the diverse (and fractured) electronic communication ecosystem.

The motivation of this work is to demonstrate how semantic classification of natural language can be used as a tool for the detection of inflammatory, inauthentic, or otherwise nefarious communication. Character-level convolutional neural networks (CNNs) are particularly well-suited for this task—as opposed to a word-level model—because they allow for non-vernacular discourse, misspelling, and other social media features (e.g., emoticons) to be learned without the constraint of fixed vocabularies (Zhang et al., 2015). We implement an adaptation of a neural network architecture recently demonstrated to be effective for text classification (Zhang et al., 2015; Józefowicz et al., 2016). The method is purely content-based and does not require any additional metadata beyond the text. To show the effectiveness of this method in relation to malicious communication and disinformation, we present a series of experimental results on semantic classification for spam emails, review bombing, political sentiment, and conversation clustering.

## 2. Related Work

The way people consume and produce information online looks radically different today than it did in the recent past (Del Vicario et al., 2016). This change has introduced a societal-level vulnerability—where foreign entities have been accused of interfering in the operations of sovereign democracies (Inkster, 2016; Okoro & Tsgyu, 2017; Woolley, 2016; Peruzzi et al., 2018). Even internally in various states, political differences have encouraged inauthentic, organized disinformation campaigns to attack brands (Visentin et al., 2019; Berthon et al., 2018). This situation has led to an increased interest in studying the dissemination of inauthentic and/or false information to find solutions that protect the integrity of online discourse and the democratic institutions that depend on it. We believe that a computational technique that can use the properties of electronic discourse for the purpose of identifying key characteristics (e.g., source, intent, and effect) can be effective for addressing this complex problem.

[1] New Knowledge, Austin, Texas, USA [2] Algorine, Inc., Austin, Texas, USA [3] Watson School of Engineering and Applied Science and the Department of Psychology: Cognitive and Brain Sciences, Binghamton University (SUNY), Binghamton, New York, USA. Correspondence to: Numa Dhamani <numa@newknowledge.io>.

*Appearing at the International Conference on Machine Learning AI for Social Good Workshop*, Long Beach, United States, 2019.

Few frameworks have been proposed to detect and/or monitor malicious communication. One existing approach is HOAXY, a platform designed to collect, detect, and visualize the online spread of misinformation and fact-checking (Shao et al., 2016). TRUTHY (another approach) uses network analysis techniques to track political memes for the purpose of detecting misinformation in U.S. political elections (Ratkiewicz et al., 2011). It includes an automated system that assesses the credibility of posts on social media using content-based features and user metadata (Castillo et al., 2011).

Motivated by computational journalism, FACTWATCHER helps journalists identify newsworthy facts supported by data that can serve as leading news stories (Hassan et al., 2014). Another attempt to fact-check simple statements employs the use of a public knowledge graph extracted from Wikipedia, framing the problem in terms of network analysis (Ciampaglia et al., 2015). There have also been efforts to create automated fact-checking systems for multimedia, where the goal is to distinguish between fake and real images using classification models based on user features and post content (Boididou et al., 2014; Gupta et al., 2013). Platforms that focus on the veracity of claims more broadly—e.g., POLITIFACT.COM and FACTCHECK.ORG—rate the accuracy of claims made in political debates, speeches, and ads. More recently, computational resources and models to detect fake news have been developed which rely on linguistic features (Pérez-Rosas et al., 2017).

Current research in this domain includes tracking and analyzing the diffusion of rumors. RUMORLENS is a suite of interactive tools to help identify rumors on social media and analyze their impact (Resnick et al., 2014). Similarly, TWITTERTRAILS allows users to automatically answer questions about rumors on Twitter (e.g., origin, propagators, main actors, etc.) and compute level of visibility (Metaxas et al., 2015). Another rumor-checking website is SNOPES.COM which focuses on urban legends and hoaxes.

# 3. Current Approach

The work surveyed above is notable for its goal of addressing a difficult problem with the use of all possible features. In the same sense, however, a limitation of this work is its reliance on complete and detailed information from the platform (i.e., most often the TWITTER platform, which makes available metadata and network graph characteristics). There is a current need, therefore, for an approach that can successfully classify content of interest across platforms, with only the text content as input and in the absence of any exogenous features—features that cannot be relied on consistently in cross-platform analysis. Our proposed method is a multi-label, multi-class semantic classifier able to successfully handle natural language across platforms. The capability to handle naturalistic social language—e.g., emoticons, slang, misspellings—at the character-level gives this framework potential to be a powerful tool in detecting malicious communication from peer-to-peer social engineering to distributed disinformation across the electronic communication ecosystem.

## 3.1. Neural Network Architecture Overview

Depictions of the architectural inspiration for the current approach[1] are presented in Figure 2 (see Zhang et al., 2015; Józefowicz et al., 2016 for in-depth architecture information). It consists of two coupled deep neural networks—a network that encodes each individual sentence and a network that classifies the entire document.

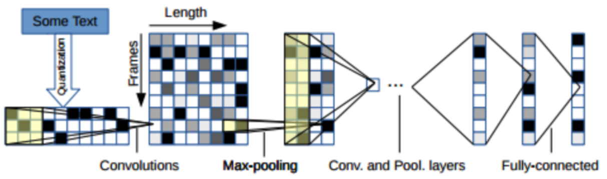

*Figure 1.* Diagram of character-level CNN model architecture. Figure reproduced from Zhang et al., 2015.

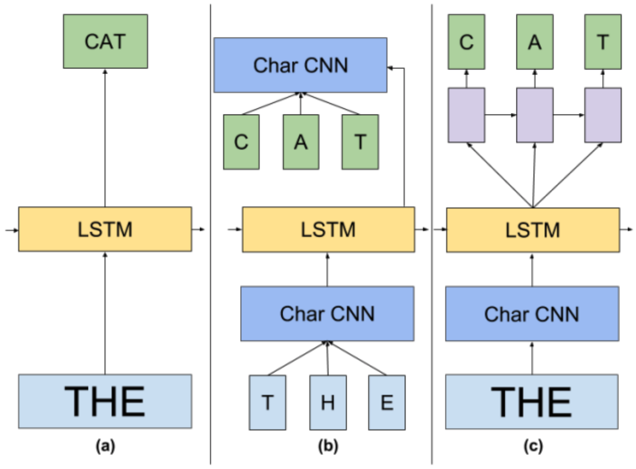

*Figure 2.* Word- and character-level models for text generation. This work uses the encoder from (b). Figure reproduced from Józefowicz et al., 2016.

The first network consists of 13 convolutional, max-pooling, dropout and bidirectional LSTM layers. Characters of an input sentence are one-hot encoded from a dictionary of 71 characters and bounded by a maximum length (tunable, depending on the application). It produces the final sentence encoding (a 512-dimensional feature vector).

[1] Source code: https://github.com/NewKnowledge/simon

The second network, which classifies the document as a whole, consists of 7 convolutional, max-pooling, dropout, and bidirectional LSTM layers. The input is again bounded by a maximum number of sentences (a tunable parameter). The final layer outputs the probabilities of classifying the document as each different type of class.

## 4. Empirical Assessment

Three examples are provided below to showcase how the CNN–LSTM classification approach can be applied to domains of interest. The final section describes how the process can be generalized to conduct unsupervised conversation clustering. The procedure is similar across applications: (1) initial weights are learned with training on the language(s) of the target domain and—if necessary—(2) transfer learning with labeled data is performed to learn the specific classes of interest (e.g., spam vs. ham, on-topic vs. off-topic review content, etc.).

### 4.1. Spam Email

Peer-to-peer social engineering attacks constitute a major threat vector at individual, business, and geopolitical levels. Perhaps the most benign example of this threat is commercial spam. Thus, a good starting point for assessment is testing the applicability of the CNN–LSTM architecture to this problem. In addition to being able to classify an email as spam or authentic, this use case can be extended to build a multi-label, multi-class classifier that differentiates between a broader set of attacks, e.g., propaganda and social engineering emails vs. phishing. We employ the transfer-learning paradigm to adapt the initial classes to a more sophisticated problem—first learning to determine if the originator of a piece of electronic communication is a friend or foe, and then learning to identify the specific class of adversarial communication best attributed to the document (see Figure 4 in the Supplementary Material).

The classifier is initially trained on the popular Enron email dataset [2], the 419 spam fraud corpus [3], and an email abuse dataset acquired from NASA Jet Propulsion Laboratory (JPL). A dataset balanced between two classes (*friend* and *foe*) was generated with 7000 samples of each class. The model converged with a test binary accuracy of 96.14%.

Then transfer learning is applied to the original binary classifier to complete the training of a full multi-class classifier, augmenting the number of handled classes from 2 to 8 with reduced labeled data and computing power requirements. The 8 classes are: friend, 419 scam, malware, credential phishing, phishing training, propaganda, social engineering, and spam. As shown in Figure 4 in the Supplementary Material, the accuracy convergences quickly to a test accuracy of 93.25%. Corresponding precision was calculated as 0.96, recall as 0.49, and F1 score as 0.65. This performance, using email text **only**, provides further evidence of the effectiveness of transfer learning in natural language processing (NLP), which still remains relatively under-explored in this space (Rodriguez et al., 2018).

### 4.2. Review Bombing

An important use-case for combating disinformation is the ability to detect inauthentic behaviors and campaigns used to threaten the integrity and good reputation of a product, company, or brand. For example, the perceived quality of products on marketplace-based websites can be manipulated through fake reviews (Goswami et al., 2017). Likewise, movie reviews can be manipulated through coordinated campaigns that have the goal to negatively impact box office numbers and revenue—so-called *review bombing*. One approach to combating this manipulation is to identify off-topic reviews, reviews that are not relevant to the product but take on peripheral issues (e.g., political stances of organizations or athletic endorsers).

To test the usefulness of CNN–LSTM based semantic classification in this domain, our method was trained on *on-topic* vs. *off-topic* movie reviews collected from a popular, crowd-sourced review website. The initial convergence results achieve a binary test accuracy of 99.5%. These results suggest that this technique is a promising approach for detecting inauthentic behaviors for manipulation of products and brands.

### 4.3. Political Sentiment

Due to the central role social media platforms fill as critical communication infrastructure, malicious actors now have the ability to construct and disseminate false narratives with broad implications for society—often by piggybacking on polarizing content or current events. This behavior has significant and harmful effects (Nied et al., 2017), muddying the ability to use social media as a means to understand broad trends in preferences and behavior. To be able to combat fraudulent or misleading narratives, it is critical to be able to characterize the content as intentionally harmful (negative/anti), intentionally favorable (positive/pro), or neutral.

Semantic classification was applied to posts discussing Howard Schultz' potential presidential election campaign using the text of the post **only**. A three-class classifier—*Pro*, *Anti*, or *Neutral*—was used. From a corpus containing approximately 6,000 supportive posts, 20,000 negative or harmful posts, and 40,000 neutral posts, we sampled 4,000

[2] www.kaggle.com/wcukierski/enron-email-dataset

[3] www.kaggle.com/rtatman/fraudulent-email-corpus

posts from each category.

Figure 5 in the Supplementary Material presents the training results on the political sentiment task. The model attained a binary test accuracy of 88.10%. Notably, the accuracy of the classification of negative and/or harmful posts is perfectly (100%) accurate. Corresponding precision, at a threshold probability of 0.5, was calculated as 0.68, recall as 0.98, and F1 score as 0.81. The ROC curve is also presented in Figure 5. We note that no hyper-parameter tuning was performed, which suggests that these metrics are amenable to further improvement.

### 4.4. Conversation Clustering

The current approach can also learn features that are useful for exploratory data analysis. The ability to examine the structure and trends of online discourse provides critical context for detecting and understanding the large-scale disinformation campaigns. We use the latent space generated by a pre-trained model to cluster conversations and visualize the results.

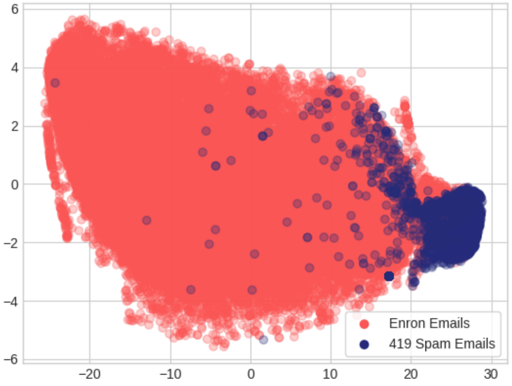

*Figure 3.* t-SNE embedding on Enron and 419 spam email datasets.

An example is shown in Figure 3, which depicts a *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) clustering of the 128-dimensional features that the model learned from training on the Enron email dataset and the 419 spam fraud corpus. The features effectively separate the two classes and illustrates the increased variability in the Enron class as opposed to the spam class.

## 5. General Discussion

### 5.1. Limitations and Alternatives

One criticism of this work is that a simpler approach to text classification might be capable of achieving similar levels of performance. Indeed, methods such as bag-of-words and bag-of-*n*grams models have comparable performance (Zhang et al., 2015) on many supervised tasks. However, we assert that the latent representation learned by the proposed approach better captures nuance in discourse than the bag-of-*n*grams representation and, consequently, is better suited for transfer learning and exploratory tasks. The choice of a character-level CNN over the common approach of using word embedding spaces allows for misspelling, slang, unconventional characters (e.g., emoticons), and anomalous or unique vocabulary usage. These advantages are uniquely suited for the purpose of analyzing and tracking discourse patterns that are not represented in mainstream discourse.

### 5.2. Ongoing Work

The applications above begin to address small parts of the disinformation problem. We think of these as pieces of a broader approach necessary to effectively combat disinformation. Towards that goal, we have an ongoing effort that examines the dynamics of language usage across communities and social platforms. The kind of document embedding used above for transfer learning and clustering can also be used to measure the similarity of language use across communities, track changes in language use over time, and identify potential sources of those changes (e.g. adopting slang present in another platform or community, indicating a diffusion of ideas between the two groups).

### 5.3. Conclusion

We present a character-level CNN–LSTM model as an effective general-purpose tool for exploration and inference tasks relevant to disinformation on social media. As demonstrated in Section 4, the model is able to learn a generic vector-space representation from a rich labeled dataset, then be utilized in label-scarce settings (common in the disinformation space) via transfer learning. The representation is also useful in an exploratory setting, which is important for understanding and adapting to the rapidly changing social media ecosystem. The flexibility of this framework to create a compact representation of natural language across a variety of communication channels by analyzing character-to-character contingencies makes it a powerful tool with broad potential to affect the quality and integrity of online discourse.

### Acknowledgments

Work was supported by the Defense Advanced Research Projects Agency (DARPA) under D3M (FA8750-17-C-0094) and ASED (FA8650-18-C-7889). Views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

## References

Berthon, P., Treen, E., and Pitt, L. How truthiness, fake news and post-fact endanger brands and what to do about it. *GfK Marketing Intelligence Review*, 2018.

Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., and Newman, N. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 743–748. ACM, 2014.

Carvalho, C., Klagge, N., and Moench, E. The persistent effects of a false news shock. *Journal of Empirical Finance*, 18(4):597–615, 2011.

Castillo, C., Mendoza, M., and Poblete, B. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pp. 675–684. ACM, 2011.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., and Flammini, A. Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):1–13, 06 2015. doi: 10.1371/journal.pone.0128193. URL https://doi.org/10.1371/journal.pone.0128193.

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. The spreading of misinformation online. *Proc. Natl. Acad. Sci. U. S. A.*, 113(3):554–559, January 2016.

Goswami, K., Park, Y., and Song, C. Impact of reviewer social interaction on online consumer review fraud detection. *Journal of Big Data*, 4(1):15, May 2017. ISSN 2196-1115. doi: 10.1186/s40537-017-0075-6. URL https://doi.org/10.1186/s40537-017-0075-6.

Gupta, A., Lamba, H., Kumaraguru, P., and Joshi, A. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pp. 729–736, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2038-2. doi: 10.1145/2487788.2488033. URL http://doi.acm.org/10.1145/2487788.2488033.

Hassan, N., Sultana, A., Wu, Y., Zhang, G., Li, C., Yang, J., and Yu, C. Data in, fact out: Automated monitoring of facts by factwatcher. *Proc. VLDB Endow.*, 7(13):1557–1560, August 2014. ISSN 2150-8097. doi: 10.14778/2733004.2733029. URL http://dx.doi.org/10.14778/2733004.2733029.

Inkster, N. Information warfare and the us presidential election. *Survival*, 58(5):23–32, 2016. doi: 10.1080/00396338.2016.1231527. URL https://doi.org/10.1080/00396338.2016.1231527.

Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. Exploring the limits of language modeling. *CoRR*, abs/1602.02410, 2016. URL http://arxiv.org/abs/1602.02410.

Metaxas, P. T., Finn, S., and Mustafaraj, E. Using twittertrails.com to investigate rumor propagation. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work &#38; Social Computing*, CSCW'15 Companion, pp. 69–72, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2946-0. doi: 10.1145/2685553.2702691. URL http://doi.acm.org/10.1145/2685553.2702691.

Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., and Quattrociocchi, W. Collective attention in the age of (mis) information. *Computers in Human Behavior*, 51:1198–1204, 2015.

Nied, A. C., Stewart, L., Spiro, E., and Starbird, K. Alternative narratives of crisis events: Communities and social botnets engaged on social media. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, pp. 263–266, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4688-7. doi: 10.1145/3022198.3026307. URL http://doi.acm.org/10.1145/3022198.3026307.

Okoro, N. and Nwafor, K. A. Social media and political participation in nigeria during the 2011 general elections: The lapses and the lessons. *Global Journal of Arts Humanities and Social Sciences*, 2013.

Okoro, N. and Tsgyu, S. An appraisal of the utilisation of social media for political communication in the 2011 nigerian presidential election. *African Research Review*, 11:115, 02 2017. doi: 10.4314/afrrev.v11i1.9.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. Automatic Detection of Fake News. *arXiv e-prints*, art. arXiv:1708.07104, Aug 2017.

Peruzzi, A., Zollo, F., Quattrociocchi, W., and Scala, A. How news may affect markets complex structure: The case of cambridge analytica. *Entropy*, 20(10), 2018. ISSN 1099-4300. doi: 10.3390/e20100765. URL http://www.mdpi.com/1099-4300/20/10/765.

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pp. 249–252. ACM, 2011.

Resnick, P., Carton, S., Park, S., Shen, Y., and others. Rumorlens: A system for analyzing the impact of rumors and corrections in social media. *Proc. Computational*, 2014.

Rodriguez, J. D., Caldwell, A., and Liu, A. Transfer learning for entity recognition of novel classes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1974–1985, 2018.

Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pp. 745–750. International World Wide Web Conferences Steering Committee, 2016.

Visentin, M., Pizzi, G., and Pichierri, M. Fake news, real problems for brands: The impact of content truthfulness and source credibility on consumers' behavioral intentions toward the advertised brands. *Journal of Interactive Marketing*, 45:99 – 112, 2019. ISSN 1094-9968. doi: https://doi.org/10.1016/j.intmar.2018.09.001. URL http://www.sciencedirect.com/science/article/pii/S1094996818300525.

Woolley, S. C. Automating power: Social bot interference in global politics. *First Monday*, 21(4), March 2016.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pp. 649–657, 2015.

## Supplementary Material

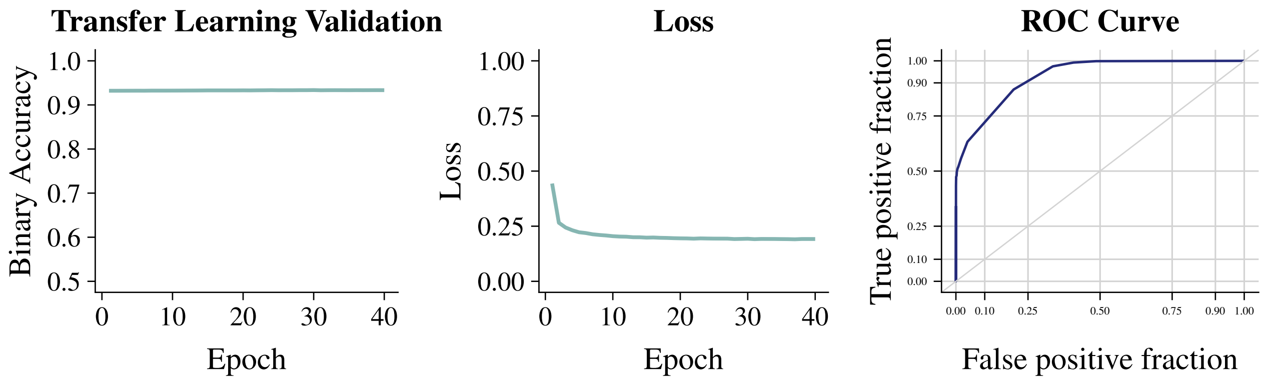

*Figure 4.* Convergence and performance results for the email spam classification task.

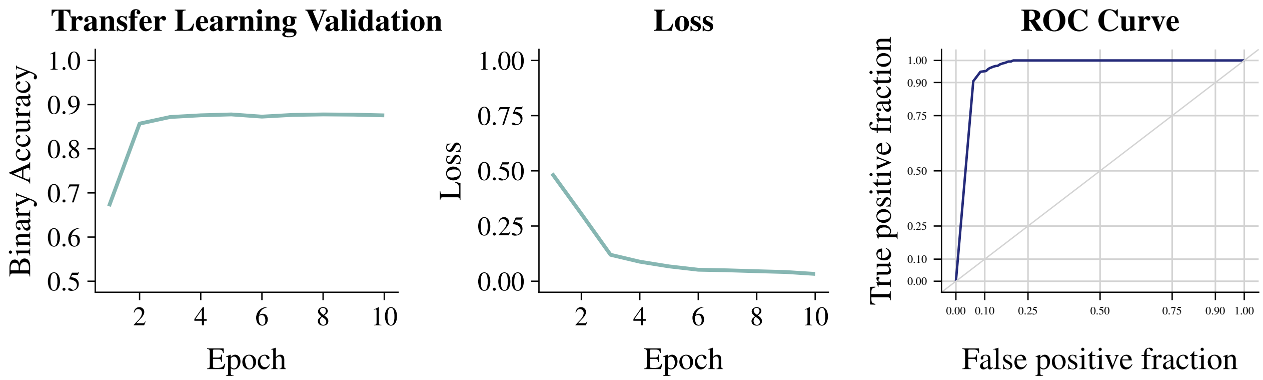

*Figure 5.* Convergence and performance results for the narrative categorization task.