Towards Deep Learning Interpretability: A Topic Modeling Approach

Yidong Chai
Tsinghua University, yidongchai@gmail.com

Weifeng Li
University of Georgia, weifeng.li@uga.edu

Towards Deep Learning Interpretability: A Topic Modeling Approach

Short Paper

Yidong Chai
Tsinghua University
Beijing, China
chaiyd14@mails.tsinghua.edu.cn

Weifeng Li
University of Georgia
Athens, GA
weifeng.li@uga.edu

Abstract

The recent development of deep learning has achieved the state-of-the-art performance in various machine learning tasks. The IS research community has started to leveraged deep learning-based text mining for analyzing textual documents. The lack of interpretability is endemic among the state-of-the-art deep learning models, constraining model improvement, limiting additional insights, and prohibiting adoption. In this study, we propose a novel text mining research framework, Neural Topic Embedding, capable of extracting useful and interpretable representations of texts through deep neural networks. Specifically, we leverage topic modeling to enrich deep learning data representations with meaning. To demonstrate the effectiveness of our proposed framework, we conducted a preliminary evaluation experiment on a testbed of fake review detection and our interpretable representations improves the state-of-the-art by almost 8 percent as measured by F1 score. Our study contributes to the IS community by opening the gate for future adoption of the state-of-the-art deep learning methods.

Keywords: deep learning interpretability; topic modeling; text mining; fake review detection

Introduction

Deep learning is the state-of-the-art machine learning method that builds upon large-scale neural networks and unsupervised representation learning (Goodfellow et al. 2016). With its excellent ability to extract useful representations from complex high-dimensional data, deep learning has demonstrated its effectiveness by achieving the state-of-the-art performance in various machine learning tasks (LeCun et al. 2015). Text mining, in particular, largely benefit from the recent development of deep learning (Zhang et al. 2015). The state-of-the-art text mining framework averages 81.9% accuracy in language understanding, 93.2% F1 score in question answering, and 92.8% in named entity recognition (Devlin et al. 2018). The IS research community has started to leveraged deep learning-based text mining for analyzing textual documents (Adamopoulos et al. 2018; Li et al. 2016).

However, one major limitation of most deep learning methods and other complex methods in big data analytics is the lack of interpretability. In particular, behind the high prediction accuracy is the "black box" of complex deep neural network architecture and optimization procedure, posing challenges in explaining how the predictions were made. The lack of interpretability might lead to a number of consequences. First, the diagnosis of the method becomes problematic, constraining understanding, correction, and further improvement of the method. Second, besides the output, limited characterization of data patterns is provided, hindering the discovery of additional insights into the context domain. Third, the reliability and validity of the method are largely unclear, preventing the method from being adopted.

While there has been preliminary attempts to improve deep learning through additive feature attribution (Lundberg and Lee 2017), the technique is not directly applicable to the text mining domain. In this study,

we seek to develop an interpretable deep learning capability in the context of text mining. Specifically, we propose a novel text mining research framework, Neural Topic Embedding (NTE), capable of extracting useful and interpretable representations of texts through deep neural networks. As suggested in its name, our proposed framework leverages topic modeling to enhance data representations with meaning. Our framework is readily customizable to solve a variety of supervised learning, semi-supervised learning, and multitask learning tasks. To demonstrate the effectiveness of our proposed framework, we conducted a preliminary evaluation experiment on a testbed of customer reviews and our interpretable representations improves the state-of-the-art by almost 8 percent as measured by F1 score. Our study makes the following contributions to IS research community. First, we contribute to the IS research community by taking the first step towards incorporating interpretability into complex deep learning applications in text mining, and opening the gate for future adoption of the state-of-the-art deep learning methods. Second, our research contributes to the IS methodology by providing a state-of-the-art text mining framework for finding useful data representations. This framework can be applied to a wide variety of text mining contexts, such as financial reports, customer reviews, enterprise blogs and more. Third, our proposed framework contributes to the big data analytics with a novel model combining the advantages of the state-of-the-art deep learning-based language models and topic models.

Related Literature

The state-of-the-art text mining is generally based on two approaches: the deep learning-based approach and topic modeling-based approach. The former seeks to optimize for various discriminative tasks (e.g., sentiment analysis) by learning the most useful representations of data coupled with designing effective deep neural networks. The latter draws upon probabilistic modeling to find the distributions (e.g., topics) that could best explain the data generative process. The rest of the section details these techniques.

Deep Learning-based Text Mining

Traditional machine learning methods builds upon feature engineering to handcraft representations of data (i.e., features) that can be used in downstream tasks (e.g., classification). While interpretable, featuring engineering could be laborious and time-consuming, and handcrafted features often have a hard time capturing all possible patterns within the data (LeCun et al. 2015). By contrast, deep learning builds upon unsupervised representation learning, where useful representations of data are automatically extracted using deep neural networks, such as convolutional neural networks (CNN) and recurrent neural networks (RNN). In the context of deep learning-based text mining, the goal of unsupervised representation learning is performing language modeling to find distributed representations that capture word semantics. The distributed word representations prove to advance the performance of various text mining tasks, including POS tagging, chunking, NER, semantic role labeling, and syntactic parsing (Collobert et al. 2011).

In the rapidly growing field of deep learning-based language modeling, we have observed three major trends. First, we find that the majority of language models leverage word sequence to extract word semantics. Earlier models leverage shallow sequential models (e.g., skipgram) to capture the semantics of each word with its neighboring words (Mikolov et al. 2013). Recently, deep neural networks such as CNN and RNN were proposed to be used as the foundation of neural language models. While CNNs are not sequential in nature, CNN-based neural language models incorporate word order through N-gram or position encoding and have demonstrated state-of-the-art performance (Devlin et al. 2018; Radford et al. 2018). RNN is a type of deep neural network where each neuron unit is connected to the next successive neuron unit, forming a sequence of neurons. In LSTM, the state-of-the-art RNN, each neuron unit is a cell comprising an input gate, an output gate, and a forget gate, regulating the flow of information along the sequence and enabling the cells to "remember" information over arbitrary length (Goodfellow et al. 2016). As such, LSTM proves to be an excellent baseline model for neural language models (McCann et al. 2017; Peters et al. 2018). Second, there has been a growing trend for incorporating context vector for inferring word semantics. One major limitation of past distributed representations is the inability to model word polysemy. Specifically, each word can only have one distributed representation regardless of the context. Recent models have therefore resolved this issue by incorporating contextual information from the sentence. One common approach is the context vector introduced in McCann et al. (2017). In this study, a separate LSTM is used to generate a context vector that captures information from the contextual sentence. This approach attained better performance in downstream tasks than methods using distributed word

representations alone. Third, while language modeling are generally unsupervised tasks, the resulting distributed representations are often used for downstream supervised learning tasks. Among the state-of-the-art neural language models, there are generally two approaches for improving downstream task performance. One approach is training task-specific distributed representations (McCann et al. 2017; Peters et al. 2018). This approach is essentially a supervised language model learning distributed representation that would best improve the specified downstream task. The other approach is the pre-training/fine-tuning paradigm, where a base neural network pre-trained to capture corpus representations in an unsupervised fashion and further fine-tuned for specific downstream tasks (Devlin et al. 2018; Radford et al. 2018). Compared to the task-specific approach, the pre-training/fine-tuning paradigm captures more generic information and is more generalizable to multitask learning, where a number of variable of interests are to be predicted.

However, one major limitation of the existing neural language models is the lack of interpretability. While the distributed representation capture word semantics, it is difficult for human to interpret the meaning from distributed representations. Specifically, it is often unclear what meaning each dimension of the representation captures. Consequently, interpreting the results of deep learning-based text mining becomes be challenging. Even when the distributed representation has compositionality, where the relative position of words in the distributed representation space suggests their relationships, there is limited interpretability because the positions of words cannot provide a straightforward explanation for its semantics. In the context of IS, such limitation not only prohibits us from gaining business insights but also discourage the adoption of the state-of-the-art deep learning methods.

As such, an interpretable distributed representation can help. While the topic word embedding generates word embeddings just like others, the attempt to assign each word embedding with a topic indicating its meaning seems to have promising results. Overall, topic modeling provides meaningful representations of texts; therefore, we review topic modeling literature in the next subsection.

Topic Modeling

Drawing upon generative modeling, topic modeling provides highly interpretable results such as topics and topic mixtures. The IS research community has extensively leveraged topic modeling for extracting insights in many contexts (Debortoli et al. 2016). The standard topic model mainly captures the underlying semantic themes, known as topics, which can be further leveraged to interpret documents (Blei et al. 2003). Specifically, each topic $\beta_k \in \{\beta_1, \dots, \beta_K\}$ is modeled as a multinomial distribution p_{β_k} over the vocabulary $W = \{w_1, w_2, \dots, w_J\}$, representing a coherent semantic meaning with the probability of each word $p_{\beta_k}(w_j)$ suggesting the word's importance within the topic. Each document $d \in D$ is then modeled as a topic mixture $\boldsymbol{\theta}_d = [\theta_d^{(1)}, \dots, \theta_d^{(K)}]$, accounting for the semantic composition of the document. The probability of each topic $\theta_d^{(k)}$ in the topic mixture suggests the topic's prevalence within the document. Within each document d, each word d_n is explained by its corresponding topic β_{z_n} generated from the topic mixture $z_n \sim \boldsymbol{\theta}_d$.

Topic modeling offers interpretability by providing a number of posteriors for directly explaining the underlying structure of the data. Let the posterior topic $p(\boldsymbol{\beta}|D)$ suggest the underlying topics of the document collection; the posterior topic mixture $p(\boldsymbol{\theta}_d|D)$ suggest the semantic composition of the document d; and, the posterior topic of each word $p(z_{d,n}|D)$ suggest the topic of d_n. The posteriors of topic models are inferred by fitting data to probabilistic models with two approaches: the *sampling* approach and the *variational* approach. The *sampling* approach rely on conditional distributions of the probabilistic model to iteratively draw samples, which are further used in the approximated integral of the posteriors (Neal 1993). In general, the *sampling* approach is computationally costly as thousands of iterations over the entire dataset are needed for the posteriors to be reasonably approximated. The *variational* approach, on the other hand, seeks to optimize variational distributions to approximate true posteriors (Wang and Blei 2013). Replacing sampling with optimization significantly reduces the computational complexity for inference; however, variational distributions could be difficult to optimize if the probabilistic model is nonconjugate (Wang and Blei 2013). With the recent development in deep learning, a novel neural variational inference approach, variational autoencoder (VAE), has been proposed to optimize posteriors in arbitrarily complex probabilistic models with high efficiency, giving rise to a number of neural variational topic models (Kingma and Welling 2014; Miao et al. 2017). VAE is based on the idea of deep autoencoder, which is a type of neural network for learning useful low-dimensional representations of data with two symmetric neural networks, an encoder network and a decoder network. The encoder network learns succinct representations by reducing dimensionality of original data and the decoder network seeks to reconstruct the original data from the low-dimensional representations. As a whole, the deep autoencoder finds low-dimensional representations that contain sufficient information to reconstruct the original data. As posteriors of latent variables are essentially succinct representations of data given the probabilistic model, a VAE obtains posteriors with two neural networks: an inference network and a reconstruction network. The inference network approximates posteriors of the latent variables with a neural network and a reconstruction network to generate data from the approximated posteriors following the probabilistic model. As neural networks are capable of approximating a wide variety of functional forms, the posteriors can be reasonably approximated despite the complexity of the probabilistic model. Hence, the VAE inference has the potential to solve many different topic model variants, regardless of conjugacy. To date a number of neural variational topic models have been proposed, we demonstrate the details of these models in Table 1.

From the review of existing neural topic models, we have three major findings. First, consistent with traditional topic modeling, documents are treated as "bag-of-words" and the word order in original documents are generally ignored. In particular, the inference network employed by previous neural topic models are traditional feed forward neural networks and multilayer perceptrons. These networks are not inherently equipped to model sequential data, as compared to sequence neural networks such as LSTM. As

Table 1. A Summary Major Neural Topic Models

Study	Inference Network	Main Latent Variable	Reconstruction Network	Supervised?
SCHOLAR (Card et al. 2018)	Multilayer Perceptron	Document topic mixture $\boldsymbol{\theta}$ with logistic normal distribution	Sparse Additive Generative model (SAGE) + Multilayer Perceptron	Yes
NTM (Miao et al. 2017)	Multilayer Perceptron	Document topic mixture $\boldsymbol{\theta}$ modeled through Gaussian Softmax	LDA with marginalized word topic assignment $\boldsymbol{z}$	No
TopicRNN(Dieng et al. 2017)	Feed forward neural network	Document topic mixture $\boldsymbol{\theta}$	RNN + LDA	No
AVITM (Srivastav & Sutton 2017)	Feed forward neural network	Document topic mixture $\boldsymbol{\theta}$ with logistic normal distribution	LDA with marginalized word topic assignment $\boldsymbol{z}$	No

mentioned previously, the word sequence contains dependency information that could help infer word semantics and has been extensively exploited in the deep learning-based language modeling literature. Equipping neural topic models with inference networks capable of modeling sequential data has the potential to further improve the approximation of posteriors (Dieng et al. 2017). Second, the main latent variable whose posteriors are approximated by the inference network are consistently document topic mixture$\boldsymbol{\theta}$. This is mainly for simplicity as the unit of analysis in topic models are documents and recovering each document from its topic mixture seems straightforward for reconstruction networks. Moreover, most neural topic models deliberately chose $\boldsymbol{\theta}$ to be logistic normal or generated from Gaussian softmax transformation as compared to the Dirichlet distribution in traditional topic models. This is because Dirichlet distribution is inherently incompatible with the variational autoencoder (Srivastava and Sutton 2017). However, the choice of main latent variable being the document topic mixture$\boldsymbol{\theta}$ further prohibits neural topic models from accounting for the dependencies among words, because reconstructing the original word sequence from topic mixture $\boldsymbol{\theta}$ is difficult. To enable neural topic model to accommodate word sequences, the main latent variable needs to be carefully selected and the word-level latent variables are particularly preferred. Third, with few exception, most existing models are unsupervised topic models. While these models have demonstrated improved results in modeling corpora, the resulting topics have the potential to enhance other downstream tasks such as text classification. Card et al. (2018) provided a set of guidelines for developing supervised neural topic model and demonstrated that the topics could further help with many predictive applications. However, the effectiveness of these models in supervised downstream tasks needs further exploration.

The state-of-the-art text mining method mostly builds upon neural language models and topic models. On one hand, the neural language models largely rely on sequential dependencies to capture word semantics in its resultant distributed representations and have demonstrated advanced effectiveness in various downstream tasks; however, interpreting these distributed representations is challenging, hindering the extraction of business insights. On the other hand, topic models have been extensively used as an interpretable text mining tools as it provides plausible explanations for the observed documents, but even with the state-of-the-art neural topic model, the fundamental "bag-of-words" perspective is prohibiting topic models to learn more informative topics by accounting for word sequential dependencies.

The Neural Topic Embedding (NTE) Text Mining Framework

In this study, we propose the development of an interpretable text mining framework that combines the advantages of the state-of-the-art neural language models and topic models. Specifically, our proposed framework seeks to achieve interpretable distributed representations of words by leveraging the state-of-the-art neural topic modeling. Moreover, our framework extends neural topic modeling by taking word sequential dependencies into consideration, thereby generating more informative topics.

Drawing upon the variational autoencoder (VAE) framework, our proposed method comprises a novel inference network and a reconstruction network. Our innovative inference network is capable of simultaneously accommodating the word sequence and modeling topics, thereby generating more informative distributed representations that can be interpreted using the topical semantics. The reconstruction network recovers the original word sequence to ensure the inference network is effectively trained. Moreover, to achieve the best performance for supervised downstream tasks, our proposed framework can be pre-trained using unlabeled corpora and further fine-tuned for downstream tasks using labeled datasets. The rest of the section describe our proposed framework in detail.

Inference Network

In the variational autoencoding (VAE) framework, the inference network mainly seeks to approximate the posterior distribution of the chosen main latent variable with deep neural networks. Instead of inferring document topic mixture $\boldsymbol{\theta}$ as the extant work, our inference network mainly approximates the sequential word topic assignment $\{z_1, \dots, z_{dn}\}$. Our choice of word topic assignment as the main latent variable has at least two notable advantages. First, as mentioned previously, having document topic mixture $\boldsymbol{\theta}$ as the main latent variable prohibits the modeling of sequential dependencies. Our selection of word topic assignment enables the inference network to model the dependency information embedded in word sequence, thereby improving the quality of our resultant data representation. Second, the resulting word topic assignment is a distributed representation of each word, with each dimension suggesting the word's association with the corresponding topic. As compared to the state-of-the-art distributed representations, our distributed representations are more interpretable: each representation quantifies semantics by showing the words associations with each semantic topic, and can therefore have readily explained by topics. Following prior literature (Card et al. 2018), the prior word topic assignment z_n follows the logistic normal distribution and is parameterized by mean $\boldsymbol{\mu}_n$ and standard deviation $\boldsymbol{\sigma}_n$.

To allow for modeling sequential dependencies, our inference network is inspired by the state-of-the-art language models, where dependency information is captured by sequence neural networks such as LSTM. In alignment with deep learning-based language models, our inference network comprises a multilayer LSTM incorporated with a contextualized design. Specifically, our inference network has a two-layer LSTM architecture. The first layer LSTM provides the context vector that is essentially a summary of the input document for contextualizing word semantics. The second layer has two LSTMs, both of which builds upon the context vector to infer topic-based word semantics. One LSTM approximates the mean of word topic assignment $\{\mu_1, \dots, \mu_{dn}\}$ and the other LSTM approximates the standard deviation of word topic assignment $\{\sigma_1, \dots, \sigma_{dn}\}$, where d is the input document and dn is the number of words in the document. With these two LSTMs, both μ_n and σ_n to relate to μ_{n-1} and σ_{n-1}, respectively. The sequential dependencies between neighboring words and their topics can be therefore modeled. To operationalize the logistic normal distribution, we leverages a Gaussian softmax transformation where we introduce an auxiliary variable x_n

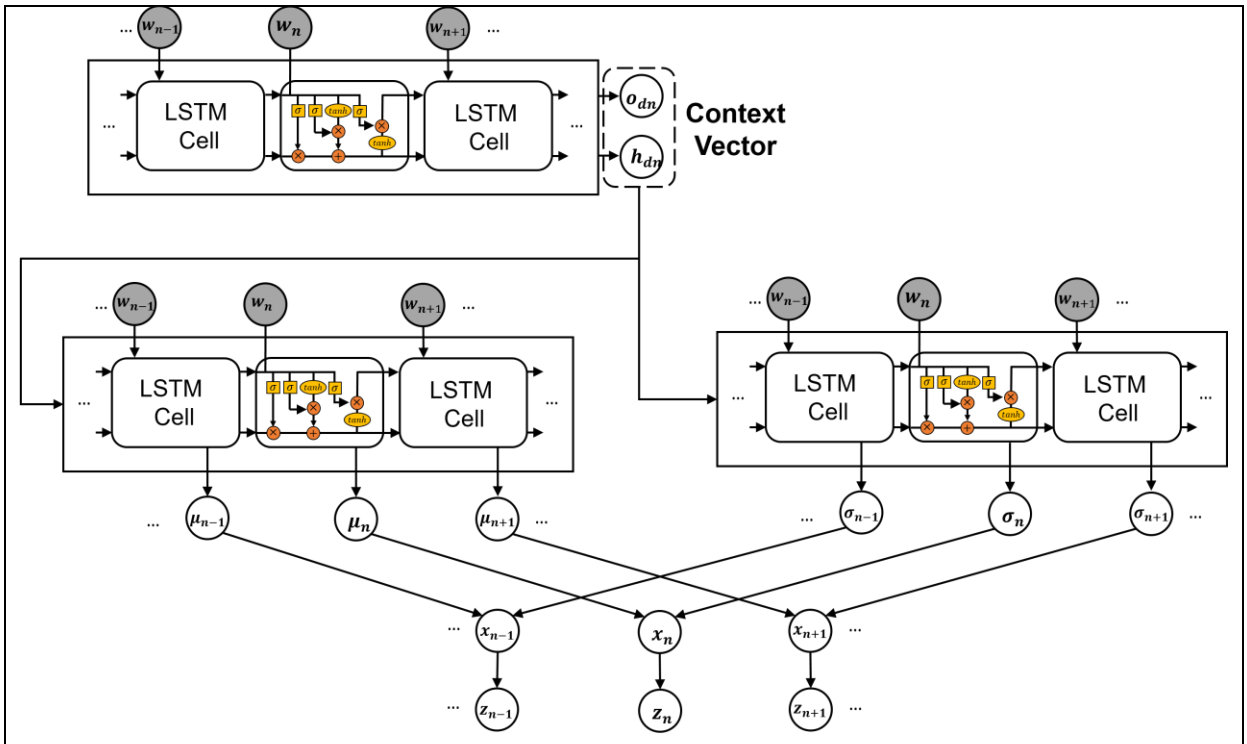

Figure 1. Architecture of Inference Network

as the un-normalized topic vector $\boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n)$, and $\boldsymbol{z}_n = SoftMax(\boldsymbol{x}_n)$. To further illustrate, the architecture of the inference network is depicted in Figure 1.

Reconstruction Network and Extension to Supervised Learning

The goal of the reconstruction network is to recover the original input from the main latent variable to ensure the effectiveness of the inference network in approximating posteriors. Our reconstruction network mainly recovers each individual word w_n based on its approximated topic assignment z_n. Unlike previous neural topic models that generate bag-of-words in reconstruction networks, our reconstruction network generates the original word sequence. This design enables our proposed VAE to train the inference network to learn sequential dependencies in the input documents and effectively generates distributed representations (i.e., word topic assignment) that captures not only topical semantics but also sequential dependencies. In accordance with the topic modeling convention, we model the generation of each word as $w_n|z_n \sim \text{Multinomial}(\boldsymbol{\beta}_{z_n})$, where $\boldsymbol{\beta}$ is the topic matrix with each column being a topic and $\boldsymbol{\beta}_{z_n}$ is the z_n-th topic represented as a multinomial distribution over the vocabulary. As mentioned previously, to generat word topic assignment, we first generate the auxiliary variable $\boldsymbol{x_n} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n)$, followed by $\boldsymbol{z_n} = SoftMax(\boldsymbol{x_n})$.

Our framework can be further leveraged for making predictions in supervised learning contexts. Predicting variables of interest y_d can be based on the word topic assignment $\{z_1, \dots, z_{dn}\}$: $p(y_{dn}|w_1, \dots, w_{dn}) = p(y_{dn}|z_1, \dots, z_{dn})$. This design is consistent with the conventional deep learning-based text mining where distributed representations (in our case, word topic assignment) are further used as low-dimensional representation of data for downstream tasks. Additionally, using word topic assignment for making prediction is also a common practice among existing supervised topic models. A number of sequential classifiers could be used here to model $p(y_d|z_1, \dots, z_{dn})$. For evaluation purposes, we use LSTM to model $p(y_d|z_1, \dots, z_{dn})$ in our study such that we can show the improvement of our proposed framework with the state-of-the-art baseline sequential classification algorithm.

Our proposed framework brings the interpretability in at least two ways. First, the proposed framework is capable of automatically extracting word dependency-enhanced topics from a corpus. The word dependency-enhanced topics suggest the semantic themes underlying the corpus, which could help interpret the corpus through a topical lens. By accounting for word dependencies among the input words, the topics generated by our proposed framework are capable of capturing more meaningful information as compared to topics extracted from traditional topic models. Moreover, our proposed framework can show how topics are generally distributed throughout the corpus, from which one can gain more insights. Second, our framework models documents with the main latent variable of word topic assignment that is readily interpretable. For each word, the word topic assignment z_n is a distribution over topics. In particular, the word topic assignment is essentially a word embedding $z_n = [z_n^{(1)}, \dots, z_n^{(K)}]$, where K is the number of topics and $z_n^{(k)}$ is the probability of word w_n has topic β_k. In comparison with the existing neural word embedding methods, the word embedding generated by our framework is a K-dimensional vector, with each dimension suggesting the word semantics' relationship to the corresponding topic.

Preliminary Evaluation and Results

To demonstrate the effectiveness of our proposed framework, we conduct preliminary evaluation in the context of fake review detection. Fake review detection has been a research question of interest to the IS research community for the past years (Kumar et al. 2018). Our testbed is mainly collected from the deceptive opinion spam dataset, which includes 800 real and 800 fake reviews for 20 Chicago hotels. We first compare our proposed method against traditional supervised topic models to demonstrate the effectiveness of incorporating sequential dependencies in texts. Then, we compare the proposed method against the state-of-the-art neural language models to show the efficacy of incorporating the topic-based interpretability.

In the first experiment, the baseline supervised topic models include the Supervised Latent Dirichlet Allocation (SLDA) with varying topic numbers. Since our proposed method is coupled with LSTM, we denote the mode as NTE-LSTM. The comparison results are summarized in Table 2.

Table 2. Fake Review Detection Results Using Conventional Topic Models				
Models	Accuracy	Precision	Recall	F1-score
SLDA-10	0.4994***	0.5846*	0.2694**	0.2571***
SLDA-20	0.5013***	0.5269***	0.2887**	0.2828***
SLDA-30	0.4975***	0.5025***	0.3304**	0.3343***
SLDA-40	0.5094***	0.5141***	0.2471***	0.2938***
NTE-LSTM	**0.8157**	**0.8066**	**0.8767**	**0.8389**

Note: * p<0.05, ** p<0.01, and *** p<0.001; SLDA-n refers to the SLDA with n topics.

Experimental results summarized in Table 2 indicate that our model (i.e., the NTE-LSTM) achieved highest value in accuracy, precision, recall, and F1-score, with 0.8157, 0.8066, 0.8767 and 0.8389 respectively. The results also indicate that 81.59% of the reviews were identified correctly; 80.66% of the reviews that identified as fake reviews were actually fake; 87.67% of the fake reviews were identified correctly. Moreover, in comparison with other configurations, our model improved the performances by statistically significant margins. In particular, in all cases, p-values for Accuracy and F1-score were less than 0.001; p-values for recall were less than 0.01; and p-values for precision were less than 0.05. Therefore, our model is more effective for fake review detection, demonstrating the effectiveness of accommodating for the sequential dependencies among words.

In the second experiment, the baseline deep models include the state-of-the-art RNN variants: Long short-term memory (LSTM) and Gated recurrent unit (GRU), both of which are widely used for text classification task. The comparison results are summarized in Table 3.

Table 3. Fake Review Detection Results Using Deep Learning Models				
Models	Accuracy	Precision	Recall	F1-score
LSTM-1	0.7963	0.7972	0.7921**	0.7927*
LSTM-2	0.7613	0.7403	0.8058	0.7703*
GRU-1	0.7588	0.7354	0.8134	0.7701*
GRU-2	0.7488*	0.7374*	0.772**	0.7538**
NTE-LSTM	**0.8157**	**0.8066**	**0.8767**	**0.8389**

Note: * p<0.05, ** p<0.01; LSTM-n refers to the LSTM with n layers; Similar for GRU-n.

Similarly, results in Table 3 show our model achieved the highest values in all measures. Moreover, our model outperformed the standard deep learning models by statistically significant margins. The p-values for F1-score were less than 0.05 in all cases; p-values for Recall were less than 0.001 in LSTM-1 and GRU-2; and p-value for Accuracy was less than 0.05 in GRU-2. Overall, our model is superior to the state-of-the-art recurrent neural networks in the context of fake review detection, suggesting the improvement due to the incorporation of topic-based interpretability.

Conclusions and Future Directions

The lack of interpretability is endemic among the state-of-the-art deep learning models, constraining model improvement, limiting additional insights, and prohibiting adoption. Our research seeks to address this issue in the context of text mining. In particular, our proposed text mining framework combines the advantages of the state-of-the-art deep learning and topic modeling, providing useful and interpretable representations of texts. We demonstrate the effectiveness of our proposed framework in the testbed of fake review detection. In the future, we plan to expand our evaluation to a number of other text mining tasks and further assess the interpretability of our proposed method.

References

Adamopoulos, P., Ghose, A., and Todri, V. 2018. "The Impact of User Personality Traits on Word of Mouth: Text-Mining Social Media Platforms. Information Systems Research Published Online in Articles in Advance," *Information Systems Research* (29:3), pp. 612–640. (https://doi.org/10.1287/isre.2017.0768).

Blei, D. M., Ng, A., and Jordan, M. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993–1022.

Card, D., Tan, C., and Smith, N. A. 2018. "Neural Models for Documents with Metadata," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1), pp. 2031–2040.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. 2011. "Natural Language Processing (Almost) from Scratch," *J. Mach. Learn. Res.* (12), JMLR.org, pp. 2493–2537.

Debortoli, S., Müller, O., Junglas, I., and Vom Brocke, J. 2016. "Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the Association for Information Systems* (39), pp. 110–135.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *ArXiv Preprint ArXiv:1810.04805*.

Dieng, A. B., Wang, C., Gao, J., and Paisley, J. 2017. "TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency," in *ICLR*, pp. 1–13.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, MIT Press.

Kingma, D. P., and Welling, M. 2014. "Auto-Encoding Variational Bayes," in *Proceedings of International Conference on Learning Representations (ICLR)*. (https://doi.org/10.1051/0004-6361/201527329).

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. "Detecting Review Manipulation on Online Platforms with Hierarchical Supervised Learning," *Journal of Management Information Systems* (35:1), Routledge, pp. 350–380. (https://doi.org/10.1080/07421222.2018.1440758).

LeCun, Y., Bengio, Y., and Hinton, G. 2015. "Deep Learning," *Nature* (521:7553), pp. 436–444. (https://doi.org/10.1038/nature14539).

Li, W., Chen, H., and Nunamaker, J. F. 2016. "Identifying and Profiling Key Sellers in Cyber Carding Community: AZSecure Text Mining System," *Journal of Management Information Systems* (33:4), Routledge, pp. 1059–1086. (https://doi.org/10.1080/07421222.2016.1267528).

Lundberg, S. M., and Lee, S.-I. 2017. "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Curran Associates, Inc., pp. 4765–4774.

McCann, B., Bradbury, J., Xiong, C., and Socher, R. 2017. "Learned in Translation: Contextualized Word Vectors," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Curran Associates, Inc., pp. 6294–6305.

Miao, Y., Grefenstette, E., and Blunsom, P. 2017. "Discovering Discrete Latent Topics with Neural Variational Inference," in *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70), Proceedings of Machine Learning Research, D. Precup and Y. W. Teh (eds.), International Convention Centre, Sydney, Australia: PMLR, pp. 2410–2419.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. 2013. "Distributed Representations of Words and Phrases and Their Compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Neal, R. M. 1993. "Probabilistic Inference Using Markov Chain Monte Carlo Methods," *Probabilistic Inference Using Markov Chain Monte Carlo* (September).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. 2018. "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (Vol. 1), pp. 2227–2237.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. 2018. "Improving Language Understanding by Generative Pre-Training,".

Srivastava, A., and Sutton, C. 2017. "Autoencoding Variational Inference For Topic Models," in *Proceedings of ICLR*.

Wang, C., and Blei, D. M. 2013. "Variational Inference in Nonconjugate Models," *Journal of Machine Learning Research* (14:1), pp. 1005–1031.