the distribution of true and generated samples [8]. In this paper, inspired by the idea of the discriminator in GANs, we propose to distinguish between real and synthetic reviews based on distributed representation of their constituent tokens. It should be noted that the method proposed in [52] to defend against model-generated reviews, examines the character distribution of synthetic and real reviews as the language model is trained at character level granularity. Therefore, the direct application of this approach on word-level generators is ruled out.

7 CONCLUSION AND FUTURE WORK

We proposed and evaluated a wide-ranging class of attacks on online review platforms based on neural language models at word-level granularity using transfer-learning. The unique attribute of our work is being domain independent and can target any arbitrary review domain even with small available review samples. The main intuition is to: (i) develop a universal model to learn general linguistic patterns in review domain and transfer this knowledge to the domain-specific language; (ii) generate high quality reviews which are competitive with real reviews and can pass the quality test by both computational-based detectors and human evaluators; (iii) demonstrate that synthetic reviews do not completely mimic the true distribution of real reviews, so this is a powerful signal to detect automated fake reviews. Our results on discriminating generated reviews are promising. In our ongoing work, we aim to study the performance of other neural network architectures like CNN in modeling synthetic review distributions and to develop a more powerful discriminator.

Acknowledgement. This work was supported in part by NSF grant SaTC-1816497. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Roee Aharoni and et al. 2014. Automatic detection of machine translated text and translation quality estimation. In *ACL*.
[2] Ebru Arisoy, Tara N Sainath, Brian Kingsbury, and Bhuvana Ramabhadran. 2012. Deep neural network language models. In *NAACL-HLT*. ACL.
[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.. In *Lrec*.
[4] Amir Fayazi, , and et al. 2015. Uncovering crowdsourced manipulation of online reviews. In *SIGIR*.
[5] William Fedus and et al. 2018. Maskgan: Better text generation via filling in the _. *ICLR* (2018).
[6] Daniela Gerz and et al. 2018. On the Relation between Linguistic Typology and (Limitations of) Multilingual Language Modeling. In *EMNLP*.
[7] Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* (2013).
[8] Ian Goodfellow and et al. 2014. Generative adversarial nets. In *NIPS*.
[9] Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv* (2013).
[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).
[11] Bryan Hooi and et al. 2016. Birdnest: Bayesian inference for ratings-fraud detection. In *ICDM*.
[12] Dirk Hovy. 2016. The enemy in your own camp: How well can we detect statistically-generated fake reviews–An adversarial study. In *ACL*.
[13] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*.
[14] Hakan Inan and et al. 2016. Tying word vectors and word classifiers: A loss framework for language modeling. *arXiv* (2016).
[15] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *WSDM*.
[16] Mika Juuti and et al. 2018. Stay on-topic: Generating context-specific fake restaurant reviews. In *ESORICS*.
[17] Parisa Kaghazgaran and et al. 2017. Behavioral analysis of review fraud: Linking malicious crowdsourcing to amazon and beyond. In *ICWSM*.
[18] Parisa Kaghazgaran and et al. 2018. Combating crowdsourced review manipulators: A neighborhood-based approach. In *WSDM*.
[19] Parisa Kaghazgaran and et al. 2019. TOmCAT: Target-Oriented Crowd Review ATtacks and Countermeasures. In *ICWSM*.
[20] Anjuli Kannan and et al. 2016. Smart reply: Automated response suggestion for email. In *KDD*.
[21] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
[22] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *Social Media Analytics: Advances and Applications* (2018).
[23] Shanshan Li and et al. 2017. Crowdsourced App Review Manipulation. In *SIGIR*.
[24] Zachary C Lipton and et al. 2015. Capturing meaning in product reviews with character-level generative text models. *arXiv* (2015).
[25] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* (2016).
[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* (2008).
[27] Mitul Makadia. 2018. What Are the Advantage of Natural Language Generation and Its Impact on Business Intelligence?, https://www.marutitech.com/advantages-of-natural-language-generation/, Last Access: 01/28/2019. (2018).
[28] Julian McAuley and et al. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*.
[29] Gábor Melis and et al. 2017. On the state of the art of evaluation in neural language models. *ICLR* (2017).
[30] Stephen Merity and et al. 2017. Regularizing and optimizing LSTM language models. *arXiv* (2017).
[31] Stephen Merity and et al. 2018. An Analysis of Neural Language Modeling at Multiple Scales. *arXiv* (2018).
[32] Tomáš Mikolov and et al. [n.d.]. Empirical evaluation and combination of advanced language modeling techniques.
[33] Tomáš Mikolov and et al. 2010. Recurrent neural network based language model. In *ISCA*.
[34] Tomas Mikolov and et al. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.
[35] Sewon Min and et al. 2017. Question answering through transfer learning from large fine-grained supervision data. *arXiv* (2017).
[36] Arjun Mukherjee and et al. 2013. What yelp fake review filter might be doing?. In *ICWSM*.
[37] Hoang-Quoc Nguyen-Son and et al. 2017. Identifying computer-generated text using statistical analysis. In *APSIPA ASC*.
[38] Matthew Peters and et al. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv* (2017).
[39] Matthew E et al. Peters. 2018. Deep contextualized word representations. *arXiv* (2018).
[40] Jakub Piskorski and et al. 2008. Exploring linguistic features for web spam detection: a preliminary study. In *AIRWeb*.
[41] Alec Radford and et al. 2017. Learning to generate reviews and discovering sentiment. *arXiv* (2017).
[42] Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *KDD*.
[43] Ehud Reiter and et al. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence* (2005).
[44] Rico Sennrich and et al. 2015. Improving neural machine translation models with monolingual data. *ACL* (2015).
[45] Iulian Vlad Serban and et al. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR, abs/1507.04808* (2015).
[46] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv* (2015).
[47] Kijung Shin, , and et al. 2017. D-cube: Dense-block detection in terabyte-scale tensors. In *WSDM*.
[48] Ilya Sutskever and et al. 2011. Generating text with recurrent neural networks. In *ICML*.
[49] Gang Wang, , and et al. 2012. Serf and turf: crowdturfing for fun and profit. In *WWW*.
[50] Anbang Xu and et al. 2017. A new chatbot for customer service on social media. In *CHI*.
[51] Arun Kumar Yadav and Samir Kumar Borgohain. 2014. Sentence generation from a bag of words using N-gram model. In *ICACCCT*.
[52] Yuanshun Yao and et al. 2017. Automated crowdturfing attacks and defenses in online review systems. In *CCS*.
[53] Jason Yosinski and et al. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*.