CSC2412: Project Proposal

Grigory Dorodnov David Landsman

October 23, 2020

With the advent of deep learning and attention-based transformer models, the field of natural language processing (NLP) is seeing an incredible improvement in the performance of language models for a variety of tasks, including information retrieval [1, 2, 3], question answering systems [4, 5], machine translation [6, 7, 8] and much more. As applications of NLP begin to extend to many different areas which make use of highly sensitive and personal data such as healthcare [9], finance [10] and personal messaging [11], it is vital to consider aspects of security and privacy in deployed models. The main question we aim to address in this project is: given the tools provided by differential privacy (DP), are we able to balance the great performance of NLP models with a strong and robust approach to privacy, in order to ensure the privacy of personal data used to train and deploy these models?

Our research goals for the project are: (1) review and summarize literature for current approaches in DP language models and word embeddings; (2) analyze and compare performance of language models and word embeddings (including neural network models and statistical models) for a specific task such as sentiment analysis or text classification, in the presence of classical differential privacy methods (for neural models - private gradient descent, for statistical models - basic mechanisms such as Laplace noise); and (3) investigate novel ways to integrate differential privacy into language models and word embeddings to improve their performance in particular tasks.

Differential privacy in NLP models is a relatively new area of research, which is not yet supported by an abundant amount of articles. Much of the research in privacy-preserving NLP focuses on incorporating cryptographic techniques such as fully homomorphic encryption [12] or secure multi-party computation [13] into the training procedure, which can lead to slower models, especially when training on large datasets. However, a decent growth of interest in applying DP methods to preserve privacy in NLP can be observed in recent years. In the following we highlight some important and interesting papers in this area. *Fernandes et al.* [14, 15] focus their attention on the problem of authorship obfusctaion and make use of the Laplace mechanism to introduce privacy in their model. *Li et al.* [16] explore an alternative method to achieve authorship obfuscation by learning text representations that are invariant to author characteristics. *Pan et al.* [17] design and evaluate novel privacy attacks on state-of-the-art NLP models including BERT and GPT, and propose a few defenses to protect models against privacy leaks. *McMahan et al.* [18] and *Li et al.* [19] propose methods for training differentially private recurrent neural language models. Finally, *Adelani et al.* [20] derive formal privacy guarantees for a general text de-identification method.

References

[1] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 55–64, 2016.

[2] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104, 2019.

[3] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 497–506, 2018.

[4] Li Dong, Furu Wei, Ming Zhou, and Ke Xu. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 260–269, 2015.

[5] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[8] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[9] Chengyi Zheng, Wei Yu, Fagen Xie, Wansu Chen, Cheryl Mercado, Lina S Sy, Lei Qian, Sungching Glenn, Gina Lee, Hung Fu Tseng, et al. The use of natural language processing to identify tdap-related local reactions at five health care systems in the vaccine safety datalink. *International journal of medical informatics*, 127:27–34, 2019.

[10] Craig Lewis and Steven Young. Fad or future? automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, 49(5):587–615, 2019.

[11] Brant Chee, Richard Berlin, and Bruce Schatz. Measuring population health using personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2009, page 92. American Medical Informatics Association, 2009.

[12] Ahmad Al Badawi, Luong Hoang, Chan Fook Mun, Kim Laine, and Khin Mi Mi Aung. Privft: Private and fast text classification with homomorphic encryption, 2019.

[13] Q. Feng, D. He, Z. Liu, H. Wang, and K. R. Choo. Securenlp: A system for multi-party privacy-preserving natural language processing. *IEEE Transactions on Information Forensics and Security*, 15:3709–3721, 2020.

[14] Natasha Fernandes, Mark Dras, and Annabelle McIver. Author obfuscation using generalised differential privacy. *arXiv preprint arXiv:1805.08866*, 2018.

[15] Natasha Fernandes, Mark Dras, and Annabelle McIver. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham, 2019.

[16] Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. *arXiv preprint arXiv:1805.06093*, 2018.

[17] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE, 2020.

[18] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017.

[19] Xinyi Li, Yinchuan Li, Hongyang Yang, Liuqing Yang, and Xiao-Yang Liu. Dp-lstm: Differential privacy-inspired lstm for stock prediction using financial news. *arXiv preprint arXiv:1912.10806*, 2019.

[20] David Ifeoluwa Adelani, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. Privacy guarantees for de-identifying text transformations. *arXiv preprint arXiv:2008.03101*, 2020.