A model to classify financial texts while protecting users' privacy

13 October 2021, by Ingrid Fadelli

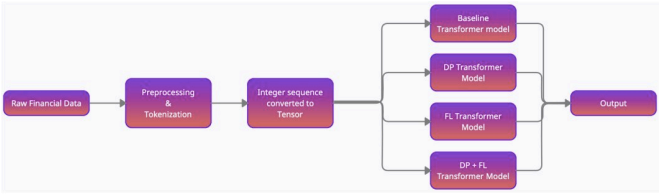

Diagram summarizing the pipeline of the model devised by the researchers. Credit: Basu et al.

Over the past decade or so, computer scientists developed a variety of machine learning (ML) models that can analyze large amounts of data both quickly and efficiently. To be applied in real-world situations that involve the analysis of highly sensitive data, however, these models should protect the privacy of users and prevent information from reaching third parties or from being accessed by developers.

Researchers at Manipal Institute of Technology, Carnagie Mellon University and Yildiz Technical University have recently created a privacy-enabled model for the analysis and classification of financial texts. This model, introduced in a paper pre-published on arXiv, is based on a combination of natural language processing (NLP) and machine learning techniques.

"Our paper was based on our previous work, named 'Benchmarking differential privacy and federated learning for BERT models'," Priyam Basu, one of the researchers who carried out the study, told *Tech Xplore*. "This work was our modest attempt at combining the domains of natural language processing (NLP) and privacy preserving machine learning."

The main objective of the recent work by Basu and his colleagues was to develop a NLP model that preserves the privacy of users, preventing their data from being accessed by others. Such a model could be particularly useful for the analysis of bank statements, tax returns and other sensitive financial documents.

"Machine Learning is majorly based on data and gives you insights and predictions and information based on data," Basu said. "Hence, it is very important for us to delve into research on how to preserve user privacy at the same time."

The framework developed by Basu and his colleagues is based on two approaches known as differential privacy and federated learning, combined with bidirectional encoder representations from transformers (BERT), which are renowned and widely used NLP models. Differential privacy techniques add a certain amount of noise to the data that is fed to the model. As a result, the party processing the data (e.g., developers, tech firms or other companies) cannot gain access to the real documents and data, as individual elements are concealed.

"Federated Learning, on the other hand, is a method of training a model on multiple decentralized devices so that no one device has access to the entire data at once," Basu explained. "BERT is a language model that gives contextualized embeddings for natural language text which can be used later on multiple tasks, such as classification, sequence tagging, semantic analysis etc."

Basu and his colleagues used the strategy they developed to train several NLP models for classifying financial texts. They then evaluated these models in a series of experiments, where they used them to analyze data from the Financial Phrase Bank dataset. Their results were highly promising, as they found that the NLP models performed as well as other state-of-the-art techniques for the analysis of financial texts, while

ensuring greater data protection.

These researchers' study could have important implications for several industries, including both the financial sector and other fields that involve the analysis of sensitive user data. In the future, the new models they developed could help to significantly increase the privacy associated with NLP techniques that analyze personal and financial information.

"Classification and categorisation based on natural language data is used in a lot of domains and hence, we have provided a way to do the same while maintaining the privacy of user data, which is highly important in finance, where the data used is highly sensitive and confidential," Basu said. "We now plan to improve the accuracy achieved by our model, while not having to lose out too much on the privacy trade-off. We also hope to explore other techniques to achieve the same as well as perform other NLP tasks like NER, Semantic analysis and Clustering using DP and FL."

More information: Privacy enabled financial text classification using differential privacy and federated learning. arXiv:2110.01643 [cs.CL]. arxiv.org/abs/2110.01643

Benchmarking differential privacy and federated learning for BERT models. arXiv:2106.13973 [cs.CL]. arxiv.org/abs/2106.13973

APA citation: A model to classify financial texts while protecting users' privacy (2021, October 13) retrieved 26 October 2021 from https://techxplore.com/news/2021-10-financial-texts-users-privacy.html