Privacy enabled Financial Text Classification using Differential Privacy and Federated Learning

Priyam Basu and **Tiasa Singha Roy**
Manipal Institute of Technology
{priyam.basu1, tiasa.singharoy}@learner.manipal.edu

Rakshit Naidu
Carnegie Mellon University
rnemakal@andrew.cmu.edu

Zumrut Muftuoglu
Yildiz Technical University
zumrutmuftuoglu@gmail.com

Abstract

Privacy is important considering the financial Domain as such data is highly confidential and sensitive. Natural Language Processing (NLP) techniques can be applied for text classification and entity detection purposes in financial domains such as customer feedback sentiment analysis, invoice entity detection, categorisation of financial documents by type etc. Due to the sensitive nature of such data, privacy measures need to be taken for handling and training large models with such data. In this work, we propose a contextualized transformer (BERT and RoBERTa) based text classification model integrated with privacy features such as Differential Privacy (DP) and Federated Learning (FL). We present how to privately train NLP models and desirable privacy-utility tradeoffs and evaluate them on the Financial Phrase Bank dataset.

1 Introduction

Divulging personally identifiable information during a business transaction has become a commonplace occurrence for most individuals. This activity can span from sharing of bank account numbers, loan account numbers, and credit/debit card numbers, to providing non-financial personally identifiable information such as name, social security number, driver's license number, address, and e-mail address. Maintaining the privacy of confidential customer information has become essential for any firm which collects or stores personally identifiable data. The financial services industry operates and deals with a significant amount of confidential client and customer data for daily business transactions. Though many organizations are taking strides to improve their privacy practices, and consumers are becoming more privacy-aware, it remains a tremendous burden for users to manage their privacy (Anton et al., 2004).

NLP has major applications in the finance industry for many tasks such as detection of entities for gross tax calculation from invoice and payroll data, categorising different kinds of financial documents based on type, grouping of financial documents based on semantic similarity, sentiment analysis of financial text (Vicari and Gaspari, 2020), conversational bots for banking systems, investment recommendation engines etc.

Text Classification can be extended to many NLP applications including sentiment analysis, question answering, and topic labeling . For example, financial or government institutions that wish to train a chatbot for their clients cannot be allowed to upload all text data from the client-side to their central server due to strict privacy protection statements (Liu et al., 2021). At this point, applying the federated learning paradigm presents an approach to solve the dilemma due to its advances in privacy preservation and collaborative training where the central server can train a powerful model with different local labeled data at client devices without uploading the raw data considering increasing privacy concerns in public.

The goal of this paper is to propose a privacy enabled text classification system, combining state-of-the-art transformers (BERT and RoBERTa) with differential privacy, on both centralized and FL based setups, exploring different privacy budgets to investigate the privacy-utility trade-off and see how they perform when trying to classify financial document-based text sequences. For the federated setups, we try to explore both IID (Independent and Identically Distributed) and non-IID distributions of data.

2 Related Work

Deep learning techniques have often been used to learn text representations via neural models by language application. The input text can give us individual demographic information about the author. Sentiment analysis can be used for the classification or categorization of financial documents. Xing

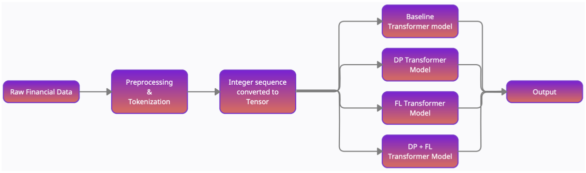
Raw Financial Data
Preprocessing & Tokenization
Integer sequence converted to Tensor
Baseline Transformer model
DP Transformer Model
FL Transformer Model
DP + FL Transformer Model
Output

Figure 1: Pipeline

et al. investigate the error patterns of some widely acknowledged sentiment analysis methods in the finance domain. Mishev et al. perform more than one hundred experiments using publicly available datasets, labeled by financial experts. In their work, Liu et al. propose a domain-specific language model pre-trained on large-scale financial corpora and evaluate it on the Financial Phrase Bank dataset. Araci presents a BERT-based model which is pre-trained on a large amount of finance-based data in his study.

Studies have been conducted on training differentially private deep models with the formal differential privacy approach in the literature (Abadi et al., 2016; McMahan et al., 2018; Yu et al., 2019). Fernandes et al. discuss the security through differential privacy in textual data. (Panchal, 2020) in his work portrays the use of DP in the generation of contextually similar messages for Honey encryption which encrypts messages using low min-entropy keys such as passwords. Federated learning is another privacy-enhancing approach (McMahan et al., 2017; Yang et al., 2019; Kairouz et al., 2021; Jana and Biemann, 2021; Priyanshu and Naidu, 2021), which relies on distributed training of models on devices and sharing of model gradients. Liu et al. show how FL can be used for decentralized training of heavy pre-trained NLP models. Basu et al. in their work have shown a detailed benchmark comparison of multiple BERT based models with DP and FL for depression detection. Jana and Biemann in their work show a differentially private sequence tagging system in a federated learning setup.

3 Dataset

The key arguments for the low utilization of statistical techniques in financial sentiment analysis have been the difficulty of implementation for practical applications and the lack of high- quality training data for building such models. Especially in the case of finance and economic texts, annotated collections are a scarce resource and many are reserved for proprietary use only. For this reason, we use the Financial Phrase Bank dataset (Malo et al., 2014) which was also used for benchmarking the pre-trained FinBERT model for sentiment analysis (Araci, 2019). The dataset includes approximately 5000 phrases/sentences from financial news texts and company press releases. The objective of the phrase level annotation task is to classify each example sentence into a positive, negative or neutral category by considering only the information explicitly available in the given sentence. Since the study is focused only on financial and economic domains, the annotators were asked to consider the sentences from the viewpoint of an investor only; i.e. whether the news may have a positive, negative or neutral influence on the stock price. As a result, sentences that have a sentiment that is not relevant from an economic or financial perspective are considered neutral.

Given a large number of overlapping annotations (5 to 8 annotations per sentence), there are several ways to define a majority vote-based gold standard. To provide an objective comparison, the authors have formed 4 alternative reference datasets based on the strength of majority agreement. For the purpose of this task, we use those sentences with 75% or more agreement. The final dataset has 3453 sentences in total out of which 60% belong to the neutral class, 28% belong to the positive class and 12% belong to the positive class.

4 Preliminaries

Today, the text is the most widely used communication instrument.For years, researchers are studies focusing on implementing different approaches that make possible machines to imitate human reading (Ly et al., 2020). Natural Language Processing(NLP) lays a bridge between computers and natural languages by helping machines to analyze human language (Manning and Schütze, 1999) .Devlin et al. developed a model which is based on bidirectional encoder representation (Alyafeai et al., 2020). RoBERTa is a modified form of BERT (Liu et al., 2019).

4.1 BERT

Transformer-based models have been used since they use a self-attention mechanism and process the entire input data at once instead of as a sequence to capture long-term dependencies for obtaining contextual meaning. Bidirectional Encoder Representations from Transformers (BERT)(Devlin et al., 2018) tokenizes words into sub-words (using WordPiece) which are then given as input to the model. It also uses positional embeddings to replace recurrence.

4.2 RoBERTa

Robustly Optimized BERT-Pretraining Approach (RoBERTa) (Liu et al., 2019) is a state-of-the-art transformer model which improves BERT (Devlin et al., 2018) that uses a multi-headed attention mechanism which enables it to capture long term dependencies. It essentially fine-tunes the original BERT model along with data manipulation and uses Byte-Pair Encoding for utilizing the character and word level representations and removed Next Sentence Prediction (NSP) to match or even slightly improve downstream task performance.

4.3 Differential Privacy

Differential Privacy (DP) is a privacy standard which allows data use in any analysis by presenting mathematical guarantee (Dwork and Roth, 2014). It provides strong confidentiality in statistical databases and machine learning approaches through mathematical definition. This definition is an acceptable measure of privacy concern (Dwork, 2008).

Definition 1.1 : *M and E denote a random mechanism and each event (output) respectively. D and D' are defined neighboring datasets having difference with one record. (ε, δ) protects confidentiality (Dwork, 2011). M gives (ε, δ)-differential privacy for and D and D' if M satsifies:*

$$\Pr\left[M\left(D\right) \in E\right] \leq e^{\epsilon} \cdot \Pr\left[M\left(D'\right) \in E\right] + \delta \quad (1)$$

where ε denotes the privacy budget and δ represents the probability of error.

4.3.1 The Privacy Budget

The privacy guarantee level of M is controlled through privacy budget of ϵ (Haeberlen et al., 2011).There are two widely used privacy budget compositions as the sequential composition and the parallel composition.

The ratio between the two mechanisms ($M(D)$ and $M(D')$) limits by e^{ε}. For $\delta = 0$, M gives ε-differential privacy by its strictest definition. In other case, for some low probability cases, (ε,δ)-differential privacy provides latitude to invade strict ε-differential privacy. ε-differential privacy is called as pure differential privacy and (ε, δ)-differential privacy, where $\delta > 0$, is called as *approximate differential privacy* (Beimel et al., 2014). Differential privacy has two implementation settings: Centralized DP (CDP) via DP-SGD and Local DP (LDP) (Qu et al., 2021).

In CDP, a trusted data curator answers queries or releases differentially private models by using randomisation algorithms (Dwork and Roth, 2014). In this article, we use DP-SGD (Differentially Private Stochastic Gradient Descent) (Abadi et al., 2016) to train our models.

4.4 Federated Learning

As conventional centralized learning systems require that all training data produced on different devices be uploaded to a server or cloud for training, it may give rise to serious privacy concerns (Privacy, 2017). FL allows training an algorithm in a decentralized way (McMahan et al., 2017, 2016). It ensures multiple parties collectively train a machine learning model without exchanging the local data (Li et al., 2021). To define mathematically, it is assumed that there are N parties, and each party is showed with T_i, where $i \in [1, N]$. For the non-federated setting, each party uses its local data and depicted by D_i to train a local model M_i and send the local model parameters to the server. The predictive data is sent only the local model parameters to the FL server. Most centralized setups have just the IID assumption for train test data but in a federated learning based decentralized setup, non-IID poses the problem of high skewness of different devices due to different data distribution (Liu et al., 2021).

In federated language modeling, existing works (Yang et al., 2018) use FedAvg as the federated optimization algorithm. In FedAvg, gradients that are computed locally over a large population of clients are aggregated by the server to build a novel global model. Every client is trained by locally stored data and computes the average gradient with the current global model via one or more steps of SGD.

Applying FL to text classification can cause problems such as designing proper aggregating algorithms for handling the gradients or weights uploaded by different client models. Zhu et al. proposed a text classification using the standard FedAvg algorithm to update the model parameter with local trained models. Model compression has also been introduced to federated classification tasks due to the dilemma of computation restraints on the client-side, where an attempt to reduce the model size on the client-side to enable the real application of federated learning was made. For overcoming the communication dilemma of FL, central server can successfully train the central model with only one or a few rounds of communication under poor communication scenarios in a one-shot or few-shot setting.

5 Experimental Results

In the scope of the study, the FinBERT (pre-trained) model is used as the base model. Two NLP models were trained by implementing DP and FL. In this section, the results presented in the tables are discussed. The results placed in the tables are the average and the standard deviation of the results obtained after running the models thrice.

The dataset was split into train set and test set with 80:20 train test ratio. BERT and RoBERTa based models were used for the language modelling part. It should be noted that the table contain the average and the standard deviation of the results obtained after running the models 3 times. Table 1 shows a comparison according to epsilon values between both the language models using Centralized DP and in a Federated Learning set up. The Opacus library was used along with PyTorch for the experiments. We implement DP, FL and DP-FL on BERT and RoBERTa for $\epsilon = 0.5, 5, 15, 20, 25$. Our baseline model (with no noise) achieves an accuracy of **67.71%** and **68.37%** on BERT and RoBERTa respectively.

In baseline mode, we can see that RoBERTa has a slight improvement over BERT because of its robustness owing to a heavier pre-training procedure. We also notice that with the increase in epsilon values, the amount of standard deviation decreases as the model approaches towards its vanilla variant (without DP noise).

Table 2 also shows us the results obtained when DP was applied in a federated learning mode, both in IID (Identical and Independently distributed) and Non-IID data silos. For Non-IID scenarios, we assume 10 shards of size 240 assigned to each client. We run it over 10 clients in total, selecting only a fraction of 0.5 in each round for training. We add DP locally, that is, to each client model at every iteration and aggregate them to perform Federated Averaging. We observe the best accuracies with RoBERTa for the centralised DP implementation, particularly with $\epsilon = 25$ with an accuracy of 62.6%. BERT in a centralised DP setting does come close at $\epsilon = 25$ with an accuracy of 60.03%. The results also show that the accuracy decreases by adding FL to the DP implementations.

Table 1: Averaged Test Accuracies of FL and DPFL models

Setup	**Epsilon(ϵ)**	**BERT**	**RoBERTa**
Centralized DP	0.5	31.5±23.94	31.36 ± 26.35
	5	37.48±20.42	38.34 ±20.08
	15	51.71 ±14.71	51.34 ± 15.45
	20	55.37 ± 5.49	55.54 ± 5.54
	25	60.03 ± 1.37	62.6 ± 4.24
DP-FL IID	0.5	14.57 ± 2.86	20.11 ± 7.68
	5	30 ±25.6	30.04 ± 28.22
	15	40.34 ± 20.55	50.26 ± 20.84
	20	51.05 ± 7.95	54.78 ± 2.99
	25	53.47 ± 6.48	61.38 ± 0.93
DP-FL Non IID	0.5	19.82 ± 5.97	33.13 ± 25.41
	5	35.74 ± 21.48	36.51 ± 26.87
	15	45.87 ± 15.56	49.83 ± 20.6
	20	52.43 ± 4.08	53.36 ± 3.27
	25	58.96 ± 2.56	60.83 ± 0.53

We also empirically observe that with increase in ϵ , accuracy of the models also increases. This happens because as the value of ϵ increases, privacy decreases with the addition of noise from a smaller range which results in smaller variance. Consequently, the accuracy of the model increases. Inherently, applying DP to deep learning yields loss of utility due to the addition of noise and clipping. We can also observe that the performance of federated language models still lies behind that of centralized ones.

6 Conclusion

Financial data is highly sensitive , hence the risks of collecting and sharing data can limit studies. Financial organizations work with a lot of confidential user data and therefore highly value protecting the data to retain the integrity of the user and we need to delve into research of private training of machine learning models to ensure this. During this study, we benchmark the utility of privacy models while attempting to preserve the performance of SOTA transformer models such as BERT and RoBERTa. Our empirical results show that the models show better performance with increasing ε as expected with the decrease in noise. The models come close to the performance of the baseline models near the higher ε values.The DP + FL shows a similar trend which showcases a greater protection feature without compromising the performance. As future work, we hope to improve our models further by hyper-parameter tuning, freezing partial layers of the NLP model and implementing focal loss on the unbalanced dataset to better the results. The complete code to this paper can be found here: https://www.github.com/tiasa2/Privacy-enabled-Financial-Text-Classification-using-Differential-Privacy-and-Federated-Learning.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing.

A.I. Anton, J.B. Earp, Qingfeng He, W. Stufflebeam, D. Bolchini, and C. Jensen. 2004. Financial privacy policies and the need for standardization. *IEEE Security Privacy*, 2(2):36–45.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Priyam Basu, Tiasa Singha Roy, Rakshit Naidu, Zumrut Muftuoglu, Sahib Singh, and Fatemehsadat Mireshghallah. 2021. Benchmarking differential privacy and federated learning for bert models. *arXiv preprint arXiv:2106.13973*.

Amos Beimel, Kobbi Nissim, and Uri Stemmer. 2014. Private learning and sanitization: Pure vs. approximate differential privacy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork. 2008. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, pages 1–19, Berlin, Heidelberg. Springer Berlin Heidelberg.

Cynthia Dwork. 2011. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *Principles of Security and Trust*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 123–148, Germany. Springer-VDI-Verlag GmbH Co. KG.

Andreas Haeberlen, Benjamin C. Pierce, and Arjun Narayan. 2011. Differential privacy under fire. SEC'11, page 33, USA. USENIX Association.

Abhik Jana and Chris Biemann. 2021. An investigation towards differentially private sequence tagging in a federated framework. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pages 30–35.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and open problems in federated learning.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection.

Ming Liu, Stella Ho, Mengqi Wang, Longxiang Gao, Yuan Jin, and He Zhang. 2021. Federated learning meets natural language processing: A survey. *arXiv preprint arXiv:2107.12603*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *IJCAI*, pages 4513–4519.

Antoine Ly, Benno Uthayasooriyar, and Tingting Wang. 2020. A survey on natural language processing (nlp) and applications in insurance.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

H. B. McMahan, Eider Moore, D. Ramage, and B. A. Y. Arcas. 2016. Federated learning of deep networks using model averaging. *ArXiv*, abs/1602.05629.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T. Chitkushev, and Dimitar Trajanov. 2020. Evaluation of sentiment analysis in finance: From lexicons to transformers. *IEEE Access*, 8:131662–131682.

Kunjal Panchal. 2020. Differential privacy and natural language processing to generate contextually similar decoy messages in honey encryption scheme. *arXiv preprint arXiv:2010.15985*.

Apple Differential Privacy. 2017. Learning with privacy at scale.

Aman Priyanshu and Rakshit Naidu. 2021. Fedpandemic: A cross-device federated learning approach towards elementary prognosis of diseases during a pandemic.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Privacy-adaptive bert for natural language understanding.

Mattia Vicari and Mauro Gaspari. 2020. Analysis of news sentiments using natural language processing and deep learning. *Ai & Society*, pages 1–7.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 978–987.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated machine learning: Concept and applications.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*.

Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. 2019. Differentially private model publishing for deep learning. *2019 IEEE Symposium on Security and Privacy (SP)*.

Xinghua Zhu, Jianzong Wang, Zhenhou Hong, and Jing Xiao. 2020. Empirical studies of institutional federated learning for natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 625–634, Online. Association for Computational Linguistics.