Post Project Paper

Abstract

The autonomous anonymization of text has become an important task for natural language processing, as the amount of private data being stored has grown in the last decades and for institutions to comply with privacy regulations such as GDPR. Near-entity models (NER) is used for token classifications tasks to decide what entity a word belongs to. In our work, we propose a pipeline involving two BERT-based NER models. The first model is fine-tuned to be a normal NER model that annotates the text with entity labels, which are then passed to the second BERT-based NER model. The second model, fine-tuned only a dataset containing confidentiality tags, classifies individual words as confidential or not, based on the entity annotations generated by the first model. This resulted in an precision of 0.66, 0.19 and 0.17 for the non-entity, non-confidential and confidential words respectively. The results showed a marginal difference in replacing or including entities in the when classifying confidential words in a text.

1 Describe

This section introduces the questions addressed in the project.

1.1 Introduction

As the volume of data stored in the world increases, the need to be able to anonymize and store private data safely has become an important task for natural language processing. With the amount of data daily being processed and stored the manual anonymization of unstructured data is becoming infeasible.

With the help of NER models, words can be classified as different named entities. This project explores if these can be used to help identifying confidential words, and if the technique of incorporating NER models into the word classification task of deciding if a word or sequence should be labeled as confidential or not. To investigate this, two language models were fine-tuned as NER models with different tasks. The first task is to classify named-entities. The other task is to classify words as non-entities, confidential or non-confidential. By recognizing the entities from a text, the idea is that the giving these entities to the second language model would improve the accuracy of the prediction of which words should be labeled as confidential.

This was done in an effort to investigate if chaining language models together would result in a better accuracy than using just one but with a broader task. To this end, two datasets, one more focused on classic NER classification and one more focused on confidentiality, were used.

1.2 Method

We created a two-stage pipeline using BERT-based models to identify and classify tokens as confidential or non-confidential. In the first stage, a NER model is fine-tuned to detect entities in text. The second stage uses these entity annotations to determine whether each token is confidential. By comparing this two-stage setup to a baseline that omits entity annotations, we can evaluate the impact of NER on confidential-token classification.

1.2.1 Datasets

The first BERT-base-cased model was fine-tuned on permutations of the TAB and CoNLL datasets. Two datasets that are gold labeled. This tested whether more robust NER classification improves sensitivity to confidential information.

The CoNLL dataset uses Inside-Outside-Beginning (IOB) tags for five named-entity classes, while the TAB dataset has four additional tags and a different format (Pilán et al., 2022). Both datasets contain gold-standard labels. To unify them, the TAB dataset was converted to the IOB scheme, resulting in 17 total classes. In addition the TAB

dataset annotates confidentiality.

1.3 Models

1.3.1 Bert Base NER Entity Classification

The first stage pipeline is a NER model based on a pre-tuned BERT model (110M parameters) that identifies and annotates entities. Two versions were tested: one fine-tuned on TAB alone, and another using both TAB and CoNLL. We used standard hyperparameters (e.g., a learning rate of 2e-5, batch size of 16, 3 epochs). Model performance was evaluated at the end of each epoch, and the best of each model was saved.

1.3.2 Annotated input text

A number of strategies on how to incorporate the entity annotations from the first model into the input for the second model. The following four strategies were used: TOKEN, [ENTITY] TOKEN, [ENTITY] TOKEN [ENTITY], and [ENTITY]. Here, *TOKEN* represents the original text token; the other variants adds or replace it with the token's entity label. The TOKEN strategy served as our baseline.

1.3.3 Bert Base NER Confidential Classification

For the second model, the same BERT architecture was adapted to classify non-named-entities, named-entities, and confidential named-entities. Its input includes both text described in Section 1.3.2 and TAB's confidentiality labels. We applied hyperparameters similar to the first stage and saved the best checkpoint each epoch. By using the TOKEN strategy and comparing it to the annotated approach, we can assess the impact of NER on confidential-token classification.

1.4 Full Pipeline

After training the two BERT-based NER models, we combined them into a unified workflow shown in Figure 1.

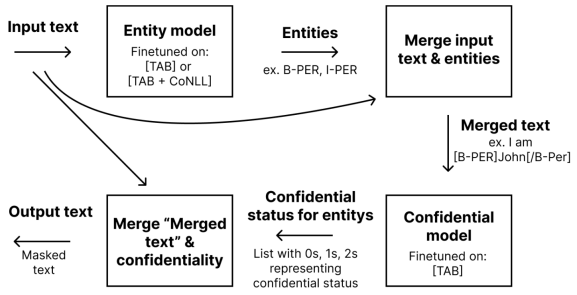

Figure 1: Overview of the Full Pipeline

1.5 Results and Analysis

This section the results for the NER and classification models, and the combination of them into one pipeline.

1.5.1 NER models

The two NER models trained for the entity classification task showed a difference between fine-tuning on CoNLL before training on TAB. The model trained only on TAB achieved an overall accuracy of 84% on the TAB validation set while the model fine-tuned on ConLL achieved an accuracy of 86%. One additional difference is that the fine-tuned model was able to capture underrepresented classes that the model only trained on TAB was unable to capture.

1.5.2 Confidential Classification

The model for classifying confidential tokens was fine-tuned with four different methods, first using only plain text as a baseline, second padding the token with entity in front, third padding around the token marking start and end of entity, and the finally exchanging the token for the entity.

Method	Input	Accuracy
1	Token	0.79
2	[Entity] Token	0.74
3	[Entity] Token [/Entity]	0.71
4	[Entity]	0.80

Table 1: Methods Overall Accuracy

The result seen in Table 1 indicates that the introduction of entities as wrappers for the words reduces the models capacity to identify confidential words. Replacing the word with its entity is on par with just the word.

1.6 Pipeline results

Combining the two models as displayed in Figure 1, achieved a precision of 0.66, 0.19 and 0.17 for the non-entity, non-confidential and confidential words respectively, in the TAB validation data set. This shows that the approach of using two models, and a pipeline like the proposed, to be an unsuccessful approach to the task of classifying confidential entities.

2 Examine

In this section, I critically reflect on my personal experience throughout the project, highlighting how we addressed dataset differences, navigated implementation hurdles, and interpreted our final results for confidentiality classification.

Differences in Datasets During development, we worked with two annotated English datasets:

- **CoNLL-2003** (Tjong Kim Sang and De Meulder, 2003): A widely used NER dataset containing four entity classes (PER, ORG, LOC, MISC) in IOB format.
- **Text Anonymization Benchmark (TAB)** (Pilán et al., 2022): Court case texts from the European Court of Human Rights. TAB includes additional entity types (e.g., QUANTITY) and uses direct token labeling rather than the IOB scheme.

From the outset, I underestimated how time-consuming it would be to convert the simpler TAB labeling to match the IOB approach in CoNLL. Because of this, certain classes were duplicated or merged into new labels. We ended up with 17 total classes once we included additional entity types, and some classes had very few training examples. Upon reconciling the annotation styles, we discovered subtle inconsistencies, such as the token "*of*" frequently receiving an `I-ORG` label regardless of context. This challenge reflects findings in the literature; for example, Ivanova et al. (2022) show that even slight differences in annotation guidelines can lead to significant performance discrepancies. Moreover, recent efforts such as the Universal NER benchmark (Mayhew et al., 2024) emphasize the importance of establishing a unified, cross-corpus schema, which reinforces the need for robust conversion processes in multidataset setups.

Implementation Challenges Another core lesson I learned was that even well-known datasets, such as CoNLL-2003, need careful inspection when integrated with other corpora. Although the Hugging Face libraries facilitated initial data loading and tokenization, reconciling the annotation schemes required custom Python scripts. Debugging these scripts was especially difficult because errors such as a mislabeled "*of*" would appear only in certain organizational names, making them easy to miss. I also realized the importance of clearly documenting the logic of every script so that each member of the team could interpret - and potentially fix - any unexpected issues in the combined data set.

Final Classification Results After completing the two-stage pipeline, where the first BERT-based model performed NER and the second predicted confidentiality labels, we evaluated the final system in three classes.

- **Label 0:** Non-Confidential, Non-Entity
- **Label 1:** Confidential, Non-Entity
- **Label 2:** Confidential, Entity

Table 2 shows the performance metrics in these classes. Although the overall accuracy was around 0.59, class imbalance and annotation inconsistencies significantly impacted precision and recall, especially for Labels 1 and 2.

Label	Precision	Recall	F1-score	Description
0	0.6552	0.8529	0.7384	Non-Conf., Non-Entity
1	0.2532	0.0167	0.0303	Conf., Non-Entity
2	0.1569	0.1288	0.1334	Conf., Entity
Accuracy = 0.5865				

Table 2: Final classification metrics for our confidential tagging system.

The above results make it evident that while Label 0 (Non-Confidential, Non-Entity) yielded decent precision and recall, the model struggled with the sparse confidential classes. This observation is consistent with the broader literature on class imbalance in NLP; as highlighted in Henning et al. (2023), class imbalance can lead to models that overly favor the majority class, thereby neglecting rarer, yet critical, categories. Furthermore, techniques such as sentence-level resampling (Wang and Wang, 2022) or modified loss functions like Dice loss (Li et al., 2020) have been proposed to address these issues. Although our current approach did not incorporate these advanced methods, their potential benefits are an important consideration for future improvements.

Personal Learning Reflections From a broader perspective, this project clearly demonstrated the deep interconnection between data quality, annotation consistency, and model design. Initially, I assumed a large pre-trained model like BERT would overcome small mislabelings or data discrepancies, but I discovered that such assumptions

were incorrect. Going forward, I plan to emphasize thorough dataset verification and consider alternative, parameter-efficient strategies (e.g., adapters or partial-freezing) to handle classification tasks with highly skewed labels. Additionally, exploring advanced techniques for class imbalance and data augmentation—as suggested by recent literature—could further improve model performance.

In summary, addressing these challenges has deepened my expertise in data preprocessing, label imbalance, and pipeline design, skills that I now consider essential for developing successful NLP systems.

3 Articulate

In this section, I summarize the most important lessons from the project, how I arrived at them, and why they matter.

Key Learnings Our work revealed that small annotation mistakes can overshadow the benefits of powerful models. For instance, over-tagging the token "*of*" as an `I-ORG` severely impacted precision and recall for confidential entities. Observing such inconsistencies reinforced how class imbalance and data-quality issues can limit performance even with sophisticated architectures.

Methods of Learning Through iterative experiments and reading on text anonymization and BERT-based classification, I saw firsthand how a multistage pipeline can magnify seemingly minor errors in data preparation. Frequent team discussions were important to quickly diagnose problems and align our approaches to converting the TAB and CoNLL-2003 data sets for consistent labeling.

Practical Relevance Ultimately, our best accuracy of around 58% highlighted the need for refined annotation schemes, class rebalancing, or advanced methods like CRF layers on top of BERT. While the results were not exceptional for confidential data, this experience sharpened my ability to evaluate scientific literature, identify pipeline bottlenecks, and propose targeted improvements for future NLP projects.

References

Sophie Henning, William Beluch, Alexander Fraser, and Annemarie Friedrich. 2023. A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–540, Dubrovnik, Croatia. Association for Computational Linguistics.

Rositsa V. Ivanova, Sabrina Kirrane, and Marieke van Erp. 2022. Comparing annotated datasets for named entity recognition in English literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3788–3797, Marseille, France. European Language Resources Association.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Xiaochen Wang and Yue Wang. 2022. Sentence-level resampling for named entity recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2151–2165, Seattle, United States. Association for Computational Linguistics.