DSBA6188 Final Project: Group 6 - Search Wizards

Name Entity Recognition for Banking Compliance and Risk

Team Members: Eric Phann (data, programming, and modeling). Jake Stallard (annotation, future considerations). Yaxin Zhao (annotation, research, model procedure). Kristen Zhang (annotation, most report writing). Sydney Kelly (annotation, future consideration).

1. Problem Statement

We decided to do a Name Entity Recognition (NER) in the provided dataset (eCFR Title 12) about banking compliance and materials. Our goal was to identify and classify the entities appearing in these materials, including the regulator, act, institution, and term. So NER would be an appropriate task. It's essential in text mining and information retrieval because a) banking compliance and risk materials always contain large amounts of textual data, which is very hard to read and manage. Doing NER on these materials can help to structure the data into a more manageable format; b) for future retrieval, accurately identifying entities in banking texts improves information retrieval capabilities. It enables more precise searching, categorization, and linking of related documents or entities.

Practically, with the NER done, the banks can efficiently recognize the key information including the regulators, act, and involved entities, and therefore have a better preparation to report compliance, conduct disclosures, and assess and manage the risk.

2. Data

Dataset Overview:

We used eCFR Title 12 dataset. The full name is the Code of Federal Regulations. It's a large dataset (16.4 MB). It's composed of the United States federal regulations covering banking, financial institutions, and related news. The information inside the dataset includes regulations and provisions about banking supervision, financial policy, monetary policy, consumer protection, etc. We obtained the dataset from eCFR, a national archive or dataset.

We didn't perform pre-processing/chunking for the dataset. The purpose of pre-processing is to reduce noise in the dataset, and standardize the text. And the purpose of chunking is to segment the text into phrases and mark boundaries of entities within the text. However, this dataset is not like a web-scraped-dataset on social media, which includes lots of noise. This dataset is already cleaned provided by eCFR, listed chapter by chapter. Therefore, we chose not to do the pre-processing. We only shuffled the dataset to get a variety of documents rather than annotating, for example, only chapter 1.

Annotated Data (100 manual annotations; 500 few-shot annotations):

We selected 100 examples to be labeled among four annotators. By developing the basic annotation guideline with ChatGPT 3.5 to illustrate the meaning of regulator, act, institution, and term, and then fulfilling the annotation guideline with examples that we came across in the process of annotation, four members annotated 100 manual shots for validation locally using Prodigy. The annotation guideline is available in the appendix. One copy of the annotation was chosen as the validation data. With the manually annotated data, the accuracy and reliability of the annotations can be assured, which is better for future training and evaluation. We also produced 500 “few-shot” annotations using the annotation guidelines and a few examples. This was done through integrating spaCy-llm with OpenAI API (ChatGPT 3.5 v3).

3. Model Procedure

Model Selection:

- few-shot-model

  Few-shot learning involves training models on a small dataset (few-shot) and then adapting them to perform tasks on new, unseen data with limited additional training examples. Banking compliance materials may contain specialized terminology and entities that are not well-represented in general language models. Few-shot learning allows for customizing the model to recognize these specific entities with minimal labeled examples. Additionally, few-shot learning facilitates rapid adaptation to new regulatory changes or emerging entities in the banking domain, making it suitable for dynamic environments.

  We used Chat-GPT 3.5 v3 to generate annotated data using “few-shot” examples and then developed a NER model using this data. SpaCy-LLM config file and few-shot examples used are available in the repo as well.

- manual-model

Manual model refers to an approach where domain experts manually annotate training data with entity labels and then train NER models using traditional supervised learning methods. In domains like banking compliance, where accuracy and precision are crucial, manual annotation ensures high-quality labeled data tailored to the specific nuances of the domain.

We used manual annotations developed by the team to develop this NER model.

- mixed-model

  A mixed model combines automated approaches (such as pre-trained language models) with manual intervention for fine-tuning or correcting predictions.

  Mixed models offer flexibility and scalability, allowing for iterative improvements through a combination of automated processing and human oversight, which is particularly valuable in dynamic and evolving domains like banking compliance.

  In this case, we used few-shot annotations as training data and manual annotations for validation data to see how performance would be among the two datasets.

Implementation Details:

- Shuffle/partition datasets and annotate as needed in Prodigy
- Use data-to-spacy recipe using relevant data to convert .jsonl to .spaCy and generate any other required or desired configs for the spaCy pipeline
- Use resulting corpus folders to develop, train, and evaluate models as needed

4. Evaluation

Quantitative Metrics: We used F1 scores to evaluate the models. F1 score is an appropriate metric which combines the two evaluation metrics: precision and recall, therefore considering both the accuracy of prediction and the ability to capture all relevant entities. In the three models that we tested, the few-shot model had the best F1 score of 0.59, the manual-model has the best F1 score of 0.13, and the mixed model has the best F1 score of 0.08.

Qualitative Analysis: We selected 5 examples (see appendix) not used for training for each of our models to annotate. There were no specific selection criteria other than examples being ideally meaningful and short (basically we chose the examples randomly). For the few-shot model, it worked better in recognizing the regulators and institutions, but lacked the recognition of acts and terms. For the manual model, it's labeling fewer entities (only 4 entities), but successfully identified one ACT entity. For the mixed model, it identified six entities (the most among the three models), successfully identified regulators and one act.

Experiment Results:

	F1 Score	Run Time	Qualitative Analysis
few-shot-model	0.59	4m22s[1]	Very lacking in labels. Worked better in recognizing regulators and institutions.
manual-model	0.13	4m36s[2]	Less labeling numbers. But labeled one act successfully.
mixed-model	0.08	4m29s[3]	Labeled six entities. Successfully identified regulators and one act.

5. Code and Reproducibility

Code Overview: Our code could be found on GitHub repository and Hugging Face. Our code (see appendix) was composed of five sections:

- Import Packages: We imported spacy and displacy.
- Import Datasets: The few-shot training data and manual validation data was imported and uploaded.
- Import spaCy configs and files.
- Training the three models.
- Re-evaluate the models with five examples.

Reproducibility: If others want to reproduce the results,

[1] Due to the large size of dataset, we only used 80 few-shot examples for training and 20 additional ones for validation.

[2] Due to the large size of dataset, we only used 80 manual examples for training and 20 additional ones for validation.

[3] Due to the large size of dataset, we only used 80 few-shot examples as training and 20 manual examples as validation.

DSBA6188 Final Project: Group 6 - Search Wizards

The following software is required: **prodigy** (for annotating), **vscode** (for local annotations and file saving), and **google colab** (for running the codes). **ngrok** is optional for group annotations.

The following libraries are required: spaCy

GPU: Colab's T4 GPU

---

6. Next Steps

Considering our dataset contains large text documents we were experiencing initial challenges with memory when trying to train our model as well as receiving overall low accuracy scores for our annotations. One way to start improving our model is to continue to meet as a group to discuss and refine our annotation guidelines even further based on our initial process. This will allow us to talk about different text formats or phrases that may seem to conflict with one or more guidelines, along with any errors that could occur. Following this, we can also think about an expansion of our dataset in order to continuously improve our model with new texts and data.

With more time, using a fine-tuning recipe in Prodigy in conjunction with SpaCy would help us take advantage of Prodigy’s active learning features. Essentially, this allowed us to review and fix annotations in the test set for which a low confidence score was given. This would help to make more efficient and meaningful improvements to our model. A worthwhile feature to take advantage of considering the length, content of, and size of our data.

Chunking would be an effective strategy for our group to separate and evaluate our data. This would allow us to essentially split our dataset into more manageable sections, which would make it much easier to improve and remove any inconsistencies that may occur leading to a more accurate model.

Another strategy is to improve our use and interpretation of the data privacy and security factors since we could potentially be dealing with a lot of sensitive information as well as personal information. Creating more secure access controls would help ensure that only authorized individuals have access to the data. Additionally, implementing encryption measures for data storage and transmission can safeguard against unauthorized access. Regular audits and reviews of security protocols can also help identify and address any vulnerabilities in the system, thereby enhancing overall data protection and compliance with privacy regulations.

DSBA6188 Final Project: Group 6 - Search Wizards

Appendix

A. Code Listings

A.1. Import packages

```
from google.colab import files
import spacy
from spacy import displacy
```

A.2. Import datasets

```
uploaded = files.upload()
```

Then chose the two files (ecfr-few-shot.jsonl, ecfr-manual.jsonl)

A.3. Import spaCy configs and files

```
%%capture
!python -m spacy download en_core_web_sm
# upload config.cfg, dev.spacy, train.spacy
uploaded = files.upload()
# upload ner.json (/labels contents)
uploaded = files.upload()
# upload config.cfg, dev.spacy, train.spacy
uploaded = files.upload()
# upload ner.json (/labels contents)
uploaded = files.upload()
# upload config.cfg, dev.spacy, train.spacy
uploaded = files.upload()
# upload ner.json (/labels contents)
uploaded = files.upload()
```

A.4. Training the model

```
gpu = spacy.prefer_gpu()
print(gpu)
```

```
!python -m spacy train ./few-shot-corpus/config.cfg --paths.train
./few-shot-corpus/train.spacy --paths.dev ./few-shot-corpus/dev.spacy --gpu-id 0
```

```
--output ./few-shot-model
!zip -r ./few-shot-model.zip ./few-shot-model
from google.colab import files
files.download("./few-shot-model.zip")
!python -m spacy train ./manual-corpus/config.cfg --paths.train
./manual-corpus/train.spacy --paths.dev ./manual-corpus/dev.spacy --gpu-id 0
--output ./manual-model
!zip -r ./manual-model.zip ./manual-model
from google.colab import files
files.download("./manual-model.zip")
!python -m spacy train ./mixed-corpus/config.cfg --paths.train
./mixed-corpus/train.spacy --paths.dev ./mixed-corpus/dev.spacy --gpu-id 0
--output ./mixed-model
!zip -r ./mixed-model.zip ./mixed-model
from google.colab import files
files.download("./mixed-model.zip")
```

A.5. Evaluating the models

```
nlp = spacy.load("./few-shot-model/model-best")
doc = nlp("Notwithstanding any other provision of this title, the NCUA may, without
any administrative due process, immediately place into conservatorship or
liquidation any corporate credit union that has been categorized as critically
undercapitalized.")
displacy.render(doc, style="ent")
```

Similar process to other models and other examples.

B. Inference examples

B.1. For the few-shot model, it worked better to recognize regulator and institution, but neglected the acts/terms. It could be due to the imbalance in labels. And overall, it’s very lacking in the number of labels overall, with example 3 not even having any labels. The examples are as below:

- Notwithstanding any other provision of this title, the NCUA **REGULATOR** may, without any administrative due process, immediately place into conservatorship or liquidation any corporate credit union that has been categorized as critically undercapitalized.
- Loans made under title III by banks for cooperatives and agricultural credit banks **REGULATOR** may be made to eligible domestic parties domiciled within any territory that may be served by Farm Credit institutions under section 1.2 of the Act and to eligible foreign parties without regard to domicile.
- If an interlocutory appeal or collateral attack is brought in any court concerning all or any part of an adjudicatory proceeding, the challenged adjudicatory proceeding shall continue without regard to the pendency of that court proceeding. No default or other failure to act as directed

in the adjudicatory proceeding within the times prescribed in this subpart shall be excused based on the pendency before any court of any interlocutory appeal or collateral attack.
- If the Board of Directors **REGULATOR** finds that a savings association is a special supervisory association under the provisions of section 8(a)(8)(B) of the FDIA (12 U.S.C. 1818(a)(8)(B)) for purposes of temporary suspension of insured status, the Board of Directors **REGULATOR** shall serve upon the association its findings with regard to the determination that the capital of the association, as computed using applicable accounting standards, has suffered a material decline;
- The conservator or receiver may enforce any contract entered into by the regulated entity pursuant to the **INSTITUTION** provisions and subject to the restrictions of section 1367(d)(13) of the Safety and Soundness Act.

B.2. For the manual model, three of five examples weren’t successfully labeled. But this time the model successfully labeled one ACT entity.

- Notwithstanding any other provision of this title, the NCUA may, without any administrative due process, immediately place into conservatorship or liquidation any corporate credit union that has been categorized as critically undercapitalized.
- Loans made under title III by banks for cooperatives and agricultural credit banks may be made to eligible domestic parties domiciled within any territory that may be served by Farm Credit institutions **REGULATOR** under section 1.2 of the Act and to eligible foreign parties without regard to domicile.
- If an interlocutory appeal or collateral attack is brought in any court concerning all or any part of an adjudicatory proceeding, the challenged adjudicatory proceeding shall continue without regard to the pendency of that court proceeding. No default or other failure to act as directed in the adjudicatory proceeding within the times prescribed in this subpart shall be excused based on the pendency before any court of any interlocutory appeal or collateral attack.
- If the Board of Directors finds that a savings association is a special supervisory association under the provisions of section 8(a)(8)(B) of the FDIA (12 U.S.C. 1818(a)(8)(B)) **ACT** for purposes of temporary suspension of insured status, the Board of Directors **REGULATOR** shall serve upon the association its findings with regard to the determination that the capital of the association, as computed using applicable accounting standards, has suffered a material decline; that such association or its directors or officers, is engaging in an unsafe or unsound practice in conducting the business of the association; that such association is in an unsafe or unsound condition to continue operating as an insured association; or that such association or its directors or officers, has violated any law, rule, regulation, order, condition imposed in writing by any Federal banking agency **REGULATOR** , or any written agreement, or that the association failed to enter into a capital improvement plan acceptable to the Corporation prior to January, 1990.
- The conservator or receiver may enforce any contract entered into by the regulated entity pursuant to the provisions and subject to the restrictions of section 1367(d)(13) of the Safety and Soundness Act.

B.3. For the mixed model, the labeling entities were the most among the three. But three of the five examples were not identified with any entity either.

- Notwithstanding any other provision of this title, the NCUA **REGULATOR** may, without any administrative due process, immediately place into conservatorship or liquidation any corporate credit union that has been categorized as critically undercapitalized.
- Loans made under title III by banks for cooperatives and agricultural credit banks may be made to eligible domestic parties domiciled within any territory that may be served by Farm Credit institutions under section 1.2 of the Act and to eligible foreign parties without regard to domicile.
- If an interlocutory appeal or collateral attack is brought in any court concerning all or any part of an adjudicatory proceeding, the challenged adjudicatory proceeding shall continue without regard to the pendency of that court proceeding. No default or other failure to act as directed in the adjudicatory proceeding within the times prescribed in this subpart shall be excused based on the pendency before any court of any interlocutory appeal or collateral attack
- If the Board of Directors **REGULATOR** finds that a savings association is a special supervisory association under the provisions of section 8(a)(8)(B) of the FDIA (12 U.S.C. 1818(a)(8)(B)) for purposes of temporary suspension of insured status, the Board of Directors **REGULATOR** shall serve upon the association its findings with regard to the determination that the capital of the association, as computed using applicable accounting standards, has suffered a material decline; that such association or its directors or officers, is engaging in an unsafe or unsound practice in conducting the business of the association; that such association is in an unsafe or unsound condition to continue operating as an insured association; or that such association or its directors or officers, has violated any law, rule, regulation, order, condition imposed in writing by any Federal banking agency **REGULATOR** , or any written agreement **ACT** , or that the association failed to enter into a capital improvement plan acceptable to the Corporation **REGULATOR** prior to January, 1990.
- The conservator or receiver may enforce any contract entered into by the regulated entity pursuant to the provisions and subject to the restrictions of section 1367(d)(13) of the Safety and Soundness Act.

C. Annotation Guidelines

Annotation Guidelines

1. **Regulatory Bodies**:
   - Organizations or agencies tasked with overseeing and enforcing regulations within a specific industry or sector, to ensure compliance with banking laws and regulations, maintaining financial stability.
   - Common Examples: Federal Reserve Board, OCC, FDIC (Federal Deposit Insurance Corporation), CFPB, and SEC
   - Unfamiliar examples: Federal Housing Finance Agency (FHFA), Federal Emergency Management Agency (FEMA), and National Credit Union Administration (NCUA).
2. **Regulatory Acts or Laws**:

- Legislative measures enacted by government bodies to regulate and govern specific aspects of an industry or sector, to establish rules and requirements for financial institutions.
- Common examples: The Dodd-Frank Act, Bank Secrecy Act (BSA), Gramm-Leach-Bliley Act (GLBA), Truth in Lending Act (TILA), and Community Reinvestment Act (CRA)
- Unfamiliar examples:
    - i. § 614.4930: The symbol "§" typically signifies "section," and the subsequent numbers and text likely denote a specific section or provision within a particular act.
    - ii. 12 CFR part 262: Those who only included the index of the act.
    - iii. Some are called “principles” or “standards”, like “U.S. Generally Accepted Accounting Principles” and “the International Financial Reporting Standards”. We still consider these acts as they are all used to regulate the market.

3. **Financial Institutions**:
    - Entities that provide financial services, such as banking, lending, investing, and insurance, to individuals, businesses, and governments.
    - Common examples: It could include common titles like bank and credit union.
    - Unfamiliar examples:
        - i. Federal Reserve Bank: It’s both a regulator and an institution. But in this annotation, we only considered it as a regulator as it’s giving more attention to regulation and supervision.
        - ii. the Corporation: In the text, the “corporation” started with an upper case “C”, it could be referred to some specific companies so we included it here.
        - iii. Sometimes the institution is called a “corporation” but it’s not an institution but a regulator: e.g., Federal Deposit Insurance Corporation (FDIC). More background knowledge and research should be done when deciding on these.
4. **Financial Terms**:
    - The terminology and concepts related to finance, banking, and investment activities.
    - Common examples: Interest rate, disclosure, asset management/arrangement, security interest, equity interest.
    - Unfamiliar examples:
        - i. Long terms which should include every word: Limited partnership interest

ii. FSI, QFC: Those look like organizations, but they are actually terms like “Financial Stability Index” and “Qualified Financial Contract”.