# Few-shot and Zero-shot Approaches to Legal Text Classification: A Case Study in the Financial Sector

**Rajdeep Sarkar**[1], **Atul Kr. Ojha**[1], **Jay Megaro**[2],
**John Mariano** [2], **Vall Herard**[2] and **John P. McCrae**[1]

[1] Data Science Institute, National University of Ireland Galway, Ireland
[2] FMR LLC, Boston, USA

{rajdeep.sarkar, atulkumar.ojha, john.mccrae}@insight-centre.org
{jay.megaro, john.mariano, vall.herard}@fmr.com

## Abstract

The application of predictive coding techniques to legal texts has the potential to greatly reduce the cost of legal review of documents, however, there is such a wide array of legal tasks and continuously evolving legislation that it is hard to construct sufficient training data to cover all cases. In this paper, we investigate few-shot and zero-shot approaches that require substantially less training data and introduce a triplet architecture, which for promissory statements produces performance close to that of a supervised system. This method allows predictive coding methods to be rapidly developed for new regulations and markets.

## 1 Introduction

Organizations that are governed by legal and regulatory statues concerning communications with the public are required to comply with principles-based content standards. As such, this involves a significant expense due to having to use highly qualified staff to review, iterate on communications internally and file content with regulators, externally. There is thus a substantial expense in terms of iteration time and specialized staff associated with this process. With recent advances in Natural Language Processing (NLP) technologies, it is increasingly becoming possible to automatically flag high-risk statements by *predictive coding* and thus reduce the cost of these manual reviews. However, each industry has specific regulatory requirements and modern NLP systems need large training sets to be effective, and as such it is challenging to develop such systems. In this paper, we focus on a single example of such a regulatory compliance in the financial domain under the US regulation FINRA 2210[1], which states that "no member may make any false, exaggerated, unwarranted, promissory or misleading statement or claim in any communication." We examine how we can train a system in the following settings: firstly in a traditional data-heavy supervised setting, where a large number of existing examples have been classified. Secondly, we investigate a zero-shot training situation, where we have asked a legal expert to provide only rough guidelines for what is not compliant with the legal code. Finally, we combine this in a few-shot setting and show that with comparatively little training data, we can achieve performance that is equivalent with the data-heavy supervised setting and thus enables text classification systems for regulatory compliance to be constructed quickly and with little effort allowing them to cover a wide range of industries and national regulatory frameworks.

## 2 Related Work

There has been some work in the area of legal text classification and the application of text classification techniques to legal texts has mostly been successful so far. Methods based on counting the words in the text and then classifying using machine learning approaches such as support vector machines (Cortes and Vapnik, 1995) for example by Sulea et al. (2017), where they applied this method to the classification of texts according to the legal area, ruling and time span of the text. Deep learning methods such as Convolutional Neural Networks (CNNs) have been shown to further improve the performance of such systems (Wei et al., 2018). More recently, the emergence of large pre-trained language models such as BERT (Devlin et al., 2019) has further increased the performance and Shaheen et al. (2020) showed that these models could be used to classify legal texts according to thousands of labels and even on multiple languages if sufficient training data exists.

A criticism of such NLP-based approaches to predictive coding, especially with the emergence of more sophisticated deep learning methods, is

[1] https://www.finra.org/rules-guidance/rulebooks/finra-rules/2210



*Natural Legal Language Processing Workshop 2021*, pages 102–106
November 10, 2021. ©2021 Association for Computational Linguistics

that they can appear to be 'black boxes', and thus there has been work in providing explainable systems (Mahoney et al., 2019) that can identify snippets and provides explanations for why they make certain predictions. Similarly, some work has gone into the investigation of specific complexities of legal texts, such as in Nallapati and Manning (2008), who showed that for some legal texts the complex combination of negative and positive statements can confused machine learning approaches. They showed that by combining these machine learning approaches with propositional logic, text classification systems could handle intricate legal wording.

## 3 Methodology

To solve the problem of legal text classification, we approach this with a triplet architecture (Wei et al., 2021) where an input sentence, $s$, is compared with a positive example $s_+$ and a negative example $s_-$ as depicted in Figure 1. We begin by describing the model architecture. Then we discuss the triplet loss used for training the network. Finally, we describe the classification model used for the final classification.

### 3.1 Model Architecture

Most existing methods of text classification only consider the local features of the samples, and their experimental results show better performance than traditional non-deep learning methods. However, in these methods, the global features of the sample are usually ignored, and these ignored global features will affect the classification accuracy. These global features are key to the use-case presented. To solve this problem, a triplet capsule network framework is proposed for text classification, to optimize results.

A triplet network consist of three instances of the same neural network with shared parameters. The network takes as input three examples in each sample. The three samples consists of the anchor, positive and negative examples. The anchor and positive examples belong to the same class, while the negative example belongs to a different class. The network outputs two values, the distance between the anchor and the positive example and the distance between the anchor and the negative example.

We design a triplet network for the sentence classification task. The network encodes each incoming sentence using Sentence-BERT (Reimers and Gurevych, 2019) encoder. Sentence-BERT captures the contextual information in a sentence in a fixed-size vector representation. The contextual sentence representation is then fed to a two-layer perceptron. The hidden layer of the perceptron has ReLU (Nair and Hinton, 2010) activation for introducing non-linearity in the perceptron.

$$\mathbf{e}_{\mathrm{s}}^1 = \text{S-BERT(s)} \tag{1}$$

$$\mathbf{e}_{\mathrm{s}}^2 = RELU(W_{\theta,1}\mathbf{e}_{\mathrm{s}}^1) \tag{2}$$

$$\mathbf{e}_{\mathrm{s}}^3 = W_{\theta,2}\mathbf{e}_{\mathrm{s}}^2 \tag{3}$$

where $W_{\theta,1} \in R^{d_{e^2} \times d_{e^1 h}}$ and $W_{\theta,2} \in R^{d_{e^3} \times d_{e^2}}$ and the parameter matrices to be learned during training. The Sentence-BERT model is also fine-tuned during the training procedure.

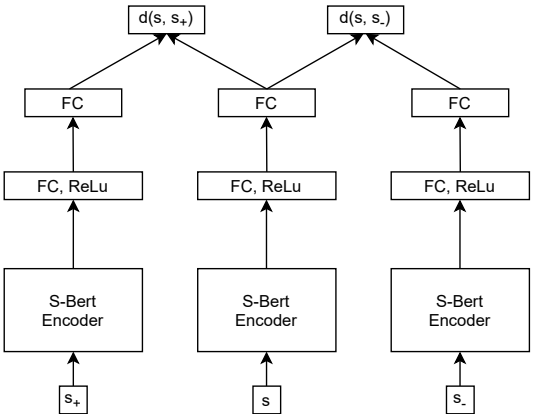


Figure 1: Overall architecture of our approach.

### 3.2 Triplet Loss

Triplet loss (Hoffer and Ailon, 2015) has been used in few-shot classification methods. Although introduced for images, it has been successfully adapted in natural language processing (Wei et al., 2021; Lauriola and Moschitti, 2021). Triplet loss enables the network to distinguish been positive and negative examples of a class. It is defined in Equation 4.

$$\mathcal{L}(D_+, D_-) = \|D_+, D_- - 1\|_2^2 \tag{4}$$

where $D_+$ and $D_-$ are defined in Equation 7 and 8 respectively.

$$d_+(s, s_+) = \left\| \mathbf{e}_{\mathrm{s}}^3 - \mathbf{e}_{\mathrm{s}_+}^3 \right\|_2 \tag{5}$$

$$d_-(s, s_-) = \left\| \mathbf{e}_{\mathrm{s}}^3 - \mathbf{e}_{\mathrm{s}_-}^3 \right\|_2 \tag{6}$$

$$D_+ = \frac{e^{d_+(s,s_+)}}{e^{d_+(s,s_+)} + e^{d_-(s,s_-)}} \tag{7}$$

$$D_- = \frac{e^{d_-(s,s_-)}}{e^{d_+(s,s_+)} + e^{d_-(s,s_-)}} \tag{8}$$

Table 1: Performance of our few-shot learning model in comparison with other supervised and zero-shot learning methods.

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| Naive Bayes | 0.78 | 0.48 | 0.60 | 0.75 |
| MLP | 0.66 | 0.70 | 0.68 | 0.75 |
| SVM | 0.76 | 0.67 | 0.71 | 0.79 |
| S-BERT | 0.72 | 0.69 | 0.70 | 0.78 |
| Laser | 0.75 | 0.68 | 0.71 | 0.79 |
| Zero-Shot | 0.48 | 0.75 | 0.59 | 0.60 |
| Few-Shot(ours) | 0.61 | 0.74 | 0.67 | 0.72 |

where $\|\cdot\|_2$ denotes the $l_2$ norm. The embeddings $\mathbf{e}_s$, $\mathbf{e}_{s_+}$ and $\mathbf{e}_{s_-}$ denote the representation of the anchor, positive and negative sentences from Equation 3 and $e$ is Euler's number. The loss objective ensures that when $\frac{d_+(s,s_+)}{d_-(s,s_-)} \rightarrow 0$, then $\mathcal{L} \rightarrow 0$. We minimize $\mathcal{L}$ to learn the parameters of our model.

### 3.3 Classification

The network learns sentence representations where examples of the same class are close together. The closeness of two sentences is measured by calculating the euclidean distance between their representations from Equation 3. For the final classification, we use a Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel. We use SVM for classification as it learns by minimizing the hinge loss which is similar to the loss used for training the triplet network. Given the sentence representation from Equation 3, the SVM outputs a probability $p$ of a sentence being a promissory sentence. The sentence is classified being promissory using Equation 9.

$$\text{class}(s) = \begin{cases} \text{promissory} & p \geq \alpha \\ \text{not promissory} & p < \alpha \end{cases} \quad (9)$$

where $\alpha$ is a hyperparameter to be set.

## 4 Experimental Setup

### 4.1 Dataset

We retrieved data from internal and external data sources in the financial services industry to create the initial data sets for the approach. After data setup, we cleaned the data to remove duplicate and irrelevant content to ensure data quality before review. Each data point was reviewed and labelled by both in-house licensed staff and contractors to confirm the interpretation of regulatory content standards.

We split the dataset into training, development and test set. The training set contains 2,016 promissory sentences and 3,260 non-promissory sentence. The test set contains 860 and 1,402 promissory and non-promissory examples, respectively. For our few-shot learning model, we sample 40 promissory and 190 non-promissory examples sentences from the training set and learn our model on this subset.

### 4.2 Baselines and Evaluation Metrics

We compare the performance of our approach with the following supervised learning methods.

- Naive Bayes: We learn a Naive Bayes classification model using TF-IDF scores of the tokens in the sentence.
- Multi-Layer Perceptron (MLP): We learn a two-layer perceptron with ReLu activation in the hidden layer using the TF-IDF scores of the sentence tokens as input features to the model.
- SVM: Similar to the MLP model, we learn an SVM model for the classification task. We set the regularization parameter $C$ and $gamma$ to 1.0 and 0.1 respectively.
- Sentence-BERT (Reimers and Gurevych, 2019): This setting is similar to our proposed approach. We encode each sentence into a fixed-sized vector using its Sentence-BERT embedding. The sentence embedding is then fed into a 3 layer fully connected neural network with ReLu activation in the first two layers. The model is learned by minimizing the cross-entropy loss of classification using the Adam optimizer.
- Laser (Artetxe and Schwenk, 2019): In this setting, we encode each sentence using its Laser embeddings. The remaining architecture remains the same as that used in the Sentence-BERT model.

In addition to the supervised approaches, we compare our few-shot learning approach against a zero-shot learning approach. Yin et al. (2019) suggested method for using pre-trained natural language inference models as sequence classifiers. Towards this end, we use BART model (Lewis et al., 2020) as our zero-shot learning model. We consider the sentences tagged as 'promissory' as the hypothesis. The probability of a sentence being the premise

| | Sentence | Model Result | Gold Label |
|---|---|---|---|
| 1 | Stocks are an income source which main street is ignoring | non-promissory | promissory |
| 2 | It is going up in all currencies | non-promissory | promissory |
| 3 | Joe Smith picks the best stock in each sector for the fund | non-promissory | promissory |
| 4 | All rights reserved. | promissory | non-promissory |
| 5 | Save more now. | promissory | non-promissory |
| 6 | There is no action required on your part. | promissory | non-promissory |

Table 2: Error Analysis: Examples where our few-shot model produces classification labels different from the gold labels.

for these tagged sentence is calculated using the BART model. We then consider the maximum of those scores, and if the maximum score is greater than 0.7, we classify the sentence as a promissory sentence.

### 4.3 Implementation Details

For the task, we use the Sentence-BERT base model. It encodes a sentence into a fixed-size vector of length 768. We set $d_{e^1}$, $d_{e^2}$ and $d_{e^3}$ to 768, 300 and 10 respectively. For every positive sentence belonging to the promissory class, we sample three sentences from the non-promissory class as negative sentences. We use grid-search on the development set to set the values of hyperparameters. The batch size is set to 16 for the triplet network and the model trained using Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-5 for 10 epochs. We set the cost parameter $C$ of SVM to 0.03 and $\alpha$ in Equation 9 to 0.005.

## 5 Results

In section, we first perform a quantitative analysis of models. We then study a few examples where our approach produces results different from the gold standard dataset.

### 5.1 Quantitative Analysis

In this work, we propose a few-shot learning method for legal text classification. Table 1 shows the performance of different model. Even when training with a limited number of examples, the few-shot learning model achieves better recall performance as compared to different supervised models. We find that the precision of our model is better than the zero-shot learning model but lower than the supervised models. Overall the F-Measure shows that similar results can be obtained with a few-shot approach and this enables the goal of rapid training of systems for different legal tasks. In the situation where the classifier is applied as a first filter, a high recall is preferable as we would rather create more work for a second manual annotation than miss some important texts.

### 5.2 Qualitative Analysis

In Table 2, we see some examples of misclassifications made by our algorithm. It is obvious that this is a very challenging task, with subtle changes in meaning being important for the classification. Examples 3 and 6 both appear to make factual statements, however, Example 3 is classed as promissory due to the context that 'Joe Smith' is likely an agent of the company. Similarly, Example 2 is difficult to classify without context and this shows that the introduction of further context is most likely to improve the effectiveness of the approach.

## 6 Conclusion

We have investigated the use of few-shot and zero-shot text classification methods for the quick development of predictive coding systems for legal texts. We found that zero-shot systems have a substantial decrease in performance relative to a supervised approach. We then developed a few-shot approach based on a triplet architecture and showed that this model is within a few percentage points of the supervised system in performance but requires much less manual annotation in order to develop the system.

## Acknowledgements

This work has been funded by FMR LLC. Researchers at the Data Science Institute are supported by Science Foundation Ireland as part of Grant Number SFI/12/RC/2289_P2, Insight SFI Centre for Data Analytics.

## References


Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

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.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition - Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015, Proceedings*, volume 9370 of *Lecture Notes in Computer Science*, pages 84–92. Springer.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ivano Lauriola and Alessandro Moschitti. 2021. Answer sentence selection using local and global context in transformer models. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 298–312. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Christian J. Mahoney, Jianping Zhang, Nathaniel Huber-Fliflet, Peter Gronvall, and Haozhen Zhao. 2019. A framework for explainable text classification in legal document review. In *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*, pages 1858–1867. IEEE.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 807–814. Omnipress.

Ramesh Nallapati and Christopher D. Manning. 2008. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 438–446, Honolulu, Hawaii. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. 2020. Large scale legal text classification using transformer models. *CoRR*, abs/2010.12871.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017. Exploring the use of text classification in the legal domain. In *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017), London, UK, June 16, 2017*, volume 2143 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fusheng Wei, Han Qin, Shi Ye, and Haozhen Zhao. 2018. Empirical study of deep learning for text classification in legal document review. In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 3317–3320. IEEE.

Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5493–5500. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. Association for Computational Linguistics.