Proceedings of CSR-NLP I @LREC 2022, pages 52–57
Marseille, 25 June 2022

Towards Classification of Legal Pharmaceutical Text using GAN-BERT

Tapan Auti[1], Rajdeep Sarkar[1], Bernardo Stearns[1], Atul Kr. Ojha[1], Arindam Paul[2], Michaela Comerford[2], Jay Megaro[2], John Mariano[2], Vall Herard[2], John P. McCrae[1]

[1] Data Science Institute, National University of Ireland Galway, Ireland, [2] FMR LLC, Boston, USA
{tapan.auti, rajdeep.sarkar, bernardo.stearns, atulkumar.ojha, john.mccrae}@insight-centre.org
{arindam.paul, michaela.comerford, jay.megaro, john.mariano, vall.herard}@fmr.com

Abstract

Pharmaceutical text classification is an important area of research for commercial and research institutions working in the pharmaceutical domain. Addressing this task is challenging due to the need of expert verified labelled data which can be expensive and time consuming to obtain. Towards this end, we leverage predictive coding methods for the task as they have been shown to generalise well for sentence classification. Specifically, we utilise GAN-BERT architecture to classify pharmaceutical texts. To capture the domain specificity, we propose to utilise the BioBERT model as our BERT model in the GAN-BERT framework. We conduct extensive evaluation to show the efficacy of our approach over baselines on multiple metrics.

Keywords: Generative Adversarial Network, Text classification, BERT

1. Introduction

Occurrence frequency of misleading statements has increased in industries such as financial, legal, social media, health, biomedical and pharmaceutical. Meanwhile natural language processing has grown rapidly in particular due to machine learning and deep learning methods that provide mechanisms to classify text automatically. Various innovative approaches have been proposed. Devlin et al. (2019) proposed a BERT (Devlin et al., 2019) based classifier for the classification problem. Jofche et al. (2021) suggested the use of transfer learning for knowledge extraction from the classified pharmaceutical text while Croce et al. (2020) proposed a Generative Adversarial Network (GAN) based model leveraging the BERT architecture for the text classification. Sarkar et al. (2021) investigated Naive-Bayes, SVM, Multi-Layer Perceptron (MLP), Sentence-BERT(S-BERT), Laser, Zero-Shot and Few-Shot approaches to legal-financial text classification.

In the United States, the Food and Drug Administration (FDA) is responsible for protecting public health by ensuring the safety, efficacy of drugs. Promotional communications must meet the following criteria, which are based upon the FDCA: be clear, accurate and truthful, not be misleading, promote only cleared or approved intended use, be supported by valid scientific evidence, and include a fair balance between benefits and risks. In this paper, we focus on classification in the pharmaceutical industry for detecting misleading claims. We have adopted and extended a triplet network classifier (Sarkar et al., 2021) and GAN-BERT. Both models are relatively unexplored in the pharmaceutical domain. Firstly, we train a system on Naive-Bayes, SVM, MLP, S-BERT (Reimers and Gurevych, 2019) and Laser (Artetxe and Schwenk, 2019). Secondly, we train a system on GAN-BERT. Finally, to factor the domain specificity, we replace the BERT of GAN-BERT architecture with BioBERT (Lee et al., 2020), which is a BERT architecture fine-tuned on biomedical text. The semi-supervised approach of GAN-BioBERT with GAN's generative nature and the results of Bio-BERT on biomedical datasets motivated us to use them to optimise our results.

2. Related Work

Application of innovative techniques such as GAN-BERT, BERT, few-shot and zero-shot learning in the biomedical domain has been successfully explored. However, the extraction classification of biomedical and pharmaceutical text is difficult due to the domain specificity of terms and the inter-dependency of such terms with other tokens in the text. In most cases, it involves training models on large volumes of labelled data that can be expensive and time-consuming. Towards this end, Flores et al. (2019) posited FREGEX to extract biomedical features using regular expression. The authors used string based algorithms for extracting tokens having similar patterns and contextual features. Similarly, Flores et al. (2020) proposed CREGEX, an innovative method for automatically generating informative and discriminative regular expression.

With the advent of deep-learning based models, Yao et al. (2019) proposed a knowledge guided convolutional neural network model utilising a rule-based feature extractor for clinical text classification. Du et al. (2019) suggested the use of a label prediction network for biomedical text classification. On the other hand, Wu et al. (2021) proposed Bio-IE, a novel method utilising a hybrid neural network for extracting relations from biomedical text. They used multi-head enhanced convolutional graph to capture the complex relations and context information resisting noise. Luo (2017)

proposed a LSTM model for learning word and contextual embeddings without the need of manual feature engineering.
Vaswani et al. (2017) posited transformer architecture for improving the representational capacity of LSTMs. Devlin et al. (2019) utilised the transformer architecture and proposed BERT to obtain contextualised representation of sentences as well as tokens in sentences. They showed the efficacy of BERT on various natural language processing tasks. Bio-BERT (Lee et al., 2020) is fine-tuned BERT which is pre-trained with bio-medical dataset for capturing the dependencies between domain specific terms.
Even though GAN-BERT gave good results on classification tasks, BERT's capacity to handle bio-medical data wasn't that great as it cannot extract those features. Due to this GAN-BERT failed on bio-medical dataset.

3. Methodology

In this section, we begin with a formal definition of the GAN architecture. We then outline the semi-supervised training of the GAN for legal pharmaceutical text classification. We begin by describing the semi-supervised training of GAN for text classification and then focus on the details of the Model Architecture used in this work.

3.1. Generative Adversarial Network

We leverage the GAN architecture for the classification task. The GAN architecture consists of two networks interacting with each other, a discriminator and a generator. The generator constructs 'fake' examples to deceive the discriminator during the classification task, while the discriminator is trained to distinguish the generated samples from the real samples present in the dataset. The generator and the discriminator are trained together in an adversarial setting. In this work, the GAN network is trained using the Minimax loss (Goodfellow et al., 2014) as outlined in Equation 1.

$$\mathcal{L} = \underset{G}{min}\ \underset{D}{max}\ (\mathbb{E}_{x\sim\mathbb{P}_{data}}[log(D(x))] + \mathbb{E}_{z\sim\mathbb{P}_z}[1 - log(D(G(z)))]) \quad (1)$$

where *D* and *G* are the discriminator and the generator network to be learned and z is the noise induced in the model to generate artificial samples using the generator.

3.2. Semi Supervised Training of Generative Adversarial Network

The goal of the classification task is to classify a sentence into one of *K* classes. Following Croce et al. (2020), we train the GAN in a semi-supervised setting, wherein the discriminator and the generator are trained together. Given a set of *K* classes for text classification, the discriminator is trained to classify a piece of text into one-of-*K* classes. In addition to the *K* classes, we add an extra *K+1* class to train the discriminator to classify the samples generated from the generator into the $K + 1^{th}$ class. Introducing the additional class enables the network to learn from unlabelled examples as well.

3.3. Model Architecture

We leverage the GAN-BERT architecture to classify legal pharmaceutical text in a semi-supervised setting. Given a text sequence $s = (w_1, w_2, ..., w_n)$ consisting of n tokens, we leverage a pre-trained BERT model to obtain a contextual representation of s. The BERT model encodes each token to a $d_{real} \in \mathbb{R}$ dimensional contextualised vector. We consider the CLS token representation of the BERT model as the representation of s. Mathematically we define it as:

$$\mathbf{e} = \text{BERT}(s_{data}) \quad (2)$$

$$\mathbf{s}_{data} = \mathbf{e}([CLS]) \quad (3)$$

where BERT denotes the BERT encoding architecture.

On the other hand, during semi-supervised training of the GAN, we sample a $d_{fake} \in \mathbb{R}$ vector as the noise vector for the generator. This noise vector is fed as input to the generator for generating adversarial examples. The generator then generates a $d_{real} \in \mathbb{R}$ dimensional vector which is then sent to the discriminator for classification. Mathematically, the generator network is defined as:

$$z \sim Uniform(0, 1) \quad (4)$$

$$\mathbf{s}_{G(z)} = MLP(z) \quad (5)$$

where the MLP is a 5 layer dense network with LeakyRelu activation function as the non-linearity in each layer. We do not introduce any activation function in the final layer of the MLP. The noise vector is sampled from a uniform distribution (0, 1).

Similar to the generator, the discriminator is also modelled as a multi-layer perceptron with 5 dense layers with LeakyRelu activation function in every layer except for the last layer. As the role of the discriminator is to classify the text into $K+1$ classes, the output from the final layer layer is passed through a Softmax layer to assign probabilities of the sentence belonging to a specific class. Mathematically we define the discriminator and the final classification as:

$$\text{logits} = MLP(\mathbf{s}) \quad (6)$$

$$\text{P}_{class} = Softmax(\text{logits}) \quad (7)$$

where P_{class} denotes the probability of a text sequence belonging to a specific class.

For the final classification of a sentence, we utilise

Equation 8 to assign a class to the sentence.

$$\text{class}(s) = \begin{cases} \text{compliant} & p \geq \alpha \\ \text{non-compliant} & p < \alpha \end{cases} \quad (8)$$

where the threshold value α is a hyperparameter to be set.

4. Experimental Setup

In this section, we begin with a description of the dataset used in this work. Thereafter, we outline the baseline methods and the evaluation metrics on which we evaluate the performance of our proposed approach.

4.1. Dataset

We curated the dataset by considering external data sources concerning the pharmaceutical domain. This is a public dataset from Warning/Untitled letters from the FDA and FTC enforcements that was taken from the public data of largest pharmaceutical companies in the US. The dataset was then sent to a team of in-house experts for filtering low-quality instances. The resultant dataset contained 3,786 compliant sentences and 345 non-compliant sentences.

We split the final dataset into 70%, 15% and 15% as training, validation and test set. The resultant training set contained 2,784 and 245 compliant and non-compliant sentences respectively, while the validation and the test set contained 501 and 50 compliant and non-compliant sentences.

The Sentences being compliant and non-compliant is subject to the FDCA based on the information mentioned in them if any. Some examples from the dataset are mentioned in Table1.

4.2. Baseline and Evaluation Metrics

We compare our proposed approach again with the following baseline methods:

- Naive Bayes: We utilise the TF-IDF scores of tokens in the sentences to train a Naive Bayes model for the classification task.
- Multi-Layer Perceptron: We use the TF-IDF scores of tokens in the sentences as inputs to a 2-Layer dense neural network, with ReLu activation in the first layer, to train the classification 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): Sentence-BERT is Transformer (Vaswani et al., 2017) based sentence encoders that capture the rich semantic information in a sentence into a fixed-size vector. We encode each sentence using the Sentence-BERT architecture and then pass the sentence embedding to a 2 layer dense network for classification.
- LASER (Artetxe and Schwenk, 2019): Similar to the Sentence-BERT baseline, we encode each sentence using its LASER embedding and pass it to a 2 layer dense network for classification.

4.3. Implementation Details

For the Sentence-BERT baseline, we use the publicly available SBERT-BASE-NLI-MEAN-TOKENS [1] as our sentence encoder. While we utilise the LASER embeddings to encode the sentences to a 1024-dimensional vector. We use the publicly available BioBERT [2] as a replacement of the BERT model in our proposed approach.

We fine-tune the model with a learning rate of 5e-5 for both the generator and discriminator and batch size of 64 for 10 epochs with Adam optimizer. For the final classification, we plot the ROC curve and based on the distribution we set the α in Equation 8 to 0.7.

5. Results

In this section we outline the results of using our approach. We begin with the quantitative analysis of the performance of our approach against the baseline methods. Thereafter, we conduct an ablation study of replacing the Bio-BERT architecture with other BERT based encoders. Following this, we analyse the performance of our model in a few-shot setting wherein our approach and other baselines are supplied with a limited number of labelled examples. Finally, we conduct a qualitative analysis of the results and study a few cases where the labels assigned by our model is different from the gold-label. Analysis of our training suggests that, the training loss for generator and discriminator reduced with each epoch, and validation and test gave good results, which overruled the speculation of overfitting arising due to dataset being small to support deep learning models.

5.1. Quantitative Analysis

In this work, we propose a GAN based approach for pharmaceutical text classification. Table 2 outlines the performance of our proposed method against the different baseline methods used for the classification task. It can be observed that GAN-BioBERT achieves the best result amongst all models. It should be noted that GAN-BioBERT has a better performance than BioBERT showcasing the efficacy of our proposed approach.

5.2. Replacing BERT architecture

In our proposed approach, we replaced the BERT architecture in GAN-BERT with Bio-BERT model. To study the effectiveness of our choice of model, we replace the Bio-BERT model with BERT, RoBERTa and DistilBERT. We observe from Table 3 that the performance degrades when Bio-BERT is replaced with other

[1] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

[2] https://huggingface.co/dmis-lab/biobert-base-cased-v1.2

Table 1: Examples from our dataset to showcase the classes.

	Sentence	Model Label
1	FCS is a severe and rare disease caused by an enzyme deficiency that leads to the buildup of chylomicrons and a high risk of life-threatening pancreatitis.	compliant
2	We are committed to collaborating with the FDA to prevent or mitigate drug shortages that impact the health of patients.	compliant
3	Rosemary is one of the best essential oils that helps with headaches.	non-compliant
4	Tested and affordable Immune Plus Mouth Spray supports natural immune defense.	non-compliant

Table 2: Performance of our proposed method against different baseline methods for pharmaceutical text classification.

Model	Precision	Recall	F1	Accuracy
Naive Bayes	**1.00**	0.02	0.04	0.91
MLP	0.80	0.66	0.73	0.95
SVM	**1.00**	0.30	0.46	0.94
S-BERT	0.87	0.66	0.75	0.96
Laser	0	0	0	0.91
Bio-BERT	0.86	**0.86**	0.86	0.97
GAN-Bio-BERT	0.96	**0.86**	**0.91**	**0.98**

Table 3: Impact of the BERT architecture employed for the pharmaceutical text classification task.

Model	Precision	Recall	F1	Accuracy
GAN-BERT	0.81	0.84	0.82	0.97
GAN-RoBERTa	-	-	-	0.91
GAN-Bio-BERT	0.96	0.86	0.91	0.98

Table 4: Performance of our proposed approach against different baselines when a limited number of training examples are present. K denotes the number of training samples from each class used for training the models.

#Examples	Model	Precision	Recall	F1	Accuracy
K=10	BERT	**0.14**	0.26	0.18	0.78
	Bio-BERT	0.12	**0.58**	**0.20**	0.57
	GAN-Bio-BERT	0.00	0.00	0.00	**0.91**
K=20	BERT	0.11	0.16	0.13	0.80
	Bio-BERT	0.12	0.48	0.19	0.63
	GAN-Bio-BERT	**0.42**	**0.70**	**0.52**	**0.88**
K=50	BERT	0.13	0.54	0.21	0.63
	Bio-BERT	0.12	0.40	0.19	0.69
	GAN-Bio-BERT	**0.33**	**0.90**	**0.48**	**0.83**
K=100	BERT	0.17	0.74	0.28	0.65
	Bio-BERT	0.20	**0.86**	0.32	0.67
	GAN-Bio-BERT	**0.71**	0.68	**0.69**	**0.95**

BERT based models. This gives us a clear indication of the benefits of choosing a BERT model finetuned.

We observe from the Confusion Matrix of Figure 1 that the model misclassified only 9 out of the total 501 test examples and out of the 43 minority class examples only 2 were misclassfied. This clearly gives us the understanding that even after dataset being imbalanced the non-complaint class is aptly distinguished.

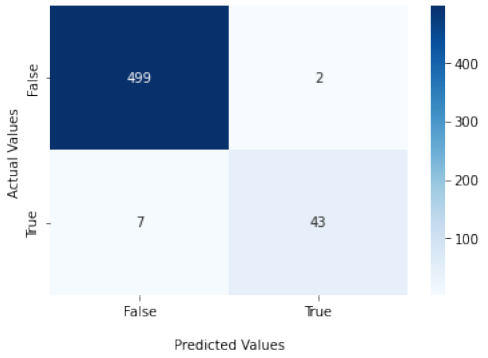

Figure 1: Confusion Matrix.

5.3. Few-Shot training

Getting labelled data can be time consuming and expensive. In this experiment, we train our proposed model with a limited number of labelled examples. We compare the performance of our proposed model again the BERT model and the Bio-BERT model.

Table 4 outlines the performance of different models when trained on a limited set of labelled data. It is interesting to notice that when there are only 10 labelled examples (K=10), GAN-Bio-BERT does not perform better than other baselines. This can be attributed to the generator generating poor quality samples, hence negatively impacting the performance of GAN-Bio-BERT. However, when the number of training samples (K) is increased, GAN-Bio-BERT outperforms different baselines by a large margin on the F1 score as well as Recall. This result demonstrates the efficacy of our proposed model on the classification task when a limited number of training examples are provided.

5.4. Qualitative Analysis

In this section, we analyse a few examples where the model assigns a different label to the sentence than the gold-label. Table 5 outlines four such cases. In the first and second examples, we can observe that the sen-

Table 5: Error Analysis: Examples where our proposed model produces classification labels different from the gold labels.

	Sentence	Model Result	Gold Label
1	Indulge in life's sweetest pleasures whenever you want.	compliant	non-compliant
2	Lower production of proinflammatory cytokines.	compliant	non-compliant
3	It is an anticholinergic medicine which helps the muscles around the airway in your lungs stay relaxed to prevent symptoms such as wheezing, cough, chest tightness, and shortness of breath.	non-compliant	compliant

tences are non-compliant but have been assigned the compliant class by the model. This might be due to the fact that these sentences seem incomplete, without more information it is difficult to say that they are non-compliant as they just state something without context. For the third sentence, we can observe that the sentences are compliant but have been assigned the non-compliant class by the model, this might be due to the context for both which implies that those specific medicines definitely work for the said symptoms, but it is extremely hard to know without having any knowledge about the medicines or context. This show that proper understanding of use cases of medicines and possible context about the sentences might help to correctly classify them.

6. Conclusion

In this work, we propose the use of predictive coding for the classification of pharmaceutical texts in the industry. We leverage the GAN-BioBERT architecture for the task and showcase its efficacy against different methods on multiple metrics. Additionally, we conduct a thorough ablation study to show the impact of our model of choice for the task.

7. 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.

8. Bibliographical References

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

Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., and Lu, Z. (2019). ML-Net: multi-label classification of biomedical texts with deep neural networks. *J. Am. Medical Informatics Assoc.*, 26(11):1279–1285.

Flores, C. A., Figueroa, R. L., and Pezoa, J. E. (2019). FREGEX: A feature extraction method for biomedical text classification using regular expressions. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2019, Berlin, Germany, July 23-27, 2019*, pages 6085–6088. IEEE.

Flores, C. A., Figueroa, R. L., Pezoa, J. E., and Zeng-Treitler, Q. (2020). CREGEX: A biomedical text classifier based on automatically generated regular expressions. *IEEE Access*, 8:29270–29280.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Jofche, N., Mishev, K., Stojanov, R., Jovanovik, M., and Trajanov, D. (2021). Pharmke: Knowledge extraction platform for pharmaceutical texts using transfer learning. *CoRR*, abs/2102.13139.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Luo, Y. (2017). Recurrent neural networks for classifying relations in clinical notes. *J. Biomed. Informatics*, 72:85–95.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, et al., editors, *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.

Sarkar, R., Ojha, A. K., Megaro, J., Mariano, J., Herard, V., and McCrae, J. P. (2021). Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 102–106, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Isabelle Guyon, et al., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wu, J., Zhang, R., Gong, T., Liu, Y., Wang, C., and Li, C. (2021). BioIE: Biomedical information extraction with multi-head attention enhanced graph convolutional network. In Yufei Huang, et al., editors, *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, December 9-12, 2021*, pages 2080–2087. IEEE.

Yao, L., Mao, C., and Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics Decis. Mak.*, 19-S(3):31–39.