Multi-label Classification using BERT and Knowledge Graphs with a Limited Training Dataset

Malick Ebiele[1,*,†], Lucy McKenna[1,†], Malika Bendechache[1,†] and Rob Brennan[2,†]

[1]ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
[2]ADAPT Centre, School of Computer Science, University College Dublin, Dublin, Ireland

Abstract

This paper provides a new approach combining BERT and Knowledge Graphs (KGs) to solve a multi-label classification problem with limited training data. The paper introduces a method of using taxonomies and a dataset with 518 entries and 340 concepts to fine-tune BERT. It also introduces a new data augmentation technique called Perfect Binary Tree (PBT)-Flow to deal with limited or imbalanced training data. The proposed approach obtained a recall@10 of 61.12%, a precision@10 of 11.86% and F1score@10 of 18.83%. While these results seem low, they are promising because of the simple architecture of the model used (BERT+2xFC), the limited size of the training data, and the large number of output concepts.

Keywords

Multi-label classification, BERT, Knowledge graphs, Data augmentation

1. Introduction

Multi-label classification is the task of assigning one or more concepts to an object or text [1]. This is a challenging task, especially with limited training data and large number of output concepts. Fortunately, the complexity of the task can be reduced by using a KG of the concepts. In fact, by leveraging the hierarchy defined in the concepts' ontology, one can considerably simplify the task at hand with little loss in semantic if the ontology is complete, well formed, semantically consistent and high quality. BERT [2] is a machine learning framework for natural language processing (NLP) that can be applied to multi-label classification. GAN-BERT is an extension of BERT by using Semi-Supervised Generative Adversarial Networks for the fine-tuning stage [3].

This paper introduces a new approach combining BERT and KGs to classify textual data and a new data augmentation technique called Perfect Binary Tree-Flow (PBT-Flow).

This paper investigates the research question: "To what extent can KGs, BERT and the PBT-Flow data augmentation technique improve the precision, recall and F1 score of multi-label classification using a limited training dataset"? In order to explore this research question, the

Woodstock'22: Symposium on the irreproducible science, June 07–11, 2022, Woodstock, NY
*Corresponding author.
†These authors contributed equally.
malick.ebiele@adaptcentre.ie (M. Ebiele); lucy.mckenna@adaptcentre.ie (L. McKenna); malika.bendechache@adaptcentre.ie (M. Bendechache); rob.brennan@adaptcentre.ie (R. Brennan)
0000-0001-5019-6839 (M. Ebiele); 0000-0002-6035-7656 (L. McKenna); 0000-0003-0069-1860 (M. Bendechache); 0000-0001-8236-362X (R. Brennan)

ARK-Virus Project [4] was selected as a use-case. The ARK-Virus Project is an extension of the ARK Platform [5] for risk management of personal protective equipment in healthcare settings during the COVID-19 pandemic. This is discussed in more detail in the Use Case and Requirements Section below (Section 3).

This paper has two main contributions. First, a method which uses KGs to simplify multi-label classification by reducing the number of output concepts. Second, the presentation of the PBT-Flow data augmentation technique for dealing with limited or unbalanced training datasets.

The remainder of this paper is structured as follows; Section 2 presents the related work. Section 3 describes the use case and requirements. Section 4 depicts the design of the proposed approach. Section 5 details the experimental settings. Section 6 provides an evaluation and Section 7 presents the conclusion.

2. Related Work

Different mechanisms have been used to solve multi-label classification problems. Rios and Kavuluru [6] proposed a model combining a Convolutional Neural Network (CNN) and a 2-Layer Graph CNN (GCNN) to perform experiments on MIMIC II and MIMIC III. Heo et al [7], on the other hand, proposed D2SBERT, a sequence of n BERT+Multilayer Perceptrons (MLPs) with an attention layer in between for medical discharge summary code prediction. Finally, Khezrian et al [8], introduced TagBERT (BERT+CNN+MLP) to produce tag recommendations for Online Q&A communities and performed experiments on the Freecode dataset. None of these previous works leveraged the possible hierarchy of the label space or did not perform an optimised data augmentation. This work differs from them at two levels. One, by leveraging the hierarchy of the label space by using KG. Two, by introducing and using an optimised data augmentation technique.

3. Use Case and Requirements

The ARK-Virus Project uses the ARK-Platform, a socio-technical risk governance system, to manage and analyse risk projects in the area of infection prevention and control. Data entered on the ARK Platform can be annotated with concepts from controlled SKOS[1] taxonomies - the ARK Risk Terminology and the ARK Health Terminology[2]. These taxonomies contain a combined total of 525 concepts plus definitions. The taxonomies use a three-layer hierarchy with the top level having a total of 141 concepts. It is also worth mentioning that these taxonomies have been built, used, and validated by domain experts over the past couple of years. The annotation of text on the ARK Platform is currently a manual process which, given the large number of concepts, can be time-consuming. Providing a set of suggested concepts, based on text entered into the ARK Platform, would be extremely useful to users. This paper demonstrates how KGs and BERT can be used together to solve this multi-label classification problem.

[1]Simple Knowledge Organization System - https://www.w3.org/TR/skos-reference/
[2]Taxonomies and platform demo available at https://openark.adaptcentre.ie/

The approach presented in this paper can be applied to other use cases. The only requirement that needs to be met to apply the approach introduced here is to have an hierarchical label space. However, the ontology of the label space should be well formed, semantically consistent, and high quality [9] with a reasonable amount of concepts in the top layer. The KG's structure can negatively affect the model performance, for example, if a broad domain is modelled with a narrow taxonomy. This will lead to a loss of the semantic which will negatively impact the performance of the proposed approach. Future work will assess how much this loss will impact the performance of the proposed approach. In this paper, the top level labels have been used for one main reason. The labels' taxonomies only have three layer hierarchy which makes the loss of semantics acceptable compared to a much complex multi-label classification task given the low resources. One could have used the second top or the second last layer or any other layer for a deeper taxonomy. The idea is to leverage the taxonomies hierarchy to simply the problem at hand.

4. Design

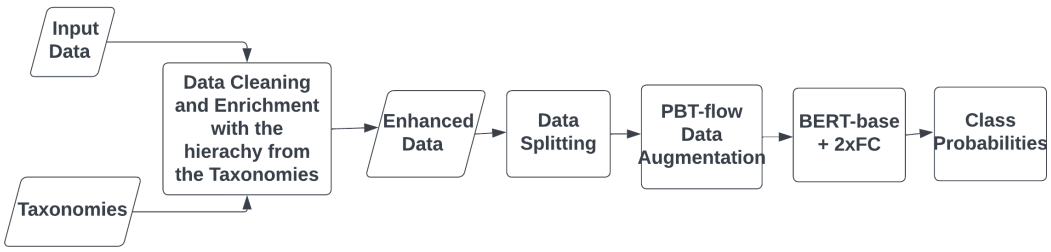

Figure 1: Knowledge Graph (KG) Enhanced BERT Training Process for a Limited Training Dataset.

Figure 1 outlines the proposed approach. First, the input data, a tabular data extracted from the ARK Platform where each entry is a collection of sentences annotated with concepts defined in the ARK taxonomies, is cleaned and enriched. The process of enrichment consists of replacing the original concepts with the top level concepts from the taxonomy hierarchy. In this case, this process reduced the total number of concepts from 340 to 116. Second, the new data of 518 entries is then split into 50%, 25% and 25% for training, test, and validation sets, respectively (see appendix A for more details). Third, the training set is augmented using the PBT-Flow technique. Fourth, the augmented data is fed into the model (BERT+2xFC). Fifth and finally, the model outputs the class probabilities. The performance is then measured using the precision@k, recall@k and F1score@k metrics with k={5,10}.

5. Experimental Settings

Given the limited training data, data augmentation was used to acquire more data. Data augmentation is a technique for increasing the diversity of training data without explicitly

collecting new data [10]. The PBT-Flow[3] Data Augmentation Technique consists of applying a set of data augmentation techniques to the input data following the Perfect Binary Tree (PBT) structure. The nodes of the tree represent the datasets and the edges indicate whether or not an augmentation technique was applied on the data. Each node has two children - one child is the result of applying an augmentation on the node (data) and the other child is a copy of the parent (no augmentation applied). Every node of the same depth is applied the same augmentation technique and each augmentation technique is applied once. The output of PBT-Flow is the concatenation of the leaves of the PBT. PBT-Flow generates a new dataset of $n2^m$ entries from an input data of n entries and a set of m data augmentation techniques.

PBT-Flow used five augmentation techniques: two Synonym Replacements [11] (k-words, with k between 1 and 10, from the original text are replaced with their respective synonyms using PPDB and WORDNET vocabulary databases , Back Translation [12] (the original text in English is translated to Deutch language then translated back), Random Swap [13] (Randomly swap k-words of the original text), Contextual Augmentation [14] (k-words in the original text are replaced with other words with paradigmatic relations). Applying the PBT-Flow technique to the original training data of 259 entries resulted in a training data of 5232 entries (after removing duplicated entries and NAs). For GAN-BERT [3], entries with less than 4 concepts were unlabelled which resulted in sets of 809 labelled and 4423 unlabelled training data.

6. Evaluation

Two approaches have been used to fine-tune BERT: a supervised approach and a semi-supervised approach based on GAN-BERT [3]. The Generator and the Discriminator of GAN-BERT are defined as discussed in [3]. However, the soft-max function has been replace by the sigmoid function and the cross-entropy-loss by the binary-cross-entropy-loss. The classifier of supervised learning model is the same as the Discriminator minus one neuron in the output layer (because the output layer of the Discriminator has $116 + 1$ neurons; one extras neuron to output the probability of the input text being fake or real). The data was augmented using the five augmentation techniques mentioned above. BERT run for 34 epochs and GAN-BERT for 11 epochs. Early stopping monitoring the validation loss with patience and min_delta equal to 5 and 5e-05, respectively, have been used. From Table 1 it can be seen that GAN-BERT+PBT-Flow model outperformed other models by 0.24% (Recall@10) up to 11.89% (Recall@5). In general, models using PBT-Flow outperformed the others. These results validate the experimental results presented in [3]. Those results showed that GAN-BERT outperformed BERT when both models are fine tuned using very limited labelled data.

One can notice that the margin of improvement in GAN-BERT is greater than in BERT. This could be due to the filtering of inputs labelled with less than 4 concepts for this model.

7. Conclusion

This paper demonstrates that combining KGs and PBT-Flow improve BERT models' performance for multi-label classification, in both supervised and semi-supervised approaches. These results

[3]Code source available at https://github.com/malick-jaures/research/tree/main/PBT-flow

Table 1
Experimental Results. (*NoAug - no augmentation applied)

	Precision@k		Recall@k		F1score@k	
Models	k=5	k=10	k=5	k=10	k=5	k=10
BERT (NoAug)	15.34	10.07	39.37	51.13	20.80	16.04
BERT + PBT-Flow	18.29	11.62	46.25	58.74	24.66	18.48
GAN-BERT (NoAug)	15.50	10.07	39.14	50.56	20.92	15.95
GAN-BERT + PBT-Flow	**19.53**	**11.86**	**51.03**	**61.12**	**26.34**	**18.83**

are interesting; combining with a threshold could suggest good concepts. While the results cannot be directly compared to the state-of-the-art models, they are similar to previously published works especially in terms of recall@10 [8, 6]. TagBERT [8] is the state-of-the-art in term of Precision@10 and F1score@10 on the Freecode dataset with 40.25% and 46.5%, respectively. On the other hand, the same model obtained a Recall@10 of 64.42% while TagCNN [15] achieved 94.9%.

In future work, we envisage to run experiments which will consist of testing our approach on public benchmarks along with TagBERT, TagCNN and other models. We also intent to retrain our model with data extracted from the ARK platform as soon as more data will be available on the platform.

Acknowledgments

This research has received funding from the ADAPT Centre for Digital Con- tent Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106 P2), co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

We would also like to express our gratitude to Dr. Brian Davis for his advice and support.

References

[1] G. Tsoumakas, I. M. Katakis, Multi-label classification: An overview, Int. J. Data Warehous. Min. 3 (2007) 1–13.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805.

[3] D. Croce, G. Castellucci, R. Basili, GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples, 2020.

[4] L. McKenna, J. Liang, N. Duda, N. McDonald, R. Brennan, Ark-virus: An ark platform extension for mindful risk governance of personal protective equipment use in healthcare, 2021.

[5] N. McDonald, L. McKenna, R. Vining, B. Doyle, J. Liang, M. E. Ward, P. Ulfvengren, U. Geary, J. Guilfoyle, A. Shuhaiber, J. Hernandez, M. Fogarty, U. Healy, C. Tallon, R. Brennan,

Evaluation of an access-risk-knowledge (ark) platform for governance of risk and change in complex socio-technical systems, International Journal of Environmental Research and Public Health 18 (2021).
[6] A. Rios, R. Kavuluru, Few-shot and zero-shot multi-label learning for structured label spaces, 2018.
[7] T. Heo, Y. Yoo, Y. Park, B. Jo, Medical code prediction from discharge summary: Document to sequence BERT using sequence attention, CoRR abs/2106.07932 (2021). URL: https://arxiv.org/abs/2106.07932.
[8] N. Khezrian, J. Habibi, I. Annamoradnejad, Tag recommendation for online q&a communities based on BERT pre-training technique, CoRR abs/2010.04971 (2020). URL: https://arxiv.org/abs/2010.04971.
[9] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey (2016). doi:10.3233/SW-150175.
[10] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. H. Hovy, A survey of data augmentation approaches for NLP, CoRR abs/2105.03075 (2021). URL: https://arxiv.org/abs/2105.03075.
[11] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, volume 28, Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.
[12] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, CoRR abs/1511.06709 (2015). URL: http://arxiv.org/abs/1511.06709.
[13] J. Wei, K. Zou, EDA: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.
[14] S. Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relations, 2018.
[15] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, 2014.

Appendix

A. Statistics of our dataset

Table 2 below gives the statistics of the number of concept per entry of the training, test, and validation sets used in the experiments above.

Table 2
Statistics of the number of concept per entry of the training, test, and validation sets

Sets	count	mean	std	min	25%	50%	75%	max
Training	259	1.9	1.4	1	1	1	2	10
Test	129	1.9	1.3	1	1	1	3	7
Validation	130	1.9	1.3	1	1	1	2	8

B. Our dataset compared to Freecode dataset

Our dataset has 518 entries with 116 unique labels while Freecode[4] dataset has 46995 entries with 9000 unique labels. This means that Freecode dataset contains about 77.6 times more labels but also 90.7 times more entries than ours. In other words, Freecode has a ratio of 5.22 entries per label while our dataset has a ratio of 4.46 entries per label. Moreover, the top 3 most representative set of labels in Freecode have respectively 1390, 711, and 571 entries. In our dataset, the top 3 most representative set of labels have respectively 60, 19, and 13 entries. The least representative set of labels in both datasets have only 1 entry.

Table 3 below gives the statistics of the number of concepts/tags per entry of our dataset and the Freecode dataset.

Table 3
Statistics of the number of concepts/tags per entry of our and Freecode datasets

Datasets	count	mean	std	min	25%	50%	75%	max
Ours	518	1.93	1.35	1	1	1	2	10
Freecode	46995	3.55	2.48	1	2	3	5	38

Table 4 below gives the statistics of the number of words per entry of our dataset and the Freecode dataset.

Table 4
Statistics of number of words per entry of our and Freecode datasets

Datasets	count	mean	std	min	25%	50%	75%	max
Ours	518	31.93	35.79	1	12	21	38	328
Freecode	46995	50.3	26.9	1	31	45	66	330

[4]Available at https://www.kaggle.com/datasets/navidkhezrian/freecode