Unsupervised Text Clusterisation to characterize Adverse Drug Reactions from hospitalization reports

Xuchun Zhang, Milou Daniel Drici, Michel Riveill

Unsupervised Text Clusterisation to characterize Adverse Drug Reactions from hospitalization reports

Xuchun ZHANG[1,2][0000−0002−8759−5240], Milou-Daniel DRICI[1,3], and Michel RIVEILL[1,2][0000−0001−6726−6637]

[1] Université Côte d'Azur, France
[2] CNRS, Inria, France
{xuchun.zhang, michel.riveill}@inria.fr
[3] CHU Nice, France
drici.md@chu-nice.fr

Abstract. The detection of Adverse Drug Reactions (ADRs) in clinical records plays a pivotal role in pharmacovigilance (PhV). Achieving near-ideal practice relies on well-trained health professionals, who are trained to identify, assess and report to health authorities ADRs occurring after drug marketing approval, including those that are infrequent. However, the number of experts trained in this practice is low and despite reporting ADRs being mandatory for healthcare professionals, pharmacovigilance still suffers from a significant under-reporting, accounting for only 5-10% of all ADRs. Yet, drug safety is crucial for assessing the benefit/risk ratio of a given drug. It is therefore important to circumvent under-reporting and to be able to collect ADRs automatically from medical reports. The most natural approach would be to train a model in a supervised manner, which requires annotation of a large volume of data, but this is unfortunately not possible. We therefore propose here an unsupervised approach to distinguish between ADRs-related and non-related reports. From a more formal point of view, we address this problem as a clustering task aiming at distinguishing medical reports containing the description of an ADR from those without.

Keywords: Text Clustering · Unsupervised Learning · Adverse Drug Reactions.

1 Introduction

Pharmacovigilance (PhV), by its definition from World Health Organisation (WHO), is "the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other medicine/vaccine related problem." [20], which concerns drug regulatory to ensure that the authorities of medical products are well studied on safety issues in everyday practice. Whereas rigorous testing must be done during the drug development program before its marketing approval, the issue of safety is not absolute. One of the reasons is that the clinical trials involves a relatively small number of quite selected

participants comparing to the large potential number of patients who will use the drug in real life. Another reason is that these trials are conducted within a limited time frame, which precludes the characterization of certain chronic adverse reactions that may occur over a longer period of time.

Managing Adverse Drug Reactions (ADRs) is one of the most important post-marketing PhV practice, since serious ADRs are thought to be responsible for 5-10 percent of hospitalisations. Pharmacovigilance aims at detecting and monitoring ADRs in real life settings, and more frequently nowadays, from hospital clinical reports or Electronic Health Records (EHRs), owing to the rich information about patient health and the structured textual content that were written by professionals working in the domain. After review, confirmation, and causality assessment by trained pharmacologists, this detection will be recorded in the National Agency for the Safety of Medicines and health products (ANSM)[4]. Then, data from the national database will be fed into the VigiBase database at the Uppsala Monitoring Centre (UMC). VigiBase is the unique global database of WHO (World Health Organization) reporting potential side effects of medicinal products. It is the largest database of its kind in the world, with over 30 million reports of suspected adverse effects of medicines, submitted, since 1968, by member countries of the WHO Programme for International Drug Monitoring (PIDM). It is continuously updated with incoming reports.

Achieving a more ideal practice relies heavily on well-trained health professionals, who are more likely to have sufficient experience to identify, assess and and report important ADRs [13]. Despite being mandatory for health care probationers to report ADRs when suspected, notifications of ADRs amount to a mere 5-10 percent of all ADRs. However the efficiency to detecting ADRs is limited due to the lack of well-trained professionals, the underreporting and the enormous amount of clinical reports at disposition.

Deep learning has boosted the development of Natural Language Processing (NLP) and showed that NLP can be a solution to practice efficiently and accurately in biological analysis, and it is getting more and more attentions from the researchers. Many shared tasks/workshops [8, 9, 19] are conducted in exploitation of possibilities of ADR detection by modern deep learning NLP techniques, which provides us with an overview of how powerful these techniques are on the annotated corpus. Despite the good performances, the state-of-art supervised NLP techniques could achieve in ADRs detection from annotated corpus, we cannot ignore that one need a large number of annotated data to train a supervised model but getting such amount of annotations is extremely expensive. On the other hand, the rapid increasing amount of EHRs without annotations are remain unexploited. To bridge this gap, we present in this paper a new unsupervised approach to help finding potential EHRs with ADRs descriptions.

[4] https://ansm.sante.fr/ and https://ansm.sante.fr/page/la-surveillance-renforcee-des-medicaments

2 Related work

Because of the rarity of EHRs related to adverse events and the limited public access to clinical records, given patient privacy and confidentiality, the first approaches used statistical analysis and feature-based methods, which try to characterize the likelihood of a candidate drug-symptom relation to be categorized as a true ADR. [10] built a knowledge base with information from the Unified Medical Language System (UMLS) to determine whether the recognized concept matches a relation of ADR. The explosion of machine learning researches in this domain has drawn the attention of creating publicly available annotated data. The 2010 Informatics for Integrating Biology and the Bedside/Vetaran Affairs (i2b2/VA) challenge [16] provides clinical records for concept extraction, assertion classification, and relation classification. The 2018 National NLP Clinical Challenges shared task (n2c2) [8] provided 505 discharge summaries for 3 different tasks: concept extraction, relation classification, and end-to-end systems construction. With similar tracks as 2018 n2c2 shared task, the MADE (Medications and Adverse Drug Events from Electronic Health Records) 1.0 challenge provides real de-identified EHRs and corresponding annotations for medications, symptoms and ADRs. The increasing accessibility of EHRs to the community has the potential to identify more EHRs with ADRs and thus makes more pharmacovigilance data available and benefits the performance of machine learning models and form a virtuous circle.

Such workshops attracts great attention from NLP research community, people in this area have proposed many supervised approaches for ADR extraction. In MADE 1.0 challenge [9], Chapman et al.[2] developed a two-stage approach by first identifying the named entities based on conditional random field (CRF), and then assigning the relevant relation type between entities based on random forest (RF) and achieve the highest score for Relation Identification (RI) task. Dandala et al. [4] adopted a combined bidirectional long short-term memory (BiLSTM) with CRF for named entities recognition (NER) and applied attention-based BiLSTM network together with medical domain ontology information from unified medical language system (UMLS) to RI task, which is the highest performing system in joint Relation Identification (NER-RI) task. Studies in n2c2 challenge [8] shows potential of deep learning models for ADR extraction. Among the best performance systems, Wei et al. [18] applied a joint-learning-based BiLSTM-CRF for both NER and RI tasks, where they conducted rule-based postprocessing to fix the obvious errors and improve the prediction. Christopoulou et al. [3] proposed a weighted BiLSTM combining a walk-based model to reasoning intra-sentence relations and a Transformer-based network to memorising inter-sentence relations. IBM Research team explored a combination of piece-wise neural networks [21] and an attention-based BiLSTM. More recently, El-allay[6] proposed a joint model with transformer and Weighted Graph Convolutional Network (WGCN) to capture ADR relations and proved its state-of-the-art performance on n2c2 dataset.

In recent years, the NLP community has demonstrated the great power of supervised machine learning techniques for ADR extraction. However, the unsu-

pervised approaches still remains uncertain and under-exploited. Pérez et al. [14] first tried analysing vector representation for ADRs from EHRs written in Spanish by linking word2vec embeddings of drug-symptom entities pair in semantical space, which shows the potential of expressing correlation between ADR and non-ADR. More recently, Bampa et al. [1] explored encoding the document type without considering too much the textual content and by clustering aggregation [7] techniques to grasp information about the phenotype of patient/document, which provided decent cluster structure for ADR analysis.

3 Method

In this section, we describe our unsupervised ADR-related records detect system. Fig 1 shows the overall structure of our model as well as its main components. By the definition of the ADR, it is obvious that its occurrence will always relate to a drug-symptom entities pair, and the contextual contents around the target drug and target symptom indicates its existence. Since the majority of clinical records were generated by hospital health care practitioners, the documents bear a well-organised structure with many medical terms like medication, chemical names, symptoms, medical observations and diagnoses etc... We assume that the source mentions for drug and symptom related entities is given and we need to find ADR-related records. Our system takes the clinical records as inputs and process the records and apply a filter algorithm to choose the potential blocks for further purpose. Then, the blocks will be tokenized and fed to the model for unsupervised learning.

3.1 Preprocessing

We assume that for any ADR, both the drug and the adverse effect are described within the same block of textual content. We defined henceforth "block" as the basic unit of textual content to analyse, which can be either whole document, paragraph, phrase, sentence, etc.

Then we can define the problem as: Let $B = \{\beta_1,\ \beta_2,\ ...\ \beta_N\}$ with N blocks of literature, each block β_i contains textual contents together with annotations for drug and for symptom entities (In a text, it is simple to locate drugs by consulting domain ontologies that explain the molecules and trade names of those who has marketing authorisation, and symptoms have also their universally codified medical definitions). Take the block "He was better controlled on **Velcade**, but developed significant **peripheral neuropathy**" as an example, where we see the drug "Velcade" and the symptom "peripheral neuropathy" in the text.

We want to separate the blocks with the description of ADR (noted as positive block β^+) from those who don't (noted as negative block β^-). As a result, the blocks that do not include any drug or symptom will be of little interest to us. We made the hypothesis that the ADR relation lies in the contextual content between drug and symptom entities, based on which, we want to reduce the influence of drug and symptom entities and increase the model's emphasis on the

context. To preprocess the drug/symptom entities, taking sentence "He was better controlled on **Velcade**, but developed significant **peripheral neuropathy**" as example, we presented four strategies:

- **Keep the entities**: "He was better controlled on **Velcade**, but developed significant **peripheral neuropathy**"
- **Replace drug entities by word 'drug' and symptom entities by word 'symptom'**: "He was better controlled on drug, but developed significant symptom"
- **Masking both drug and symptoms entities**: "He was better controlled on [MASK], but developed significant [MASK]"
- **Remove the drug/symptom entities**: "He was better controlled on, but developed significant"

We took finally the "removing drug/symptom entities" strategy to preprocess text with entities information as its best performance among the models.

3.2 Unsupervised BERT based ADR block detection

The pre-trained language models, including BERT (Bidirectional Encoder Representation from Transformer) [5], a two-stage Transformer[17]-based natural language representation framework proposed by Google Brain in 2018, shows in recent years its great potential in extraction of features from textual content, which push significantly the state-of-the-art performance in many aspects in NLP domains. We here utilised the pre-trained BERT models without fine-tuning it since the latter requires a huge corpus to support.

BERT-based transformation model split each word in input text into word-piece tokens and takes the tokenized words sequence as its own input to encode each input text into vectors of the same size in the same semantic space, which means that the basic BERT-based models embed each word-piece token but not the whole sequence. As to infer a single representation for one block, we chose to applied pooling to the embedded tokens. Asides from the original BERT model, we also tried Sentence-Bert (SBERT) [15] that take BERT as basic component and considered training it with a siamese and triplet networks, in order to catch representation not for words but for the whole sequence and thus it can map directly a sentence like input to the vector space with common similarity measures like cosine-similarity. In our case, we used the pre-trained sentence encoder part from Sentence-Bert to encode the block content into one single vector representation.

We make the hypothesis that, the description of ADR-related information lies between the medication and symptom entities. We therefore chose the textual content for each block by removing all drug or symptom associated entities as the input for all language models. This input is processed by the tokenizer of the model and then the model itself to represent each input in their own way. As we mentioned above, for the output of BERT-based models, since they encode every word-piece tokens of a block into vectors with same size, we need to apply

an extra average pooling to it to obtain one vector representation for one block as the SBERT model does. Once getting all vector representations from the language model, we applied KMedoids cluster algorithm to create clusters of similar blocks. KMedoids clustering is similar to the popular KMeans clustering algorithm, both aim to reduce the distance between points labeled as belonging to a cluster and a point designated as the cluster's centre, we choose the former due to its robustness to noise and outliers than the latter and also its flexibility with arbitrary dissimilarity measures. The ideal practice is to obtain a cluster with only positive blocks and another with only negative ones. The structure of the model is shown in Figure 1

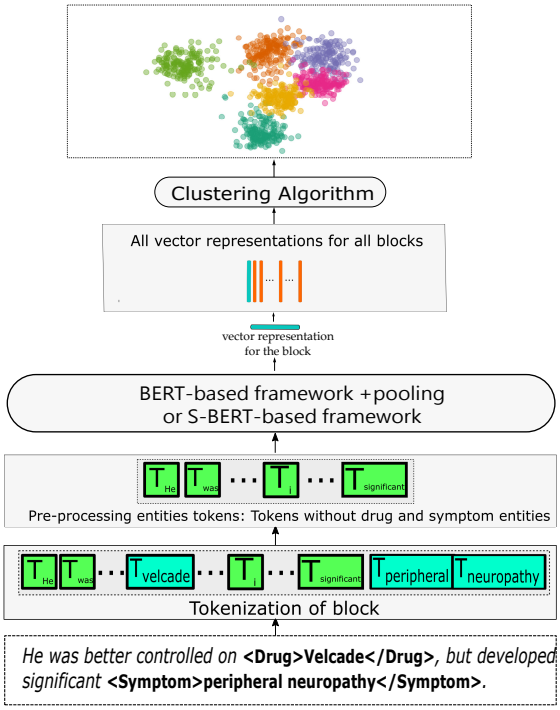

Fig. 1. The structure of our model. For each blocks, only the tokens of contextual content around drug/symptom named entities are selected for BERT-related embedding, for BERT-like models we applied a pooling strategy between tokens from the last layer to obtain a single vector representation for the block as the Sentence-BERT like models do, and then the output embedding vectors will go through the clustering algorithm to get the cluster assignment for each one of them.

4 Experimental results

4.1 Datasets

We chose here block in sentence-level, which leads to a relatively small span of text been chosen comparing to a whole length of documentation, which means also the tokenized sequence nearly exceeds the length limit of BERT-based models. We also note that we didn't take the irrelevant examples as input for our models, such as blocks with no entities or blocks with only one type of entity. The statistic of the dataset used is shown in Table 1

MADE dataset This is the data used in MADE challenge, whose corpora collected from 21 randomly selected cancer patients at the University of Massachusetts Memorial Medical Center, with the annotations of drugs, symptoms and ADRs. We choose this dataset considering the nature of source being real life clinical notes and its high quality of annotations. We sampled two sets of blocks that contains both drug and one symptom (since in unsupervised system, we have no idea in advance that whether the symptom is adverse effect of the drug or the cause of taking the drug or even an irrelevant symptom) by the help of entity type information for drugs and symptoms in EHR and we can extract two datasets as following:

- **MADE multi-d-s** Each block contains at least one drug and one symptom. All examples not containing at least one drug and one symptom were removed. This dataset contains long blocks of the EHR corpus with an almost balanced distribution between negative and positive blocks.
- **MADE 1d1s** For the dataset above, we extract those who has exactly one drug and one symptom, which called "1d1s" as "perfect situation", a nearly balanced dataset with short blocks from well-written EHR corpus.

CADEC CSIRO Adverse Drug Event Corpus (CADEC) [11] dataset is a rich annotated corpus of medical forum posts "Ask a Patient", which is dedicated to ADR-related consumer reviews on medications, which is mostly written in colloquial language and often deviates from the formal rules of English grammar and punctuation. The annotations contain entities such as drugs, ADRs, symptoms and diseases related to their respective concepts in SNOMED Clinical Terms and MedDRA. We performed the same pre-selection as we did for MADE data, we kept all blocks that has at least one drug and one symptoms.

4.2 Evaluation metrics

We chose precision, recall and F1-score as categories of metric to evaluate the result produced. The fraction of documents retrieved that are relevant to the query, is known as precision, which can be given by the formula $Precision = \frac{TP}{TP+FP}$, where TP and FP represent the number of real positive examples and

Table 1. MADE and CADEC dataset statistics

Dataset	β^+	β^-	**Sum**
MADE 1d1s	301	416	717
MADE multi-d-s	651	571	1222
CADEC	1107	79	1186

real negative examples among all that have been retrieved as positive examples. Recall is the number of correct results divided by the number of expected results, whose formula is $Recall = \frac{TP}{TP+FN}$, where FN indicates the number of retrieved negative examples that are really positive ones. F1-score is the harmonic mean of recall and precision, with the formula as $F_1 = \frac{2 \times precision \times recall}{precision + recall}$

4.3 Experimental Settings

As we mentioned in section 3.2, we mainly used three models to encode information from text: 1) the original BERT model "bert-base-cased". 2) the BioBERT [12] model, who uses the same structure as the BERT model pretrained and fine-tuned on biomedical corpora instead of employing general domain text corpora, to create a BERT model that specialises in describing features in biomedical literature. We introduced BioBERT here to verify if domain-specific knowledge has great impact on represented latent ADE information in the textual content. We used in our expriments the model "biobert-base-cased-v1.1". And 3) the Sentence-BERT (SBERT) "sentence-transformers/bert-base-nli-mean-tokens". In order to get fully representation for whole block for the first two models 1) and 2), in order to obtain the corresponding block representation in high quality, we applied average pooling to the output from short block in **MADE 1d1s** dataset and extract the "cls" token for long block in **MADE multi-d-s** and **CADEC** dataset. We choose cosine similarity as the metric and use KMedoids as clustering algorithm due to its flexibility with this measure, and set number of clusters as 2 for the purpose of evaluation.

4.4 Experimental results

We chose a fully supervised yet simple approach, a Bag of Words + Logistic Regression Classifier as the baseline of upper bound and a Bag of Words + completely random classifier as lower bound. The results are demonstrated in Table 2

From the Table of 2 we can see that for the **MADE 1d1s** dataset, compared to the lower bound, the representation provided by basic BERT model itself is not enough to capture the essential information about ADR. However, BioBERT wins BERT for its biomedical domain specified dictionary which helps it to represent better the medical text, with a highest recall value among the unsupervised methods, which means it is more reliable when we focus on retrieving more examples from the real positive ones. On the other hand, comparing to

Table 2. Comparison with supervised baseline and our unsupervised approach, we report the average *Precision*, *Recall* and F_1 scores of 5-fold cross validation. The results for unsupervised approaches (*) are always followed by a KMedoids clustering

Category	Exp	MADE 1d1s			MADE multi-d-s			CADEC		
		Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Supervised	BOW+LR	0.702	0.797	0.746	0.809	0.847	0.828	0.939	0.993	0.965
Unsupervised	$BERT^*$	0.549	0.463	0.502	0.591	0.634	0.612	0.950	0.520	0.672
	$BioBERT^*$	0.651	**0.663**	0.657	0.653	**0.673**	**0.663**	0.938	0.570	0.709
	$SBERT^*$	0.733	0.615	**0.669**	0.666	0.593	0.627	0.958	0.492	0.650
Supervised	BOW+Random	0.514	0.529	0.522	0.509	0.440	0.472	0.946	0.509	0.662

BioBERT, the SBERT embedding + clustering strategy could achieve a higher F1 score (0.669 vs 0.657 for BioBERT) with a higher precision (0.733 vs 0.651 for BioBERT) but lower recall value. This has showed us that the domain specific pre-trained language model BioBERT and sentence embedding framework SBERT has great potential in representing adverse event lied in textual data.

As for the **MADE multi-d-s** data, the augmented number of textual data with longer length boost the performance for supervised baseline. Introducing more textual content means more access to potential information, but also leads to more irrelevant content being considered. As we can see from the same table, the SBERT still stay strong in processing the sentences and thus achieve the best performance in unsupervised methods, but we can also see the drop of precision comparing to the **MADE 1d1s** dataset. The performance of BERT improves a little thanks to its general domain dictionary gathering more information from longer texts. Moving from short block to long block does introduce more resources which can be helpful in representing ADRs, but also makes BioBERT + average pooling representation difficult to tell the ADR information from the text, taking here the traditional "cls" token representation from last hidden layer showed us the best performance among the unsupervised methods. For CADEC dataset whose corpus contains more informal structures and spells and extremely unbalanced example distribution, the results seem less stunning as for the MADE data, but we can still observe that the strength of BioBERT in capturing features to represent a ADR correlated semantic content.

We have also explored taking not the context around entities but only masking the entities as input for BERT models and performed exactly the same pipeline as we did before, and it turns out that fully removing entities remains better with respect to all datasets, which lead us to the point that the content around entities did infers the information. Even more, we trained also LR classifier with the three BERT embeddings whose results grand us confidence that this kind of representation did grasp important information in distinguish ADR and non-ADR relations. Overall, the representation provided by BERT is a helpful representation as features for ADR-related block classification, during which the whole progress is fully unsupervised, which proves potential for more future explorations.

4.5 Ablation studies

Preprocessing of entities We choose the entities preprocess strategy through an ablation study on **MADE 1d1s** dataset and **MADE multi-d-s** dataset for their high quality of corpus in Table 3. For BioBERT, we take average pooling for **MADE 1d1s** dataset and extract "cls" token to represent blocks for **MADE multi-d-s** dataset. Both language models (*) are followed by KMedoids clustering. From the results we can the "Removing entities" strategy wins in aspect of high F1 score values in every cases, which supports our hypothesis, i.e. the contextual content around drug/symptom entities contains the information to characterize ADR.

Table 3. Ablation Preprocess

Model	Preporcess	MADE 1d1s			MADE multi-d-s		
		Prec	Recall	F1	Prec	Recall	F1
*BioBERT**	Kepp entities	0.581	0.404	0.477	0.583	0.604	0.593
	Replace by entity type	0.625	0.580	0.602	0.521	0.492	0.506
	Masking entities	0.690	0.605	0.645	0.620	0.585	0.602
	Remove entities	0.651	0.663	**0.657**	0.653	0.673	**0.663**
*SBERT**	Kepp entities	0.465	0.313	0.374	0.609	0.528	0.566
	Replace by entity type	0.733	0.615	0.669	0.638	0.574	0.604
	Masking entities	0.665	0.511	0.578	0.693	0.559	0.619
	Remove entities	0.681	0.685	**0.683**	0.666	0.593	**0.627**

Pooling strategy for BERT-based models We study the influence of different pooling strategy for the BERT-based language models to obtain a single vector that represent well the whole input sequence from a block. In particular we take into account two mainstream pooling strategies: calculate the average over the input tokens for each block, or extract the special "cls" token directly from model output as the summary for all tokens. We study performance for both BERT and BioBERT on **MADE 1d1s** dataset and **MADE multi-d-s** dataset, the preprocess of entities used here is "Removing entities" and the clustering method is KMedoids. Table 4 shows the results on two datasets. We observe that for short blocks in **MADE 1d1s** dataset, taking the average pooling aide in describing the ADR-related information while "cls" pooling are more suitable for representing long blocks in **MADE multi-d-s**.

4.6 Transfer to real EHR in French

We have also tried our method on data provided by Centre Hospitalier Universitaire de Nice (CHU-Nice) [5], where 41 random selected anonymized blocks

[5] https://www.chu-nice.fr/

Table 4. Ablation Pooling

Model	Pooling strategy	MADE 1d1s			MADE multi-d-s		
		Prec	Recall	F1	Prec	Recall	F1
BERT	cls	0.113	0.200	0.144	0.591	0.634	**0.612**
	avg	0.549	0.463	**0.502**	0.683	0.528	0.596
BioBERT	cls	0.625	0.623	0.624	0.653	0.673	**0.663**
	avg	0.651	0.663	**0.657**	0.604	0.634	0.619

from real EHRs obtained from the psychiatry department. Since these records are written in French while our base datasets are in English, translation from French to English was applied for all records by API from DeepL Translator[6] and we did the same processing as mentioned in section 3.2, including removing entities information from the text as input for language models and to applying average pooling for outputs of BERT/BioBERT, in order to obtain vector representation for each of the translated French block. We didn't utilise BERT built for French because we haven't had enough medical corpora to pretrain and to fine-tune such a BERT model into a BioBERT equivalent. We also didn't chose the multi-language BERT because the fact that multi-language BERT model's dictionary must contains multi-language sub-words, which still occupy resources for representing text considering the limited size for such dictionary. Besides, there is no milti-language Biomedical BERT model is available for us to represent the textual data. In order to get the best possible result, we make use of representation and clustering information got from experiments for **MADE 1d1s** dataset and applied to translated French block representations, whose results can be shown as in Table 5, where we can observe that BERT and BioBERT's representation support well even for the translated records. The reason of a lower performance for BioBERT may due to the different medical domains on which the training corpus of **MADE** data and the testing corpus are focused respectively. The former are derived only from cancer patient and the latter are from psychiatry domain, which introducing misunderstanding for BioBERT's representations. As for SBERT, the automatic translation may break the consistence of literature and lead to a misunderstanding to the texts given. We also note that in this experiment the test data is very small, which may cause insufficient information about the experiment. But still, we can see the potential of transferring what we can extracted by using our method from a known large corpora to test a translation engine processed data, and by choosing the training data with corpus in similar medical domains, the performance of BioBERT model could be promising.

5 Conclusion

Unsupervised learning can be a powerful resource in post-marketing pharmacovigilance, as it is able to leverage the large amount of data produced by daily

[6] https://www.deepl.com/translator

Table 5. Comparison with supervised baseline and our unsupervised approach for MADE-1d1s data applied to translated CHU data, we report the average *Precision*, *Recall* and F_1 scores of 5-fold cross validation. The results for unsupervised approaches (*) are always followed by a KMedoids clustering

Model	Category	MADE-1d1s to CHU		
		Prec	Recall	F1
BOW+LR	Supervised	0.628	0.696	0.660
$BERT^*$	Unsupervised	**0.564**	0.626	**0.593**
$BioBERT^*$	Unsupervised	0.524	**0.652**	0.581
$SBERT^*$	Unsupervised	0.350	0.225	0.279
BOW+Random	Supervised	0.450	0.391	0.419

trials on larger populations and avoid the significant cost of data annotation required to train a supervised model.

We have proposed a model that uses transformers (Bert or Siamese-Bert) to obtain a latent representation of the text after an easy to implement preprocessing. This is based on the removal of tokens related to the relation to be extracted. In this case: drugs and symptoms. The latent representation obtained by the transforms is sufficient to separate the text blocks into two classes (with or without the searched relation). The use of a transform model trained with a business vocabulary (BioBert in this case) further improves the homogeneity of the clusters produced.

We also performed a small experiment to test the possibility of applying the extracted representation and clustering in another language than the language of the training text by automatically translating the original texts (here French) into the language of the previously constructed model (English). To our great surprise, despite the rather large difference in style between medical reports written in French and those written in English, the results are very promising.

We still have to validate our approach to other types of relations, whether in the medical domain or in other domains that require to classify documents in relation to a given relation.

References

1. Bampa, M., Papapetrou, P., Hollmen, J.: A clustering framework for patient phenotyping with application to adverse drug events. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). p. 177–182 (Jul 2020). https://doi.org/10.1109/CBMS49503.2020.00041
2. Chapman, A.B., Peterson, K.S., Alba, P.R., DuVall, S.L., Patterson, O.V.: Detecting adverse drug events with rapidly trained classification models. Drug Safety **42**(1), 147–156 (Jan 2019). https://doi.org/10.1007/s40264-018-0763-y
3. Christopoulou, F., Tran, T.T., Sahu, S.K., Miwa, M., Ananiadou, S.: Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. Journal of the American Medical Informatics Association **27**(1), 39–46 (Jan 2020). https://doi.org/10.1093/jamia/ocz101

4. Dandala, B., Joopudi, V., Devarakonda, M.: Adverse drug events detection in clinical notes by jointly modeling entities and relations using neural networks. Drug Safety **42**(1), 135–146 (2019). https://doi.org/10.1007/s40264-018-0764-x
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs] (May 2019), http://arxiv.org/abs/1810.04805, arXiv: 1810.04805
6. El-allaly, E.d., Sarrouti, M., En-Nahnahi, N., Ouatik El Alaoui, S.: An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation. Journal of Biomedical Informatics **125**, 103968 (Jan 2022). https://doi.org/10.1016/j.jbi.2021.103968
7. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. ACM Transactions on Knowledge Discovery from Data **1**(1), 4 (Mar 2007). https://doi.org/10.1145/1217299.1217303
8. Henry, S., Buchan, K., Filannino, M., Stubbs, A., Uzuner, O.: 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association **27**(1), 3–12 (10 2019). https://doi.org/10.1093/jamia/ocz166, https://doi.org/10.1093/jamia/ocz166
9. Jagannatha, A., Liu, F., Liu, W., Yu, H.: Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug Safety **42**(1), 99–111 (2019). https://doi.org/10.1007/s40264-018-0762-z
10. Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E.M., Kors, J.A.: Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics **15**(1), 64 (Mar 2014). https://doi.org/10.1186/1471-2105-15-64
11. Karimi, S., Metke-Jimenez, A., Kemp, M., Wang, C.: Cadec: A corpus of adverse drug event annotations. Journal of Biomedical Informatics **55**, 73–81 (2015). https://doi.org/https://doi.org/10.1016/j.jbi.2015.03.010, https://www.sciencedirect.com/science/article/pii/S1532046415000532
12. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
13. Organization, W.H., et al.: The importance of pharmacovigilance. Safety monitoring of medicinal products p. 48 p. (2002)
14. Perez, A., Casillas, A., Gojenola, K.: Fully unsupervised low-dimensional representation of adverse drug reaction events through distributional semantics. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). p. 50–59. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), https://aclanthology.org/W16-5106
15. Reimers, N., Gurevych, I.: 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). pp. 3982–3992. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410
16. Uzuner, O., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**(5), 552–556 (06 2011). https://doi.org/10.1136/amiajnl-2011-000203, https://doi.org/10.1136/amiajnl-2011-000203
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017),

https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

18. Wei, Q., Ji, Z., Li, Z., Du, J., Wang, J., Xu, J., Xiang, Y., Tiryaki, F., Wu, S., Zhang, Y., Tao, C., Xu, H.: A study of deep learning approaches for medication and adverse drug event extraction from clinical text. Journal of the American Medical Informatics Association **27**(1), 13–21 (May 2019). https://doi.org/10.1093/jamia/ocz063, citation Key: 10.1093/jamia/ocz063 tex.eprint: https://academic.oup.com/jamia/article-pdf/27/1/13/34152042/ocz063.pdf
19. Weissenbacher, D., Sarker, A., Magge, A., Daughton, A., O'Connor, K., Paul, M.J., Gonzalez-Hernandez, G.: Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In: Proceedings of the Fourth Social Media Mining for Health Applications (SMM4H) Workshop & Shared Task. p. 21–30. Association for Computational Linguistics (Aug 2019). https://doi.org/10.18653/v1/W19-3203, https://aclanthology.org/W19-3203
20. WHO: Who—pharmacovigilance. https://www.who.int/teams/regulation-prequalification/regulation-and-safety/pharmacovigilance (2017)
21. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1753–1762. Association for Computational Linguistics, Lisbon, Portugal (Sep 2015). https://doi.org/10.18653/v1/D15-1203, https://aclanthology.org/D15-1203