Unsupervised Text Clusterisation to characterize Adverse Drug Reactions from hospitalization reports

Xuchun Zhang, Milou-Daniel Drici, Michel Riveill

4rd International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI' 2022), 19-21 October 2022, Corfu, Greece

Type of Presentation:
Oral: In-person: ☒
☐Poster: ☐ Virtual in Zoom: ☐
The same: ☐

Topic:
Machine Learning
Applied Artificial Intelligence

Unsupervised Text Clusterisation to characterize Adverse Drug Reactions from hospitalization reports

Xuchun ZHANG [1,2], Milou-Daniel DRICI [1,3], and Michel RIVEILL [1,2]
[1] Université Côte d'Azur, France
[2] CNRS, INRIA, France
[3] CHU Nice, France
E-mail: {xuchun.zhang, michel.riveill}@inria.fr, 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." [1], 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 involve 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.

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)[1]. 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 report important ADRs [2]. 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 number of clinical reports at disposition.

Deep learning has boosted the development of Natural Language Processing (NLP) and showed that

[1] https://ansm.sante.fr/

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 [3, 4, 5] 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.

2. Related work

Because of the rarity of annotated EHRs related to adverse events and the limited public access to clinical records, given patient privacy and confidentiality. Since the first approach which try to characterize the likelihood of a candidate drug-symptom relation to be categorized as a true ADR [6], this domain has get more and more attention from NLP research community. The 2018 National NLP Clinical Challenges shared task (n2c2) [7] provided 505 discharge summaries for 3 different tasks: concept extraction, relation classification, and end-to-end systems construction, where among the best performance systems, Wei et al. [8] 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. [9] 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 piecewise neural networks [10] and an attention-based BiLSTM. More recently, El-allay [11] 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 With similar tracks as 2018 n2c2 shared task, the MADE (Medications and Adverse Drug Events from Electronic Health Records) 1.0 challenge [4] provides real de-identified EHRs and corresponding annotations for medications, symptoms and ADRs. For MADE data, Chapman et al. [12] 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. [13] 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.

In recent years, the NLP community has demonstrated the great power of supervised machine learning techniques for ADR extraction. However, the unsupervised approaches still remain 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. [15] explored encoding the document type without considering too much the textual content and by clustering aggregation [16] techniques to grasp information about the phenotype of patient/document, which provided decent cluster structure for ADR analysis.

3. Method

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. 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 most 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.

3.2. Unsupervised BERT based ADR block detection

The pre-trained language models, including BERT (Bidirectional Encoder Representation from Transformer) [17], a two-stage Transformer-based [18] 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) [19] 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.

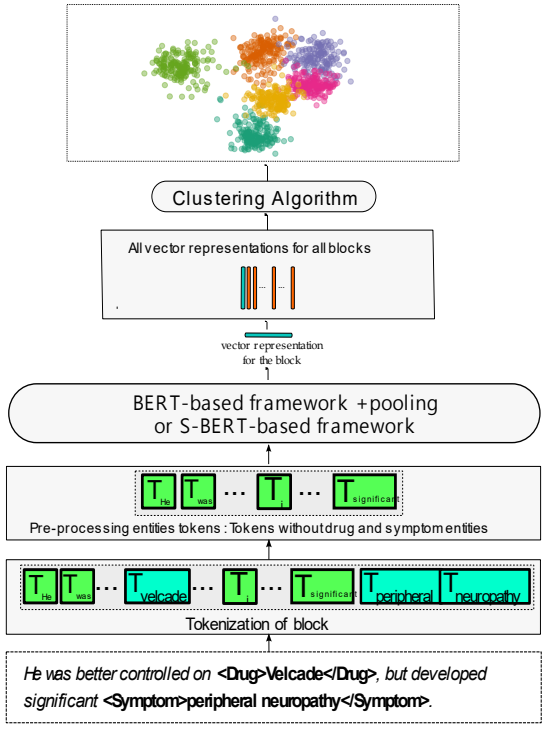

Fig. 1. The structure of our model. For each block, only the contextual tokens around drug/symptom named entities are selected for BERT-related embedding. For BERT-like models we applied a pooling strategy between tokens from the last layer to obtain a single vector representation for the block as the 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.

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 token 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 labelled 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.

4. Experimental results

4.1. Dataset

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 also means 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. We used here two datasets: MADE [4] dataset. and CSIRO Adverse Drug Event Corpus (CADEC) [20] dataset. The former is the data used in MADE 1.0 challenge [4], whose corpora collected from 21 randomly selected cancer patients at the University of Massachusetts Memorial Medical Center, with the annotations of drugs, symptoms and ADEs. We choose this dataset considering the nature of source being real life clinical notes and its high quality of annotations.

For MADE data, we sampled two sets of blocks that contains both drug and 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:** All examples that contains only drugs or only symptoms were removed and thus we got a dataset where each block has at least one drug and one symptom. This dataset contains long blocks from EHR corpus with a nearly balanced distribution with 571 negative blocks and 651 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 (with a number of 416 negative blocks and 301 positive ones) dataset with short blocks from well-written EHR corpus.

The CADEC Dataset has a rich annotated corpus of medical forum posts "Ask a Patient", which is dedicated to ADE-related consumer reviews on medications with length of several sentences. These posts are mostly written in colloquial language and often deviate from the formal rules of English grammar and punctuation. The annotations contain entities such as drugs, ADEs, symptoms and diseases related to their respective concepts in 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 symptom

4.2. Experimental Settings

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.

As evaluation metrics, we chose then precision, recall and F1-Score as categories of metric to evaluate the result produced. The fraction of documents retrieved that are relevant to the ADE, is known as precision, which can be given by the formula $\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$, where TP and FP represent the number of real positive examples and 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 $\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$, 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 $F_1 = 2 \times precision \times recall/(precision + recall)$. We computed the average score of a 5-fold cross validation as the final score for each term.

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 [21] model, who uses the same structure as the BERT model pre-trained 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 experiments the model "biobert-base-cased-v1.1". And 3) the Sentence-BERT (S-Bert) [19] "sentence-transformers/bert-base-nli-mean-tokens". To get fully representation for whole block for the first two models 1) and 2), as well as obtaining 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, to agree with the ADE-related/non-ADE blocks.

4.3. Experimental Results

The experimental results are shown in the Table 1, from which 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

Table 1. Comparison with supervised baseline and our unsupervised approach, we report the average Precision, Recall and F1 scores of 5-folds cross validation. The results for unsupervised approaches (*) are always followed by a KMedoids clustering

Category	Exps	MADE 1d1s			MADE multi-d-s			CADEC		
		Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
Supervised Classifier	BOW+LR	0.702	0.797	0.746	0.809	0.847	0.828	0.939	0.993	0.965
Unsupervised Clustering	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
	S-BERT*	0.733	0.615	0.669	0.666	0.593	0.627	0.958	0.492	0.650
Supervised Classifier	BOW+Random	0.514	0.529	0.522	0.509	0.440	0.472	0.946	0.509	0.662

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 BioBERT, the S-Bert 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 Table 2: Comparison with supervised baseline and our unsupervised approach, we report the average Precision, Recall and F1 scores of 5-folds cross validation

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 S-BERT 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 with 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 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.

5. Conclusion

Unsupervised learning can be a powerful resource in post-marketing pharmacovigilance, as it can exploit the big amount of data produced by daily trials of a larger populations and avoiding simultaneously the big cost of annotating data. We proposed a model to make use of modern text features extraction technique with BERT based models and explored the possibility of clustering ADR-related representations together in semantic space. The results indicate that with only contextual tokens as input, the model representation, especially those who obtained from domain-specific pretrained model like BioBERT, can be helpful in classifying ADR-related textual blocks with non-ADR blocks, especially for corpus like EHRs.

References

[1] Institute of Medicine (US) Committee on Quality of Health Care in America, To Err is Human: Building a Safer Health System, L. T. Kohn, J. M. Corrigan and M. S. Donaldson, Eds., Washington (DC): National Academies Press (US), 2000.

[2] W. H. Organization and others, "The importance of pharmacovigilance," *Safety monitoring of medicinal products*, p. 48 p., 2002.

[3] A. Jagannatha, F. Liu, W. Liu and H. Yu, "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*, vol. 42, p. 99–111, 2019.

[4] A. Jagannatha, F. Liu, W. Liu and H. Yu, "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*, vol. 42, p. 99–111, 2019.

[5] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. J. Paul and G. Gonzalez-Hernandez, "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*, 2019.

[6] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen and J. A. Kors, "Knowledge-based extraction of adverse drug events from biomedical text," *BMC Bioinformatics*, vol. 15, p. 64, March 2014.

[7] S. Henry, K. Buchan, M. Filannino, A. Stubbs and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association*, vol. 27, pp. 3-12, October 2019.

[8] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, C. Tao and H. Xu, "A study of deep learning approaches for medication and adverse drug event extraction from clinical text," *Journal of the American Medical Informatics Association*, vol. 27, p. 13–21, May 2019.

[9] F. Christopoulou, T. T. Tran, S. K. Sahu, M. Miwa and S. Ananiadou, "Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods," *Journal of the American Medical Informatics Association*, vol. 27, p. 39–46, January 2020.

[10] D. Zeng, K. Liu, Y. Chen and J. Zhao, "Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, 2015.

[11] E.-d. El-allaly, M. Sarrouti, N. En-Nahnahi and S. Ouatik El Alaoui, "An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation," *Journal of Biomedical Informatics*, vol. 125, p. 103968, January 2022.

[12] A. B. Chapman, K. S. Peterson, P. R. Alba, S. L. DuVall and O. V. Patterson, "Detecting Adverse Drug Events with Rapidly Trained Classification Models," *Drug Safety*, vol. 42, p. 147–156, January 2019.

[13] B. Dandala, V. Joopudi and M. Devarakonda, "Adverse Drug Events Detection in Clinical Notes by Jointly Modeling Entities and Relations Using Neural Networks," *Drug Safety*, vol. 42, p. 135–146, 2019.

[14] A. Perez, A. Casillas and K. Gojenola, "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)*, Osaka, 2016.

[15] M. Bampa, P. Papapetrou and J. Hollmen, "A Clustering Framework for Patient Phenotyping with Application to Adverse Drug Events," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020.

[16] A. Gionis, H. Mannila and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, p. 4, March 2007.

[17] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017.

[19] N. Reimers and I. Gurevych, "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)*, Hong, 2019.

[20] S. Karimi, A. Metke-Jimenez, M. Kemp and C. Wang, "Cadec: A corpus of adverse drug event annotations," *Journal of Biomedical Informatics*, vol. 55, p. 73–81, 2015.

[21] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, p. 1234–1240, 2020.