Public Health and Informatics J. Mantas et al. (Eds.) © 2021 European Federation for Medical Informatics (EFMI) and IOS Press. This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0). doi:10.3233/SHTI210368

Visualising Patterns Associated with Adverse Drug Reactions in French Forums

Nour ALLAM^{a,b}, Bissan AUDEH^a, Marie-Christine JAULENT^a and Cedric BOUSQUET^{a,1}

^aSorbonne Université, INSERM, Université Paris 13, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, Paris, France ^bESIEE Paris graduate school of engineering, Noisy-le-Grand, France

Abstract. As social media are an interesting source of information for pharmacovigilance, we implemented a novel visualisation method for pharmacovigilance specialists applied to French discussion forums. A word embedding model was trained on posts to facilitate the identification of patterns associated with adverse drug reactions.

Keywords. Natural Language Processing, Pharmacovigilance, Word Embedding, Machine Learning, Drug Surveillance

1. Introduction

Nowadays, it has become frequent for patients to describe adverse drug reactions in discussion forums. As a result, an immense amount of posts might contain precious information for pharmacovigilance. Reviewing such voluminous data by pharmacovigilance specialists is almost impossible without the use of natural language processing and machine learning techniques. Before applying these techniques, detecting and visualizing patterns in data is a major step in exploratory data analysis. This step allows better understanding of data which leads to better strategy for further knowledge extraction and interpretation. We propose an unsupervised learning approach for visualizing patterns associated with adverse drug reactions. Our visualisation is based on the output of word2vec [1], a word embedding language model that regroups tokens that co-occur in similar contexts. We used a sample from a dataset consisting of co-occurences between drugs and medical conditions previously annotated using a supervised approach [2]. Then we performed 2D representations of word embeddings of these annotations to figure if it is possible to highlight relevant patterns associated with adverse drug reactions.

2. Methods and Results

A total of 665.662 messages were extracted from the medical discussion forum Doctissimo² from January 2020 to June 2020. After preprocessing and tokenizing the

¹ Corresponding Author, Cédric Bousquet; E-mail: cedric.bousquet@chu-st-etienne.fr.

² French medical discussion forum <www.doctissimo.fr>.

posts, we applied a bigram function to merge words commonly used together. Then, we trained a Skip-gram model [1] using Python 3 with the Gensim package. After running a grid search for hyper parameters, we selected an embedding size of 300 dimensions with windows of 7 tokens, a minimum count of 7 tokens, a negative sampling [3] of 10 and 7 iterations. The remaining parameters were set to defaults. A t-SNE [4] visualisation allows us to see clusters of unigrams/bigrams that occur in the same context. For example, Figure 1 shows two clusters associated with two French medicinal products: Doliprane® (paracetamol) and Loxapac® (loxapine).

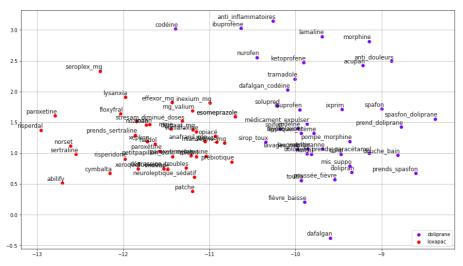


Figure 1. T-SNE visualization of the 40th nearest unigrams/bigrams to 'doliprane' and 'loxapac'

3. Discussion and Conclusion

These preliminary results show us that using word embeddings with medical forums could be a promising approach for pharmacovigilance. However, performing additions and subtractions on word embeddings was disappointing because some linear relations between the word pairs ADR1:Drug1 and ADR2:Drug2 were not consistent. We believe that further research is necessary to figure in which settings such approach is reliable to infer knowledge with the proposed visual representation.

References

- Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013 Sept 7.
- [2] Morlane-Hondère F, Grouin C, Zweigenbaum P. Identification of drug-related medical conditions in social media. In: Proceedings of the international conference on language resources and evaluation (LREC'2016), Portorož, 2016 May 23–28.
- [3] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality, 2013 Oct 16.
- [4] Van Der Maaten L, Hinton G. Visualizing data using t-SNE, J. Mach. Learn. Res. 9. 2008 Nov.