Repositioning evaluated drugs with textual reviews, giving a new enhanced distribution of drugs (REDRUGEDD)

Group 17: Hannah Stone 285434, Jasmine Tantituvanont 485731, Sarah Busch 437214, Nika Kovacic 660742, Ina Klaric 793890
Cognitive Science and Artificial Intelligence
Tilburg University

Abstract

Drug repositioning concerns the process of allocating an existing drug to a new therapeutic use and has gained significant interest due to its potential in reducing costs and expediting the drug development timeline. In this paper, we propose a semantic-based approach using the SBERT network in order to generate sentence embeddings to spatially determine drug similarities through perceived drug effects. The results of this study has the prospective possibility of contributing to the field of computational methods for drug discovery and repurposing.

Keywords: drug repositioning; natural language processing; semantic text analysis; sentence embeddings; user reviews; computational methods

Introduction

Drug repositioning is an area of great interest due to its potential in reducing the cost and time involved in bringing new treatments to market (Tobinick, 2009). Also known as drug reprofiling, repurposing or redirecting, this field is concerned with the discovery of alternative therapeutic uses for already existing and approved drugs. Typical drug repositioning strategies are split into drug-based or disease-based with drug-based focused on the mode of action of a drug versus disease-based being focused on side effects and disease indication similarities (Jarada et al., 2020). Computational methods for the purpose of the discovery of drug repositioning candidates is a modern but increasingly popular technique due to the rise of biological big data and bioinformatics knowledge. The most commonly used computational methods include: network-based approaches, text mining-based approaches and semantic-based approaches (Xue et al., 2018). Our proposed method for the discovery of drug repositioning candidates is a semantic-based approach utilising sentence embeddings to spatially determine similarities between drugs in the context of online user reviews. The user reviews collected from Drugs.com will serve as the semantic representation of each drug. To create the sentence embeddings for analysis, we have chosen to use the Sentence-BERT (SBERT) network which is a modified BERT network which has an improved runtime efficiency while maintaining accuracy in creating high-dimensional sentence embeddings (Reimers & Gurevych, 2019). Initially, a different model, word2vec, was proposed to be used in order to generate the embeddings. The reason why SBERT was chosen over word2vec was because of the need to generate embeddings that were representative of larger pieces of text. Word2vec specialises in generating word embeddings whereas SBERT is specifically developed for the purpose of generating sentence or small text size embeddings (Mikolov et al., 2013). As we wanted an overall representation of the semantics of all the reviews for a particular drug, it was determined that SBERT was a more appropriate model of choice. For our experiment, we will be using a pre-trained SBERT network which has been trained on over 1 billion training pairs. This will ensure that the model already has a wide range of semantic understanding. The SBERT pre-trained models have high generalisability so it has been determined that it is more advantageous to use a pre-trained model versus training one ourselves. For our analysis, we will create visual plots of how the generated sentence embeddings are spatially distributed and cluster groups of drugs using the Hierarchical Density-Based Spatial Clustering of Applications with Noise method (HDBSCAN). This is our chosen clustering method for several reasons. Firstly, it is able to determine clusters of varying densities so it can find clusters of varying shapes and sizes despite potential noise in the data. Also, it can determine the number of clusters appropriate for the data itself based on factors such as data density (McInnes et al., 2017). Plotting the produced review embeddings and its clusters should provide insight into the effectiveness of the use of SBERT as a method of determining drug similarity based on user reviews.

Methods

Our drug repositioning technique of semantic text analysis utilises a machine learning algorithm to find similarities between the perceived effects of drugs. Previous attempts have been made to utilise online user review data to gain further insight into the wider consequences of various drugs, such as side effects and public health monitoring (Gräßer et al., 2018). For the purpose of our experiment, the dataset used has been acquired from Gräßer et al.'s work titled 'Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning' which consists of over 200,000 reviews from Drugs.com. Reviews were concatenated and grouped together by drug name so that all reviews corresponding to a particular drug formed one large review. This information was placed in a dataframe which resulted in 2865 unique drug-review entries. The review data was converted into lowercase letters to reduce variation and decrease the vocabulary size for efficiency. Further pre-processing of the text data such as punctuation removal or lemmatization was not necessary due to the nature of the SentenceTransformers framework being created to handle this level of data variation so that a better semantic representation of each sentence could be captured. To generate the vector representation for each review, the SBERT network accessed through the SentenceTransformers framework was used. The model requires sentences as input and has a word limit of 128 words per sentence. As a result, each review was split into separate sentences using regular expressions. Due to the reviews of each drug being formed into one large review, this can result in very large strings representing one drug. Therefore, one drug may have over 100 sentences representing it. In order to generate one embedding that represents a single drug, an average of all the sentence embeddings was calculated. This resulted in a single 768-dimension vector for each drug. At this point, the semantic similarity between two drugs can be found using the cosine similarity function.

Table 1: Cosine similarities of random drug pairs

Drug 1	Drug 2	Cosine similarity
Sumatriptan	Meclizine	0.8519
Finacea	Avandia	0.5313
Conestat alfa	Acetaminophen	0.3231
Lysine	Jalyn	0.6152

For the purpose of visualisation and clustering, dimensionality reduction was performed on the resultant vectors. The t-distributed stochastic neighbour embedding method (t-SNE) was used to produce 2-dimensional vectors from the large 768-dimensional embeddings.

Results

To discover potential drug clusters, HDBSCAN was used. The results of this were plotted on a scatter graph with each cluster represented by a different colour. Two randomly selected drugs from each cluster are also annotated.

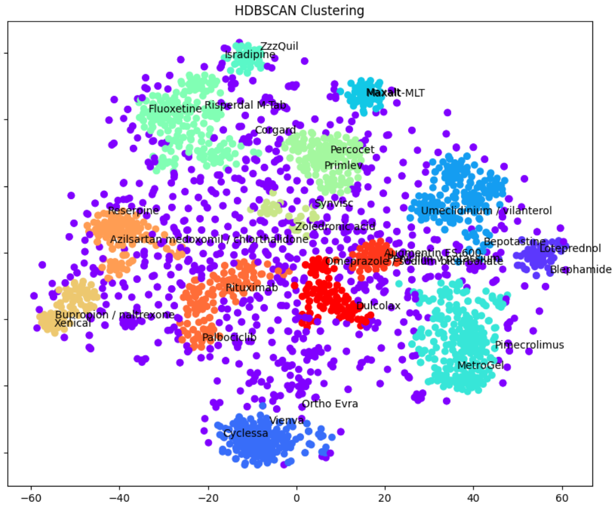

Figure 1: Scatter plot of drug embeddings with clusters differentiated by colour

Table 2: A random sample of drugs contained in a random sample of clusters

Cluster	Drugs
1	Alesse
	Aygestin
	Lessina
2	Sodium oxybate
	Silenor
	Vanatrip
3	Zipsor
	Ibuprofen
	Relafen

As our experiment utilises an unsupervised approach, manual observation was carried out.

Table 2 shows three random drugs contained within a sample of three clusters. The formal accuracy of the semantic analysis cannot be determined without a dataset containing drugs labelled with existing known secondary purposes. However, it is notable that out of the drugs shown in Table 2, Cluster 1 contains drugs which are all female contraceptives, Cluster 2 contains drugs which all act on the nervous system and Cluster 3 drugs are all anti-inflammatories.

Discussion

As demonstrated in Figure 1 and Table 2, our semantic analysis approach has demonstrated that it was able to group drugs with a similar purpose spatially closer in comparison to drugs with very dissimilar purposes and drug action. The implementation for this experiment has several areas in which it could be improved or adjusted. For example, the SBERT network used to generate the sentence embeddings was based on a generic pre-trained model. This could be fine-tuned using more domain specific training data which could potentially create more representative sentence embeddings (Hao et al., 2020). Additionally, the HDBSCAN method could be fine-tuned in order to determine what is the most optimal clustering size by using a cost function involving the cluster membership scores of each data point.

Conclusion

Averaged high-dimension vector representations of drug reviews has given an indication that the use of user drug review data could allow for an improved approach in the discovery of drug repositioning candidates. The potential use for this technique would be to allow researchers to find what drugs are classified to be the most similar to the one of interest. This would narrow the field of investigation and increase the speed at which repurposed drugs are found. From only a surface overview of the resulting embeddings, it is visible that this method has some degree of accuracy in determining drug similarities which could lead to observations into possible overlaps in therapeutic uses for certain drugs.

References

Tobinick, E. L. (2009). The value of drug repositioning in the current pharmaceutical market. *Drug News & Perspectives*, *22*(2), 119. https://doi.org/10.1358/dnp.2009.22.2.1303818

Jarada, T. N., Rokne, J. G., & Alhajj, R. (2020). A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics*, *12*(1). https://doi.org/10.1186/s13321-020-00450-7

Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of Drug Repositioning Approaches and Resources. *International Journal of Biological Sciences*, *14*(10), 1232–1244. https://doi.org/10.7150/ijbs.24612

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1908.10084

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *arXiv (Cornell University)*. Cornell University. https://arxiv.org/pdf/1301.3781

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). *Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning*. https://doi.org/10.1145/3194658.3194677

McInnes, L., Healy, J. J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

Hao, Y., Dong, L., Wei, F., & Xu, K. (2020). Investigating Learning Dynamics of BERT Fine-Tuning. In *International Joint Conference on Natural Language Processing* (pp. 87–92). https://www.aclweb.org/anthology/2020.aacl-main.11.pdf

Contribution Statement

Hannah Stone: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization

Jasmine Tantituvanont: Conceptualization, Investigation

Sarah Busch: Conceptualization, Investigation

Nika Kovacic: Conceptualization, Investigation

Ina Klaric: Conceptualization, Investigation

Statement of Technology

The dataset was acquired from Gräßer et al., 2018's paper titled 'Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning'. The data contains publicly available textual reviews from both registered and anonymous users. The dataset is permitted for research use. All figures were created ourselves. Figure 1 was created using matplotlib in Python.

All code was written ourselves. All writing was written ourselves with no paraphrasing tools. Microsoft Word was used to format the paper. Scribbr.com was used as a reference manager while working.