Comparative analysis of information-theory-based statistical methods and transformer-based machine learning techniques for scientific literature classification

Arsentii Ivasiuk[3,4 +], Ihor Stepanov[1,4 +], Stanislav Zubenko[1 +], Alina Frolova[1,2 *]

[1]Institute of Molecular Biology and Genetics of NASU, Kyiv, Ukraine; [2]Kyiv Academic University, Kyiv, Ukraine; [3]Bogomoletz Institute of Physiology, Kyiv, Ukraine; [4]Taras Shevchenko National University of Kyiv, Ukraine

*Corresponding author: fshodan@gmail.com
+Authors equally contributed to work

Introduction

Production of scientific information is growing exponentially. One of the first studies regarding scientific literature production was conducted by De Solla Price, who used publication data collected over the 100-year period 1862–1961 to calculate a doubling time. The results showed 13.5 years for doubling the scientific corpus with 5.1% annual growth rate (de Solla Price, 1965). Development of information technologies created conditions for acceleration of scientific production rate, which made scientific information more accessible, but also introduced new challenges. It becomes harder to analyze scientific literature not only for individuals, but for large organizations.

Our research is focused on the biomedical domain, which is one of the largest and most rapidly developing. Accessibility of biomedical literature through databases such as Medline and research activity in biomedicine are useful for practical implementation of natural language processing (NLP) techniques. In total Medline has more than 27M citations, and more than 800 000 articles are added to this database every year (Medline, 2021).

We implemented an information theory statistical approach and compared it with modern transformers on relevant practical tasks – classification of biomedical papers related to Drug-Induced Liver Injury (DILI) as part of the CAMDA 2021 Challenge 2. It is a clinically significant condition and is one of the reasons for registration of many potential drug candidates failures. Scientific papers are one of the main sources of information related to DILI, thus collecting and processing huge amounts of biomedical literature can help pharma companies, research organizations and regulators find information that can help future drug development and regulation. In the past the potential of statistical approaches were limited due to lack of data, but today we can study them more thoroughly and compare them to novel deep learning approaches.

Data and methods

Datasets

The positive reference data set comprises ~14,000 DILI related papers referenced in the NIH LiverTox database, which has been validated by a panel of DILI experts. This is complemented by a realistic, non-trivial negative reference set (provided by CAMDA Challenge) of ~14,000 papers that is highly enriched in manuscripts that are not relevant to DILI but where obvious negatives and any positives we could identify have been removed by filtering for keywords and through well established language models, followed by a selective manual review by DILI experts at the FDA. Training set contains approximately equal numbers of examples in the positive and negative groups, the same is true for the validation data set.

We also used several additional datasets for more robust benchmarking of the methods. As a negative reference dataset we used IMDB reviews (Kaggle, 2021). This is a dataset for binary sentiment classification, it consists of 25,000 positive class reviews and the same amount for negative class reviews. We also prepared the dataset of 10812 entries from the Medline database using keyword filtering to select articles related to longevity research.

We tried augmentation of data using the LiverTox database references, but to keep the training dataset balanced we used Hallmarks Of Cancer corpus (Baker at al., 2017) of 1852 PubMed publication abstracts. We filtered the LiverTox database to keep only articles that are not present in the DILI initial or validation dataset (~2,000 articles remained).

Pointwise mutual information/term frequency-inverse document frequency method

As our first approach to document classification we used a combination of two statistical methods: pointwise mutual information (Bouma, 2009) and term frequency-inverse document frequency (Havrlant et al., 2017). For each example in the dataset we concatenated Title and Abstract into one block of text, then we calculated pointwise mutual information score (pmi) for this block of text:

$$pmi(class, word) = log\left(\frac{p(class, word)}{p(class) \cdot p(word)}\right)$$

$$npmi(class, word) = -\frac{pmi(class, word)}{log(p(class, word))},$$

Where p(class) - is the class probability in the dataset, p(word) - word frequency in the dataset, p(class, word) - word frequency in the examples with the class label. npmi is the normalized pmi. We interpret npmi(class, word) as the importance of a word for the class. Then we calculated term frequency/inverse-document frequency:

$$tf = 1/n,\ idf = log\left(\frac{N}{N_{word}}\right),\ tf - idf = n_{word} \cdot tf \cdot idf$$

Where n – is the number of words in the given text, N – is the number of documents, N_{word} – is how many documents a given word is present in (N_{word} <= N), n_{word} – occurrences of the word in the given text.

tf-idf measures the importance of a word for the text. The value of tf-idf is higher when a word occurs in a few documents and a given text is small, which means that the word is most crucial for determining text meaning in the context of all texts.

pmi measures the importance of a word for the class. pmi will be highest when probability of the class is low and frequency of the word in the whole dataset is small and frequency of the word in the class is high, then a given word is a very important signature for the class.

Based on the above heuristics we decided to measure the importance of the word by multiplying pmi by tf-idf. Words with very low pmis (very low frequencies in general) will not contribute significantly to our decision, but tf-idf is very high for such a word and by multiplying pmi by tf-idf we can give more significance for the word.

In order to make a decision about the class we calculated sum of pmi*tf-idf for all words in the text. First we calculated a score for pmis with the first class, then with the second. class was predicted based on the highest sum.

$$F(C) = \left(\sum_{i=1}^{n} (tf - idf)_{word} \cdot npmi(C, word_i)) \right)$$

Where C – is one of the classes. Predicted class then defined as:

$$C_{pred} = argmax(F(C))$$

For division of text into words we used tokenization from spaCy python library, we filtered stop words, chose words that consisted of alphabetical characters only and translated them to lowercase.

For running the implementation of the described information theory statistical-based model we used a computer with following parameters: OS - Pop!_OS 20.04 LTS x86_64, CPU - Intel i5-9300H (8) @2.4GHz (the code was run on one core). Submissions for the CAMDA challenge were sent with the following email – igor.stepanov2000@gmail.com.

Transfer learning transformer-based SciBERT model

As a competitive solution for our classification NLP task we used a state-of-the-art approach – transfer learning general purpose model. As a base model we used SciBERT (Beltagy et al., 2019). SciBERT is a pre-trained language model for the science domain. We decided not to use BioBERT (Lee at al., 2020) because even though it is trained on biomedical data specifically, it has a major disadvantage of using original BERT vocabulary, which is very limited in the amount of chemical substances specific terms.

We concatenated article titles and abstracts, but since BERT architecture has a limit of 512 tokens, we truncated sequences longer than this limit and padded sequences that are shorter with pad token. No additional data pre-processing was done.

For training we tuned several parameters: number of training epochs, batch sizes and gradient accumulations steps. Best result was achieved with 3 epochs, batch size of 8 and gradient accumulation steps of 4. To speed up training we used build-in optimizations of the huggingface library to convert parts of calculations to FP16.

For running the SciBERT model we used a computer with the following parameters: OS - Ubuntu 18.0, GPU - Tesla K80 GPU, CPU - 2 X CPU Intel(R) Xeon(R) CPU @ 2.30GHz cores. Submissions were sent with the following email - dantistnfs@gmail.com.

Source code availability:

- https://colab.research.google.com/drive/133FITGZRWbijECy1_JMKSrVmUwJwZDH1
- https://github.com/platycristate/ptah

Results

Table 1. Methods performance on the initial and validation sets.

	pmi/tf-idf	SciBert	pmi/tf-idf	SciBert
	Initial set		Validation set	
Accuracy	0.9419	0.9731	0.9412	0.9740
F1 score	0.9438	0.9728	0.9424	0.9743
Recall	0.9519	0.9762	0.9325	0.9730
Precision	0.9358	0.9693	0.9526	0.9755
FP rate	0.065	0.0152	-	-
FN rate	0.05	0.0117	-	-
Training time (min)	0.243	41	-	-
Classification time (min)	0.012	10	0.063	4

From **Table 1** we can see that the MI model has recall larger than precision (for initial set). In order to understand this we analyzed incorrectly classified examples. Some of these examples from the negative reference have words like “hepatitis” in the titles and these words are more frequent in the positive reference. It leads to lower tf-idf scores for the positive class and larger for the negative.

We checked how our model performs with and without tf-idf score. When we used only pmi scores, the accuracy on a validation dataset was 0.9294, F1 score 0.9321, recall 0.9067, and precision 0.9589. If we compare these results with the results of the model that used a conjunction of pmi and tf-idf scores (Table 1), we can see a significant drop in recall, because our model misses lots of positive examples due to the fact that words that are very important for the positive class are, at the same time, very rare, for example, names of diseases related to liver injury. And tf-idf score can increase the significance of rare words. In the situation when pmi and term frequency was used without an idf score we have got the following results on validation dataset: accuracy 0.9267, F1 score 0.9306, recall 0.8910 and precision 0.9739. Because words that are very rare and are important for the DILI class will not have a significant influence on the decision about class, therefore it becomes generally more difficult to classify an example as a positive class and it leads to higher precision. If we compare this situation with previous, words which have large influence on choosing DILI class and have high frequency will be dominant in choosing DILI class, but this situation is rare, so we will have lower recall, but higher precision. In conclusion, we can say that our model, which utilizes pmi/tf-idf, has balanced recall and precision. The model finds more positive examples by considering the importance of words (tf-idf).

We also tested our statistical-information theory based model on the IMDB review dataset with 80% of the data as a training set. We got 84% accuracy and 83% precision, which is similar to results of classical machine learning algorithms and algorithms that are based on word embeddings (GloVe, Word2Vec) but our model was behind deep learning approaches that use transformers.

In order to simulate a real situation with a high percentage of negative classes we tested our model on the dataset consisting of articles related to longevity research, where our model classified only 2.8% of articles as related to DILI.

After augmenting our starting data with filtered LiverTox database and Hallmarks of Cancer corpus we tested methods again, but got slightly lower scores than before. We think this happened because the LiverTox database does not directly correspond to DILI positive reference articles. However, we also think this could be due to mistakes in the initial and validation datasets, which we briefly highlight in the Discussion section.

Discussion

The main problem with our statistical method is that it requires lots of examples that are needed for creation of the pmi/tf-idf dictionary. We can overcome this problem by extending the training dataset or by trying to predict pmi/tf-idf score for the word based on the examples with the calculated pmi/tf-idf scores, for example, by training a feedforward neural network that takes as an input word embedding of the word and returns pmi/tf-idf score for all classes. By doing so we can try to build a link between semantics of the word and its dependency on some class.

Another idea about improvement of our model is a combination of semantics and pmi/tf-idf. We tried to implement this idea using a feedforward neural network as our classifier; as an input we used a text embedding vector concatenated with pmi/tf-idf scores for different classes, text embedding is the sum of word embeddings. This model shows better results than the statistical method only if most of the words from the dataset are present in the pmi/tf-idf dictionary. When we added words from the test dataset to our dictionary, we were able to get 97.5% accuracy. Further investigation is needed.

Interestingly, while investigating FP and FN cases in the initial dataset we found that there are a number of articles from the positive set that were strongly identified as negative cases by both methods. Many of such cases indicated errors in positive set with articles completely not related to DILI (for example, PMID: 460061, “Quantum fluctuations in radiographic screen-film systems”). Same was applicable to the negative reference set. We believe we are able to achieve an even better score once the database is corrected. The cases where methods were unsure about the class of the article showed us that it is extremely difficult to differentiate between a true DILI article and applied research, where a particular drug was used on the patient (e.g.cancer treatment).

We would also like to point out potential problems with the datasets provided by CAMDA challenge. First, both initial and validations sets contained the same amount of positive and negative classes, which hardly represent real situations. We tried to overcome this limitation by introducing new datasets, but further extension is needed. In addition, we noticed that only positive reference contained articles with titles only and based on our investigation it is also applicable to the validation set, so neural network based methods might simply learn to classify absence of abstract as DILI article.

We propose to use our statistical method as a first approach to text classification because our method is very fast, easily interpretable and can be scaled by extending the training dataset. Compared with models such as SciBERT that require computational and informational resources for pretraining, statistical approach does not require any additional data sources and pre-training operations. Moreover, most text classification tasks do not need very high accuracy and complicated language structure analysis. Our method is very useful for dataset analysis because we can estimate the word distribution among classes and their influence in the text. This information can be used for feature selection by more complicated models such as neural network classifiers that utilize word embeddings. In some cases, however, when the task is multiclass classification or when classification requires a deeper understanding of abstract concepts, transformers, convolutional, recurrent NNs should be used because they can perform detailed analysis of the sentence structure.

Conclusion

Our statistical information-theory based method is a reliable approach that shows robust performance with balanced precision and recall. This method can be further improved by adding prediction of pmi/tf-idf scores for the words and by using word embeddings together with pmi/tf-idf score for the prediction of the class using vanilla feedforward neural networks. This statistical method is very useful for topical classification. When more intricate analysis of the text is required, one should use more sophisticated methods.

References

- de Solla Price, D.J., 1965. Is technology historically independent of science? A study in statistical historiography. *Technology and Culture*, pp.553-568.
- Bouma, Gerlof. "Normalized (pointwise) mutual information in collocation extraction." *Proceedings of GSCL* (2009): 31-40.
- Havrlant, Lukáš, and Vladik Kreinovich. "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)." *International Journal of General Systems* 46.1 (2017): 27-36.
- https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
- Beltagy, Iz, et al. “SciBERT: A Pretrained Language Model for Scientific Text.” ArXiv:1903.10676 [Cs], Sept. 2019. arXiv.org, http://arxiv.org/abs/1903.10676.
- Baker, Simon, et al. "Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer." *Bioinformatics* 33.24 (2017): 3973-3981.
- Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics* 36.4 (2020): 1234-1240.
- Medline: https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
- Kaggle IMDB dataset: https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews