Content-based subject classification at article level in biomedical context

Eric Jeangirard

Content-based subject classification at article level in biomedical context

Eric Jeangirard[1]

[1]French Ministry of Higher Education, Research and Innovation, Paris, France

April 2021

Abstract

Subject classification is an important task to analyze scholarly publications. In general, mainly two kinds of approaches are used: classification at a journal level and classification at the article level. We propose a mixed approach, leveraging on embeddings technique in NLP to train classifiers with article metadata (title, abstract, keywords in particular) labelled with the journal-level classification FoR (Fields of Research) and then apply these classifiers at the article level. We use this approach in the context of biomedical publications using metadata from Pubmed. Fasttext classifiers are trained with FoR codes and used to classify publications based on their available metadata. Results show that using a stratification sampling strategy for training help reduce the bias due to unbalanced field distribution. An implementation of the method is proposed on the repository https://github.com/dataesr/scientific_tagger

Keywords: open science, subject classification, fasttext, word embeddings, fields of research

1. Introduction

Classifying scholarly literature in subjects or fields can have multiple applications, for example helping for search and discovery in bibliographic and bibliometric tools, creating indicators to understand how a research area is structured, or detecting new trends. In our case, we are building tools to support and help the steering of the Open Science public policy in France. In the French Open Science Monitor (Jeangirard 2019), a first attempt to classify publications is proposed in order to monitor the open access trends at the discipline level. With the COVID-19 crisis, the need to open up science in the health field has been reaffirmed. A classification at a lower level in the biomedical area is then needed to help to steer the Open Science public policy.

Several data providers already provide classifications, but one of the rules we have adopted in this work (COSO 2018) is to apply open science principles to monitor open science. Accordingly, we want to use and re-release open data, and use and redistribute open code. In this sense, using classifications from non-open and non-libre sources was not an option.

Different approaches for subject classifications have been set up, both at the journal level and the article level. At the journal level, expert-based classification is possible, as the ARC journal list ((ARC) 2018) with university consultations and disciplinary groups feedback. More algorithmic approaches, for example using journal-journal citations pattern (Leydesdorff and Rafols 2009) have been implemented. Going to a publication-level classification brings several advantages: first, it should allow classifying every type of document, published in a well-established journal or as a preprint. Also, it should identify more accurately articles published in multidisciplinary journals like Science or Nature (even though a single article can by itself be classified as multidisciplinary). At the article level, clustering techniques are often proposed, based on citation links like (Boyack et al. 2011) or (Waltman and Eck 2012). More recently, content-based deep learning techniques have been applied to Wikipedia articles (Semberecki and Maciejewski 2017).

In this paper, we investigate an article-level classification method based only on metadata like title, abstract and keywords (in the future, we plan to do so using the full-text). This method is not depending on citation indexes, and that could be, maybe, transposed to classify other scholarly objects than scholarly publications.

2. Method

2.1 Data sources

We need a data source for the publications in order to extract their metadata. For the moment, we have chosen a simple, reliable, and easy source to harvest, which is Pubmed. Pubmed metadata can easily be harvested through a public API, and the available metadata is rich, including affiliations, abstracts, keywords, and MeSH (Medical Subject Headings) (Medicine, n.d.).

There is also a need of a labelled database to train machine learning algorithms. That is the tough part as there is no open, recent, comprehensive, article-level tagged publications database. That is why we have chosen to introduce a proxy to go in-between journal-level and article-level tagging.

Through Excellence in Research for Australia (ERA), the Australian Research Council (ARC) released the ERA 2018 Journal List ((Mercieca and Macauley 2008), ((ARC) 2018)). This list associates, for more than 25,000 journals, up to 3 Field of Research (FoR) codes, or 'MD' for multidisciplinary journals. This journal-level information can be transposed to an article-level database (all the

articles with the same ISSN will be assigned the same fields), which can be used to train algorithms based on the other article-level metadata available, in particular the title, abstract, keywords, and MeSH.

The Fields of Research (FoR) is a hierarchical classification, with 2-digits, 4-digits, and 6-digits classes. We use the 2-digits and 4-digits FoR codes from the ERA 2018 journal list. More than 150 4-digits FoR codes exist, and most of them are not relevant for a health-specific classifier.

This source presents the merits of being quite recent, with a hierarchical structure, and produced by a public organization under a cc-by licence.

Selecting the FoR codes that are relevant for Pubmed papers (and biomedical in general) is not an obvious task. Of course, we could limit ourselves to FoR code 11 "Medical and Health Sciences" but that would miss a lot of fields, in particular from Biology, Chemistry, and Psychology. Asking for expert inputs could have been an option but it seems there is no strong consensus on the perimeter. So we eventually chose a quantitative approach, looking at the distribution of the FoR codes in Pubmed data (based on the ISSN).

2.2 Training

Getting labelled data at the article level is very costly, because it needs a high level of domain-specific expertise for a huge amount of publications. At the same time, labelled data does exist at the journal level. We try here to leverage that data to produce an article-level classifier.

The idea is to extract relevant pieces of information from available metadata (title, abstract, keyword, MeSH) to classify a publication. We assume that the publications of journals tagged with a given Field of Research, "Chemical Sciences" for example, will contain field-specific words and n-gram in their metadata that will be caught by a machine learning classifier. As a consequence, if an article from, for example, a "Multidisciplinary" journal contains enough words specific to "Chemical Sciences", then we guess that a machine learning approach will be able to classify it as "Chemical Sciences" rather than "Multidisciplinary". In that case, the classification at the article level would be different from the one at the journal level.

We propose to train multiple machine learning models, one for each metadata type: title, abstract, keywords, MeSH, and journal title.

The design of the training dataset is key in the final model relevance and performance. We evaluate two techniques to set up the training dataset. First, with a simple random sampling. This way the distribution of the classes in the training dataset is the same as in the whole data set. In a second approach, with a stratification sampling, each class being represented equally. The aim of this second sampling technique is to lower the risk of bias: indeed, if the dataset is very unbalanced, with one class being largely predominant, the risk the classifier overfits this class can be high. Imagine an extreme case in which 99% of the

cases are tagged with label A, and 1% of the cases with label B. Then a dummy classifier that always predicts A would have a 99% precision but would miss all the B cases. The idea of the stratified sampling is to train the algorithm with 50% of A cases and 50% of B cases to try to make the algorithm learn more relevant features.

Machine learning models are built with fasttext (Joulin et al. 2016) from Facebook. It uses word embeddings techniques like in word2vec models used in (Semberecki and Maciejewski 2017). Contrary to the latest deep learning models, fasttext is extremely efficient on CPU and very fast, and so is very adapted to low budgets environments.

2.3 Prediction

For the field prediction at the article level, we propose to combine the outputs of each of the 5 models (for each metadata available) with a voting system, giving slightly more weight to the journal title in case of equality. Only the prediction with an associated score above a given threshold is taken into account. As an example, if the model for title and keywords predict "Chemical Sciences" (FoR 03) and the model based on abstract and the one based on journal-title predict "Biochemistry and Cell Biology" (FoR 0601), the heuristic to pick up the selected field chooses the second one. But in another case, if the model for journal-title predicts "Multidisciplinary" (MD) whereas models for keywords, MeSH and title predict "Psychology and Cognitive Sciences" (FoR 17), then the heuristic returns the latter, giving a result different from the pure journal-title information.

3. Results

3.1 FoR codes selection

After matching the ISSN from the Pubmed metadata and the ERA journal list, we end up with an article-level database, enriched with Fields of Research (matched at the ISSN level)

Table 1 : Sample of the available data

pmid	issn	journal_title	FoR
31739602	2076-2607	Microorganisms	-
31178264	1950-6007	Biomedicine & pharmaco-therapy	Pharmacology and Pharmaceutical Sciences
31218652	1874-9356	Folia microbiologica	Microbiology ; Medical Microbiology
31669771	1556-3871	Heart rhythm	Biomedical Engineering ; Cardiorespira-tory Medicine and Haematology
31473396	1095-8630	Journal of environmental management	Multidisciplinary

We use a sample of 500,000 records published in 2019 from the Pubmed metadata to evaluate the following statistics.

First, it appears that 18% of the publications in PubMed cannot be attached directly to a FoR code: indeed their ISSN is not in ERA data, so there is no correspondence between the ISSN and FoR codes.

Table 2: Number of FoR matched to publications in Pubmed (sample from 2019)

Number of FoR	% of publications
No FoR	18%
1 FoR	37%
2 FoR	23%
3 FoR	22%

At the 2-digits FoR code level, all the 150+ FoR codes are present in the Pubmed data (from FoR 11 “Medical and Health Sciences” for 52% of the papers to FoR 19 “Studies in Creative Arts and Writing” for less than 0.1% of the papers). Note that a paper can be attached to multiple FoR codes (up to 3).

Table 3: Distribution of the 2-digits FoR codes in Pubmed (sample from 2019)

FoR	FoR code	Percentage of publications
Medical and Health Sciences	11	51.8%
Biological Sciences	06	16.8%
Chemical Sciences	03	11.3%
Engineering	09	8.9%
Multidisciplinary	MD	8.1%
Psychology and Cognitive Sciences	17	7.2%
Physical Sciences	02	3.2%
Agricultural and Veterinary Sciences	07	3.1%
other	x	$< 3\%$

Concerning the two main codes FoR 11 "Medical and Health Sciences" and FoR 06 "Biological Sciences", we look deeper into the 4-digits FoR codes. For the others, we selected only the FoR codes representing more than 3% of the papers in Pubmed. With the same logic, for the 4-digits FoR codes in 11 - "Medical and Health Sciences" and 06 - "Biological Sciences", we selected those representing more than 2% of the papers, and grouping the others into "Other Medical and Health Sciences" and "Other Biological Sciences".

After this selection process, we end up with the 17 fields presented in Table 4 to classify publications in Pubmed.

Table 4: Distribution of the selected fields in PubMed (sample from 2019)

Class	FoR code	Percentage of publications
Clinical Sciences	1103	19.9%
Chemical Sciences	03	11.3%
Other Medical and Health Sciences	-	10.3%
Engineering	09	8.9%
Multidisciplinary	MD	8.1%
Public Health and Health Services	1117	7.4%
Psychology and Cognitive Sciences	17	7.2%
Biochemistry and Cell Biology	0601	5.4%
Neurosciences	1109	4.5%
Other Biological Sciences	-	4.5%
Pharmacology and Pharmaceutical Sciences	1115	4.1%
Oncology and Carcinogenesis	1112	3.7%
Cardiorespiratory Medicine and Haematology	1102	3.3%
Physical Sciences	02	3.2%
Agricultural and Veterinary Sciences	07	3.1%
Paediatrics and Reproductive Medicine	1114	2.6%
Microbiology	0605	2.1%

With this field selection, some publications do not get a field anymore (the publications whose ISSN are not the selected scope). However, this loss is very low (less than 1%) as shown in Table 5.

Table 5: Number of fields matched to publications in Pubmed

Number of fields	2-digits FoR	Selected fields
0	17.9%	18.7%
1	37.1%	56.6%
2	22.9%	20.9%
3	22.1%	3.8%

3.2 Metadata availability in Pubmed

In our approach, we assume rich metadata is available, in particular metadata that can help infer a scientific discipline: title, abstract, keywords, MeSH (and journal title). In the general case (on Crossref for example), for most of the publications, the abstract and keywords are not available. We look here at Pubmed metadata.

Table 6: Metadata availability in Pubmed (sample from 2019 publications)

Metadata	Availability
Abstract	87.5%
Keywords	64.2%
MeSH	69.1%
at least 1 among abstract, keywords, MeSH	96.6%
at least 2 among abstract, keywords, MeSH	80.9%
all 3 metadata	43.3%

So the coverage is far from perfect, but in more than 80%, 2 out of 3 key metadata are available.

Field-wise, there is of course a variety of situations as shown in Figure 2. However, most of the selected fields have a good metadata availability, except from "Physical Sciences" (FoR 02) with less than 30% of publications with keywords available (but abstract is available in more than 95%). As a consequence, the situation is not perfect (100% availability would be a better scenario) but rich metadata (at least partially) remain available in the vast majority of the cases.

3.3 Training data

We used two training datasets of each 850,000 publications. Each of them is split for training and testing (90%, 10%). The first dataset is a random sample

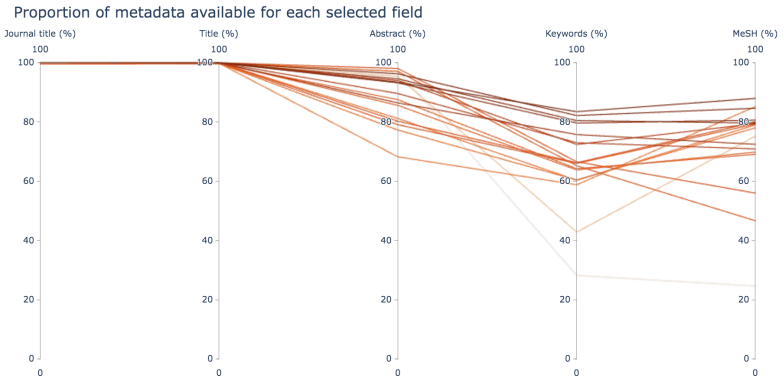

Figure 1: Proportion of publications with metadata available in Pubmed for the selected fields

from the Pubmed metadata (with at least a selected field assigned). The second dataset is a stratified sample, each selected field representing an equal part of the total.

For each dataset, we trained 5 *fasttext* models, for each metadata type: title, abstract, keywords, MeSH, and journal-title. The training parameters are presented in Table 7, the other being set to default values. All the possible parameters and their effects are detailed in fasttext online documentation (Facebook, n.d.).

Table 7: fasttext parameters used for training

Parameter	Value
epoch	50
wordNgrams	2
loss	ova
minCount	20

Several metrics can measure the performance of a classifier. In particular, the precision computes the fraction of predicted classes that are relevant and the recall that computes the fraction of relevant classes that are successfully predicted. The f1 score combines precision and recall (it is the harmonic mean of precision and recall). We report the f1 score of each model in Table 8.

Table 9: f1 score of trained model

Sampling technique	Journal title model	Title model	Abstract model	Keywords model	MeSH model
Random	99.7	40.8	44.6	43.6	42.8
Stratified	99.9	52.4	56.3	53.6	52.5

We observe that the stratified sampling approach gives overall better performance on each model.

We also notice that models based on journal-title have almost a perfect f1 score. That can be explained as the model simply tries to replicate a simple correspondence between ISSN (and so journal titles) and assigned FoR. The only advantage of the machine learning approach rather than the simple ISSN - FoR correspondence is the generalization for out-of-sample journals. The model is still able to predict a field, even for journals that are not part of the ERA 2018 journal list.

Apart from the journal-title model, the performance for the other model could seem low. Actually, we have to keep in my mind that a perfect f1 score is not even the objective, as, in that hypothetic case, the prediction would be exactly the same as the one based only on journal-title.

3.4 Gaining insights on calibrated fasttext embeddings

Fasttext is a word embeddings model, meaning that each word is represented by a numeric vector, in our case in dimension 100. The strength of this type of model is that the numeric distance between these vectors can be interpreted as a semantic distance.

The fasttext library comes with two handy functions to explore the word embeddings: *get_nearest_neighbors* and *get_analogies*. The first function lists the words whose embeddings are the closest to the input. For instance, using the model calibrated on titles, the 3 nearest neighbors of "infants" are "pregnancies", "childhood" and "children".

```
model.get_nearest_neighbors("infants")[0:3]
```

```
[(0.9249277710914612, 'pregnancies'),
 (0.912368893623352, 'children'),
 (0.9030213356018066, 'childhood')]
```

The other function enables the user to play around with word analogies. Fasttext documentation (Facebook, n.d.) gives the example of the triplet ("berlin", "germany", "france"), which can be interpreted as: "What is to France what Berlin is to Germany ?". In fasttext documentation, the first result given by the model they use is "paris". We played the same game with the model calibrated on titles, with the triplet ("hypertension", "heart", "brain"). That is to say,

according to the model we calibrated, what is to the brain what hypertension is for the heart? The two first results are “stroke” and “aneurysms”.

```
model.get_analogies("hypertension", "heart", "brain")[0:3]
```

```
[(0.9254266619682312, 'stroke'),
 (0.913159966468811, 'cerebral'),
 (0.9120014905929565, 'aneurysms')]
```

3.5 Classification inference

We applied the classification method on 45,000 publications from Pubmed with a French affiliation. For each one, we computed the field inferred with only the journal title and the field predicted using the combination of the 5 models (one for each metadata). For each model, we keep only the predictions with a probability above 0.5. The probability of each tag is directly computed by the fasttext library. We then look at the transition matrix between journal-based and article-based classification.

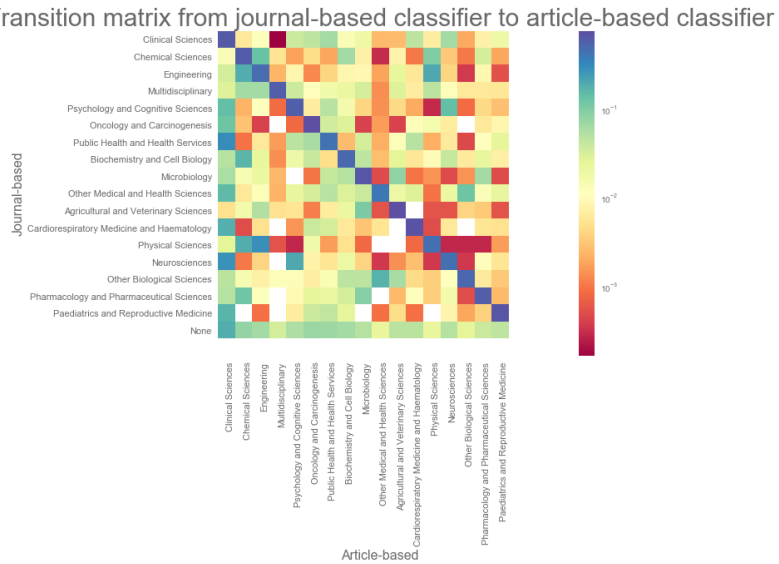

Figure 2: Transition matrix from journal-based to article-based classifier

As expected, in the majority of the cases, the article-based prediction is the same as the journal level one. However, a few things can be noticed. All the articles with no prediction at the journal level are classified with the article-level approach. The likelihood of a label change between the journal level and the article level is higher for some fields than others (in particular, “Multidisciplinary” or “Other Medical and Health Services”). For example, a publication whose title is ‘Topography and behavioral relevance of the global signal in the human

brain.', published in 'Scientific reports' was classified "Multidisciplinary" with a journal-based model and became classified "Psychology and Cognitive Sciences" with the combination of the models. Also, some transitions are more frequent, like "Physical Sciences" to "Engineering" or "Neurosciences" to "Psychology and Cognitive Sciences".

4. Discussion and conclusion

In this paper, we show how we combined existing and open data sources (Pubmed, Fields of Research) to train a machine learning model able to classify publications in the area of medicine and health. We proposed a mixed approach between journal-based and article-based classification and apply it to the metadata of French publications in Pubmed. This work was done as a pre-requisite for the construction of the French Open Science Monitor in Health and Medicine that will be released later. We propose a python implementation in https://github.com/dataesr/scientific_tagger

4.1 Findings

We first have shown that a combination of 17 Field of Research (2-digits and 4-digits) have good coverage of the publications in Pubmed, and propose to use these 17 fields for building a biomedical publication classifier.

We also underlined the role of the construction of the training dataset and put in place a technique to handle unbalanced class distribution using stratification.

With a few examples, we show the ability of fasttext models to manage to encode in their word embeddings at least part of the meaning of the words used in publications metadata (in the title in particular).

Finally, we have constructed a heuristic to predict a subject at the article level, using pieces of information from the available metadata (title, abstract, keywords, MeSH, and journal title). This method is most of the time in line with a pure journal-based classification, but brings two main advantages, being able to predict a class even if the journal is unknown and handling a bit better the articles published in Multidisciplinary journals.

4.2 Limitations and future research

The main limit of this work is the difficulty to evaluate the relevance of the final classification. Indeed, only a manual check on a representative sample of several hundreds of publications would allow to really measure the validity of the approach. That is part of future work we would like to conduct, like (Bornmann 2018) but on a larger scale to get insights on the potential bias of our method.

Another way to test the relevance of the output classes would be to use citation patterns to confirm or infirm the class's predictions like in (Wang and Waltman

2016).

Finally, we have to keep in mind that at the time of writing of (Joulin et al. 2016), fasttext was very efficient and on par with more complex deep learning architecture, but this area is evolving extremely fast and more recent NLP techniques should be investigated as well.

Software and code availability

The source code is released under an MIT license in the GitHub repository https://github.com/dataesr/scientific_tagger

References

(ARC), Australian Research Council. 2018. "ERA 2018 Journal List." https://www.arc.gov.au/excellence-research-australia/era-2018-journal-list.

Bornmann, Lutz. 2018. "Field Classification of Publications in Dimensions: A First Case Study Testing Its Reliability and Validity." *Scientometrics* 117 (1): 637–40. https://doi.org/10.1007/s11192-018-2855-y.

Boyack, Kevin W., David Newman, Russell J. Duhon, Richard Klavans, Michael Patek, Joseph R. Biberstine, Bob Schijvenaars, André Skupin, Nianli Ma, and Katy Börner. 2011. "Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches." Edited by Colin Allen. *PLoS ONE* 6 (3): e18029. https://doi.org/10.1371/journal.pone.0018029.

COSO, French Open Science Committee. 2018. "Feedback on EC Open Science Monitor Methodological Note." https://www.ouvrirlascience.fr/feedback-ec-science-monitor/.

Facebook. n.d. "Fasttext Documentation." https://fasttext.cc/docs/en/supervised-tutorial.html.

Jeangirard, Eric. 2019. "Monitoring Open Access at a National Level: French Case Study." In *ELPUB 2019 23d International Conference on Electronic Publishing*. OpenEdition Press. https://doi.org/10.4000/proceedings.elpub.2019.20.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "Bag of Tricks for Efficient Text Classification." *arXiv:1607.01759 [Cs]*, August. http://arxiv.org/abs/1607.01759.

Leydesdorff, Loet, and Ismael Rafols. 2009. "A Global Map of Science Based on the ISI Subject Categories." *Journal of the American Society for Information Science and Technology* 60 (2): 348–62. https://doi.org/10.1002/asi.20967.

Medicine, U. S. Nation Library of. n.d. "Medical Subject Heading (MeSH)." https://www.nlm.nih.gov/pubs/factsheets/mesh.html.

Mercieca, Paul, and Peter Macauley. 2008. "A New Era of Open Access?" *Australian Academic & Research Libraries* 39 (4): 243–52. https://doi.org/10.1080/00048623.2008.10721362.

Semberecki, Piotr, and Henryk Maciejewski. 2017. "Deep Learning Methods for Subject Text Classification of Articles." *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, ACSIS, 11: 357–60. https://doi.org/10.15439/2017F414.

Waltman, Ludo, and Nees Jan van Eck. 2012. "A New Methodology for Constructing a Publication-Level Classification System of Science: A New Methodology for Constructing a Publication-Level Classification System of Science." *Journal of the American Society for Information Science and Technology* 63 (12): 2378–92. https://doi.org/10.1002/asi.22748.

Wang, Qi, and Ludo Waltman. 2016. "Large-Scale Analysis of the Accuracy of the Journal Classification Systems of Web of Science and Scopus." *Journal of Informetrics* 10 (2): 347–64. https://doi.org/10.1016/j.joi.2016.02.003.