# Classifying encyclopedia articles: Comparing machine and deep learning methods and exploring their predictions

Alice Brenon, Ludovic Moncla, Katherine Mcdonough

# Classifying encyclopedia articles: Comparing machine and deep learning methods and exploring their predictions

Alice Brenon[a,b,*], Ludovic Moncla[a], Katherine McDonough[c]

[a] *Univ Lyon, INSA Lyon, CNRS, UCBL, LIRIS, UMR5205, F-69621*
[b] *ICAR, CNRS, UMR5191, 69342*
[c] *The Alan Turing Institute, UK*

**Abstract**

This article presents a comparative study of supervised classification approaches applied to the automatic classification of encyclopedia articles written in French. Our dataset is composed of 17 volumes of text from the *Encyclopédie* by Diderot and d'Alembert (1751-72) including about 70,000 articles. We combine text vectorization (bag-of-words and word embeddings) with machine learning methods, deep learning, and transformer architectures. In addition evaluating these approaches, we review the classification predictions using a variety of quantitative and qualitative methods. The best model obtains 86% as an average f-score for 38 classes. Using network analysis we highlight the difficulty of classifying semantically close classes. We also introduce examples of opportunities for qualitative evaluation of "misclassifications" in order to understand the relationship between content and different ways of ordering knowledge. We openly release all code and results obtained during this research[1].

*Keywords:* classification, supervised machine learning, deep learning, encyclopedia, computational humanities, networks

## 1. Introduction

Understanding the organization of knowledge over time is a major area of research that spans the disciplines [1, 2, 3, 4]. Historians, for example, study the kinds of documents where humans store knowledge, how knowledge is divided into categories, what the relationship is between those categories, and how they change over time (see, for example, [5, 6]). Sometimes documents explicitly label parts of their content: since the scientific revolution this has been common for reference works like encyclopedias. In texts like these, the goal is to support searching for information within broad concepts that represent a particular system for organizing knowledge. Because they tend to be organized in such systems by their editors, encyclopedias are a uniquely interesting site for exploring how the classification of knowledge changes over time. However, there are some exceptions. First, not all encyclopedias use explicit categories: they can resemble dictionaries thanks to their alphabetical organization. Second, a specific set of categories is not always applied systematically across articles. Encyclopedias are often multi-volume works in which editorial practices can evolve. Finally, sometimes people apply new categories of knowledge to historical texts. These new organizational schema can have different purposes. They can seek to simplify and reduce the number of classes of knowledge, or, alternatively, they can impose a radically different vision of how humans organize knowledge. Using computational text analysis to discover patterns in encyclopedic discourse sheds light on historical and more recent classification practices and the relationship between article classification and content.

Our experiments in this paper contribute to developing methods for such a study of encyclopedic knowledge. We focus on the famous Enlightenment text edited by Diderot and d'Alembert: the *Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers par une société de gens de lettres* (1751-1772),

*Corresponding author: alice[dot]brenon[at]insa-lyon.fr
[1] `https://gitlab.liris.cnrs.fr/geode/EDdA-Classification`

henceforth *EDdA*. This is part of a larger investigation which compares the content and organization of French encyclopedias from the late seventeenth century until today, specifically identifying and analyzing geographical discourse.[2] This particular encyclopedia's organization was designed to offer several ways to navigate its content: alphabetical order (like dictionaries), classification based on a new representation of human understanding, and cross-references. To outline the principles of the classification strategy, the *Système figuré des connoissances humaines* diagram[3] was published in the first volume. But this should be seen more as a potential navigational aid rather than a set of rules. *EDdA* is at its heart an experimental text, and the approach to classification was very much in character. In an oft-cited passage from d'Alembert's *Discours préliminaire*, he puts the *Système* into context saying that "the form of the encyclopedic tree will depend on the vantage point one assumes in viewing the universe of letters. Thus one can create as many different systems of human knowledge as there are world maps having different projections...." (15)[7]. We embrace this way of engaging with *EDdA*'s articles, but there is ultimately a tension between the task of evaluating machine-generated classifications and an understanding that there is not necessarily one, correct classification for a given article.[4]

In addition to the intentional multiplicity of pathways into its content, *EDdA*'s presentation of ordered knowledge has gaps that somewhat hinder our ability to easily use it as machine-readable data. First, there are a large number of unclassified articles. Furthermore, there is stylistic variation in the way that classes were indicated for each article. For this reason, twentieth-century researchers have attempted to standardize these classes and to predict the classes for originally unclassified articles. For example, the *Édition Numérique Collaborative et CRitique de l'Encyclopédie*, or the ENCCRE project[5] manually grouped articles into "domains" to cope with both of these challenges. And the ARTFL project[6] has tested automatic methods for normalizing and predicting articles classes (See 2. By superimposing new order on *EDdA*, we gain the ability to compare it to other encyclopedias in future work.

As part of this larger research agenda, in this article we present a comparative study of supervised Machine Learning (ML) and Deep Learning (DL) methods for the task of classifying encylcopedia articles. This builds on recent work in this area[9]), but extends the task to Deep Learning and experiments with different training data. *EDdA* contains more than 7,000 classes of knowledge, including many textual variations for expressing the same class. Starting from ENCCRE's 44 domain ensembles rather than ARTFL's 2,620 *normclasses* (or indeed the more varied original classes) is a dramatic restriction of the representation of knowledge in *EDdA*, but it is a first step in testing whether we can reproduce knowledge domains that do have some relationship to the original categories of knowledge and were selected by experts in eighteenth-century literature and history.

Section 2 reviews similar studies in the computational humanities and social sciences working on document classification. We also discuss the language-specific issues we face in this research with historical French. Section 3 describes the *EDdA* corpus, pre-processing steps, and describes the method. Section 4 presents the experiments and the results. We evaluate the results and analyze them at a high level (all classes) and with a case study examining only the results for the Geography class. We create and analyze a graph representation of the results from the SGD+TF-IDF model and explore lexical similarities between classes. Section 5 discusses the results and their implications and, as a comparison to our results from supervised classification, presents a preliminary experiment to organize *EDdA* articles using unsupervised clustering. Section 6 reviews limits and opportunities of classifying encyclopedia articles using our methods.

## 2. Related work

Document classification is a general problem in text analysis. Classification might mean assigning documents to a topic (infrastructure or foreign policy), a type of content (news, advertisement), or a type of

[2] `https://geode-project.github.io`

[3] `https://encyclopedie.uchicago.edu/content/syst%C3%A8me-figur%C3%A9-des-connaissances-humaines-de-lencylcop%C3%A9die`

[4] We have not taken the approach of examining Geography as a concept, for it was highly unstable, and we are only working with one text, but in the future computational approaches to conceptual history that focus on understanding how words associated with concepts evolve may be useful [8].

[5] `http://enccre.academie-sciences.fr/encyclopedie/`

[6] ARTFL, American and French Research on the Treasury of the French Language: `https://artfl-project.uchicago.edu/`

author/speaker (Labor or Conservative). Text corpora similar to encyclopedias include collections of political speeches (like Hansard, the US Congressional Record, or the *Archives parlementaires* for France). Here we survey existing literature that classifies large historical text corpora using different methods.

*Classifying encylopedias.* In exploring methods for classifying *EDdA* articles, we follow in the footsteps of the ARTFL team. In their 2009 paper Hornton et al [10] tested Naïve Bayesian classification on two tasks: 1) classifying the originally unclassified articles and 2) applying this model on the already classified articles to compare the results. This second task also enabled them to explore which words were most important for the classification result. While the paper did not include a formal evaluation of the performance of the model, it did offer an important close reading for a selection of the results. In their 2016 paper Roe et al [9] used Latent Dirichlet Allocation (LDA) topic modelling to analyze automatically-identified groups of articles, and to compare these to the original classes. This research posited that the LDA-identified topics could be understood as discourses that were woven throughout *EDdA*, and which do not always neatly map onto the original classes. In many ways, our work is motivated by this earlier research. We aim to establish a baseline for the classification task which can be improved on in the future, and which can be compared when using different classification metadata to fine-tune models (e.g. original classes, ARTFL simplified *normclasses*, or ENCCRE domain ensembles.

We also take inspiration from researchers working with other encyclopedias. The Nineteenth-Century Knowledge project explored rule-based and ML methods[7] to index 400,000 articles across 4 editions of the Encyclopedia Britannica [11].[8] Because Britannica editors did not use the same article classes over time, matching articles with Library of Congress Subject Headings enables cross-edition comparison and therefore improved discovery.

*Classifying other texts.* Beyond encyclopedias, humanities research has largely used text classification for subject or genre detection ("is this historical fiction or biography?" - see the Underwood examples just below) and author/group identification ("was this speech given by a Labour or Conservative MP?" [12]).

The popularity of LDA topic modeling for assessing the content of large text data is at least in part explained by the fact that it does not require pre-existing metadata or new annotations describing documents or document sections that can be used as training data: it is quicker to implement. In her analysis of British parliamentary speeches (Hansard), Guldi [13] employs topic modeling to "critically search" for "tensions and turning points" in political debates in the UK. Baron et al [14] use topic modeling as a jumping off point from which to measure the "novelty" and "transience" of speeches made during the first years of the French Revolution. This is useful because while the speeches are usually attributed to a specific deputy and are dated, there is no other metadata about each speech.

Using both LDA and other ML models, Underwood examines the history and instability of literary genre [15, 16, 17]. He offers that computational methods are useful because it can "register and compare blurry family resemblances that might be difficult to define verbally without reductiveness" (6) [16]. Such a quantitative, predictive approach to text classification enables computational humanities research like ours to think through the results in a different kind of interpretative environment.

What does this all mean for encyclopedias written in eighteenth-century France, and how does it impact our experiment design and interpretation? First, we emphasize again that encyclopedia classes are, like genre, culturally-constructed categories that change over time (even within the volumes of one publication!). Second, our ability to recreate these classes using models sheds light on the extent to which they hold fast to certain linguistic features and points us to specific subsets of the work that conform or do not conform to the predictions (e.g., by evaluating true positives vs. false positives).

### *2.1. Working in French*

Finally, our research uses texts written in French, with a smattering of other languages (especially Latin and Greek), during the eighteenth century [18]. We use some language-dependent methods on language models pre-trained on French documents. For example, we use the French version of FastText with CNN and LSTM experiment but also multilingual BERT and CamemBERT (which is pre-trained on French texts).

[7] `https://cci.drexel.edu/mrc/research/hive/`
[8] `https://tu-plogan.github.io/index.html`

It can no longer be said that French is a low-resource language in Natural Language Processing, but lack of linguistic diversity in NLP still plays a role in experiment design. Perhaps even more important is the historical nature of our texts. We therefore still face hurdles in model performance that do not exist when one is working with short, modern, English texts [19, 20].

## 3. Data and Methodology

### *3.1. From corpus to dataset*

*EDdA* data is available from two sources: ENCCRE and ARTFL. The ENCCRE project's major output is a collaboratively enriched digital edition based on one of the sets held at the Bibliothèque Mazarine[21]. Using their online platform, it is possible to browse a page image alongside a transcript of the articles on that page and also access related scholarly commentary. Similarly, ARTFL also provides online access to *EDdA* text via Philologic[9], a powerful tool for searching large TEI corpora. Articles in *EDdA* usually include a headword, a classification, grammatical details, the main text body as well as the text of sub-articles, an author's 'signature,' and cross-references. An article's headword is usually (but not systematically) followed by a classification (again, usually, but not always) within parentheses. For example, in figure 1, the article about the city of Evian is classified as "Modern Geography." These classifications printed in the text are the domain of knowledge assigned by the editors at the time of publication.

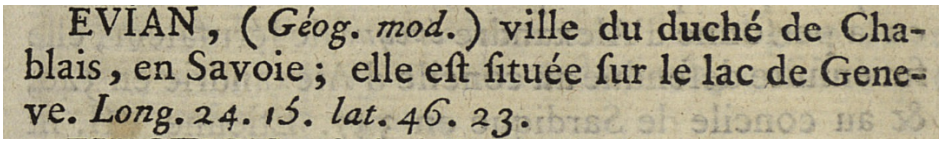

EVIAN, (*Géog. mod.*) ville du duché de Chablais, en Savoie; elle eſt ſituée ſur le lac de Geneve. *Long.* 24. 15. *lat.* 46. 23.

Figure 1: Article "Evian", vol. VI, n° 417.

Unsurprisingly for a publishing project that spanned many years, *EDdA* contains more than 7,000 unique forms of classifications. Among these are variations on abbreviations, spelling, punctuation used to set off a classification from the headword and the main text, and other patterns of language used to identify domain at the beginning of an article. Normalizing these classifications for digital editions and research has been the object of previous research, both by ARTFL and ENCCRE. ARTFL has reduced this number to 2,620 forms (corresponding to the *normclass* attribute in their TEI encoding), and then used ML methods to predict this normclass for every article. ENCCRE manually created a list of 2,160 classes (*désignants explicités*). In order to improve searches for similar articles in their online interface, ENCCRE resolved these classes to 327 *domaines* (domains) and again to 44 *ensembles de domaines* (domain groups). For example, in the article *EVIAN*, the classification *(Géog. mod.)* was expanded to a normalized label of *Géographie moderne*, and then combined in a *Géographie* domain group.

One of the challenges with this approach is that 12,635 articles do not have an explicit classification in the text. However, ENCCRE succeed in reducing this number to 2,392 after a manual correction step that consists of identifying implicit classifications based on on the body of the article. In contrast, the ARTFL approach to normalize classifications depends on automatic methods. *EVIAN* is therefore associated with the *normclass Géographie moderne*.[10] Out of the 77,085 articles in the ARTFL version of *EDdA*, 55,248 can be classified in this way among 2,620 *normclasses* and 21,837 articles remain unclassified.

In this work, we use the ARTFL *EDdA* corpus to focus exclusively on the 17 volumes of text articles (77,085 articles) encoded in XML-TEI format. Each file maps to one article and contains article-level metadata in addition to the body of text of the article such as the headword, the author(s), the tome number, the article position in the flow of articles (counted from the beginning of each tome), the name of original classification, and the *normclass*.

To prepare the data for our experiments, we enrich the ARTFL metadata with a slightly simplified version of the ENCCRE domain groups: we combine certain domain groups related to professions in an umbrella

[9] `https://github.com/ARTFL-Project/PhiloLogic4`

[10] `https://artflsrv03.uchicago.edu/philologic4/encyclopedie1117/navigate/6/421/`

group called *Métiers*. We therefore work with 38 classes rather than 44. These steps dramatically reduce the complexity of the classification task, which is a major concern for us. However, there remains significant imbalance in the number of articles per domain group (see Figure 2). For example, *Géographie* contains 13,289 articles while the next largest group only includes 6,901 articles. Some groups contain fewer than 1,000 articles (*Mathématiques*, *Musique* and *Arts et métiers*). To reduce bias that might be introduced by this imbalance, we created three samples using different "ceilings" for the number of articles in each group. This allows us to experiment with limiting the difference between domain groups (see section 4) to see if this imbalance has an effect on each method's performance.

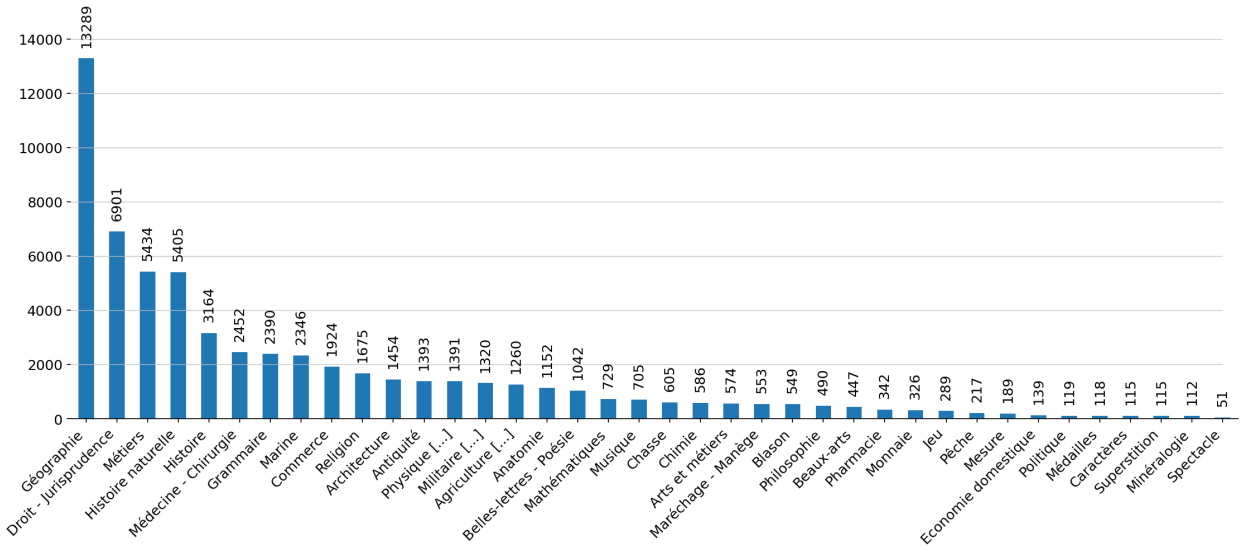



Figure 2: Number of articles in ENCCRE domain groups.

### *3.2. Vectorization*

In this paper, we compare different approaches to classifying documents in order to evaluate future encyclopedia article classification research. The method combines vectorization (or feature extraction) and classification algorithms. The first approach treats documents (e.g. the articles) as "bags-of-words" [22] where the collection of words (tokens or lemmas) in a corpus constitutes a vocabulary used as the basis for a large-dimension vector space where each document is represented. The component of a document's vector on each dimension is an integer greater than or equal to zero representing the number of occurrences in it for the corresponding element of the vocabulary (its frequency). This approach produces a large and sparse vector representation because each article includes only a small number of words that appear in the total words present in a corpus. Then, TF-IDF (term frequency-inverse document frequency) is a type of bag-of-words representation that weights words by their frequency in a corpus as well as within a document. Similar to bag-of-words, this method produces a very large and sparse vector representation which tends to alter the performances of ML algorithms. Additionally, this usually increases the space and time complexity of ML models. Finally, word embedding methods capture the context of words within a vector space. These representations are pre-trained on a large corpus by self-supervised deep learning. There are two architectures for training word embeddings: CBOW and skip-gram. The first predicts a word in relation to its context. Inversely, skip-gram predicts the context of a word. In contrast to bag-of-words methods, word embeddings produce dense vectors with fewer dimensions. However, methods like Word2Vec [23] are not well suited to tasks using longer texts (¿ 10-15 words). For this reason, we use Doc2Vec [24], which is better adapted for vectorizing longer documents. This is useful in our case because articles (which we treat as documents) vary greatly in size, from just a few words to tens of thousands of words. Word embedding methods such as Word2Vec are considered "static" because they produce a single vector representation combining different contexts of a word within the same vector. Newer methods like BERT (Bidirectional Encoder Representations from Transformers) [25] use contextual word embeddings where each word has a

representation as a function of the way it appears within a phrase, or a window of tokens. BERT uses transformer neural networks and the concept of masking to predict the next word(s). In addition to BERT (pre-trained multilingual version), we test CamemBERT [26], a pre-trained language model for documents in French.

### *3.3. Supervised classification*

Supervised classification uses vector representations of articles in the training set along with the class associated with each article (i.e., the ENCCRE domain group) as input. For this study, we test different classification algorithms: ML methods and deep learning methods using different neural network architectures. For the ML methods, we test naïve bayes, logistic regression, random forest, and Support Vector Machines (SVM). Naïve bayes can only be used with bag-of-words vector representations (bag-of-words and tf-idf) because, unlike word embeddings, the vectors do not contain negative values. We also test Stochastic Gradient Descent (SGD) as an optimization method combined with SVM. SGD is better adapted to low-density data with many dimensions such as bag- of-words vector representations.

In addition to classic ML methods, we also test CNN (Convolutional Neural Network) and BiLSTM (Bidirectional Long Short Term Memory) deep learning models. CNNs are based on two operations: convolution and max-pooling. Convolution extracts features from the data and max-pooling compresses those results. BiLSTM are based on Recurrent Neural Networks (RNNs) and are well adapted for analyzing sequences of data; they have both backward and forward information about the sequence. Each neuron is replaced by a memory unit which contains the neuron and a recurrent self-connection. It allows the model to capture dependencies over a longer sequence than classical neural networks. To train deep learning models for classification, there are two approaches for vectorization. First, it is possible to set an embedding layer whose weights at the outset are selected randomly. Alternatively, it is possible to use a pre-trained word embedding model. For both, it is necessary to add an embedding layer to a neural network. In our experiments, we use the Keras library for CNNs and BiLSTMs, and we use the French version of FastText[27] for the pre-trained word embedding layer. Finally, as a complement to the static word embeddings such as Doc2Vec and FastText, contextual language models like BERT can be fine-tuned on our corpus. This transformer-based architecture obtains better results in state-of-the-art research on several text analysis tasks such as document classification.

## 4. Experiments

Our experiments compare approaches to classifying *EDdA* articles using vectorization and supervised classification. We test the following combinations:

1. Bag-of-words vectorization and classic ML algorithms (Naive Bayes, Logistic Regression, Random forest, SVM and SGD);
2. Vectorization using static word embeddings (Doc2Vec) and classic ML algorithms (Logistic regression, Random Forest, SVM et SGD);
3. Vectorization using static word embeddings (FastText[fr]) and deep learning algorithms (CNN and BiLSTM);
4. An *end-to-end* approach using pre-trained contextual language models (BERT, CamemBERT) with fine-tuning to adapt the model for our task.

For the ML algorithms, we use Scikit-learn[11] and *GridSearchCV()* for determining hyperparameters. For the CNN model we use a classic architecture with an embedding layer with a vector size of 300 dimensions, a convolution layer with a Relu activation function, a max pooling layer, and a softmax output layer. For the BiLSTM model, we test several architectures and use here an embedding layer with a vector size of 300 dimensions, a bidirectional LSTM layer (with 20% of recurrent dropout), a max pooling layer, two dense layers with a Relu activation function and a dropout of 0.5 between each, and a softmax output layer. For the BERT models, we fix a batch size of 8 (due to memory requirements) and 4 epochs.

[11] `https://scikit-learn.org/stable/`

### *4.1. Datasets and data preparation*

We use two datasets to train and evaluate our models (train and test). These two datasets (Table 1) contain articles labeled by ENCCRE. For this paper, we use only one class[12] per article. But in reality 3,654 (about 5%) of *EDdA* articles are assigned to multiple classes. Before proceeding with our experiments, we pre-process the collection. First, we remove unclassified articles and articles with fewer than 15 words. This produces a dataset of 58,509 articles assigned to 1 of the 38 classes. We reserve 20% of the articles for a test set. The remainder are used for training. Next, we remove punctuation and stop words (articles, prepositions).

| Datasets | # articles |
|---|---|
| complete corpus | 74,190 |
| pre-processed corpus | 58,509 |
| train (all) | 46,807 |
| train (max 1,500) | 27,381 |
| train (max 500) | 14,058 |
| test | 11,702 |

Table 1: Breakdown of the complete corpus, the pre-processed corpus, and the train and test sets.

Given the imbalance in the number of articles between classes, we wish to evaluate the impact of using samples of articles from those classes that are over-represented. We hypothesize that this will reduce their impact on the less represented classes. We compare the results for three different limits (up to 500 articles, 1,500 articles, and no limit). These sub-datasets represent 14,058 articles total for sets from all classes up to 500 articles for each class, 27,381 for the sets of up to 1,500, and 46,807 for all articles (See Table 1.)

### *4.2. Classification evaluation*

To evaluate the classification results, we use precision, recall, and f-scores. In order to obtain a general result for all classes, we also examine the mean of weighted f-scores obtained for the 38 classes. Table 2 presents mean f-scores for the different methods.

| Classifier | Vectorizer | F-score | | |
|---|---|---|---|---|
| | | (1) | (2) | (3) |
| Naive Bayes | Bag-of-words | 0.63 | 0.71 | 0.70 |
| | TF-IDF | 0.74 | 0.69 | 0.44 |
| Logistic Regression | Bag-of-words | 0.74 | 0.77 | 0.79 |
| | TF-IDF | 0.77 | 0.79 | 0.81 |
| | Doc2Vec | 0.64 | 0.69 | 0.77 |
| Random Forest | Bag-of-words | 0.57 | 0.54 | 0.16 |
| | TF-IDF | 0.55 | 0.53 | 0.16 |
| | Doc2Vec | 0.63 | 0.66 | 0.60 |
| SGD | Bag-of-words | 0.70 | 0.73 | 0.75 |
| | TF-IDF | 0.77 | 0.81 | 0.81 |
| | Doc2Vec | 0.68 | 0.72 | 0.76 |
| SVM | Bag-of-words | 0.71 | 0.75 | 0.78 |
| | TF-IDF | 0.77 | 0.80 | 0.81 |
| | Doc2Vec | 0.68 | 0.74 | 0.78 |
| CNN | FastText | 0.65 | 0.72 | 0.74 |
| Bi-LSTM | | 0.69 | 0.79 | 0.80 |
| BERT Multilingual (*fine-tuning*) | - | 0.81 | 0.85 | 0.86 |
| CamemBERT (*fine-tuning*) | - | 0.78 | 0.83 | 0.86 |

Table 2: Mean f-scores for different models for the test set with a sample of a maximum of 500 articles (1), 1500 (2), and no limit (3).

The *Random Forest* method obtains the worst results whatever the vectorization method or the sampling (between 16% and 66%). The *Naive Bayes* method obtains results between 44% and 74% with a very

[12]For the remainder of the article the word *class* will refer to the ENCCRE domain group.

significant impact of the sampling for the TF-IDF vectorization. The *Logistic regression*, *SGD* and *SVM* methods obtain very similar results and the best ones are those associated with a TF-IDF vectorizer (around 80%). Surprisingly, the Doc2Vec word embedding representation produces results slightly below the bag-of-words representations. The scores increase as a function of the sample size. Therefore, this approach likely requires a larger dataset to perform well. Deep learning approaches using neural networks (CNN and BiLSTM) obtain comparable scores (between 65% and 80%), but are slightly below the best classical ML methods associated with a TF-IDF vectorizer. Class balancing has a larger negative effect than for classical methods due to the reduction of the dataset. Fine-tuning language models such as BERT Multilingual and CamemBERT on our classification task obtains slightly better, but still comparable results to classic ML methods (with 86% as the best f-score). The results between the two models (BERT Multilingual and CamemBERT) are very close in terms of global average but differ slightly for each class and also based on the sample limit. Figure 3 shows f-scores obtained for BERT Multilingual (blue curve) and CamemBERT (orange curve) for each class (grey curves refer to the other methods). Classes are sorted from left to right according to their prevalence in the sample (highest number of articles on the left and lowest on the right). We can notice that both methods obtained similar scores for the majority of classes and that they only differ for classes with few articles. For example, the class *Economie domestique* obtains 44% with BERT versus 61% with CamemBERT, class *Politique* 53% versus 8%, class *Minéralogie* 70% versus 37%, and class *Spectacle* 62% versus 0%. Since the largest gaps are for underrepresented classes, the impact on the overall average is small and is offset by smaller gaps on the most populated classes. This figure also highlights that the class *Arts et métiers* obtains bad scores with all methods. This domain is often confused with *Métiers*. Ambiguity and similarity between classes will discussed on Section 4.3.

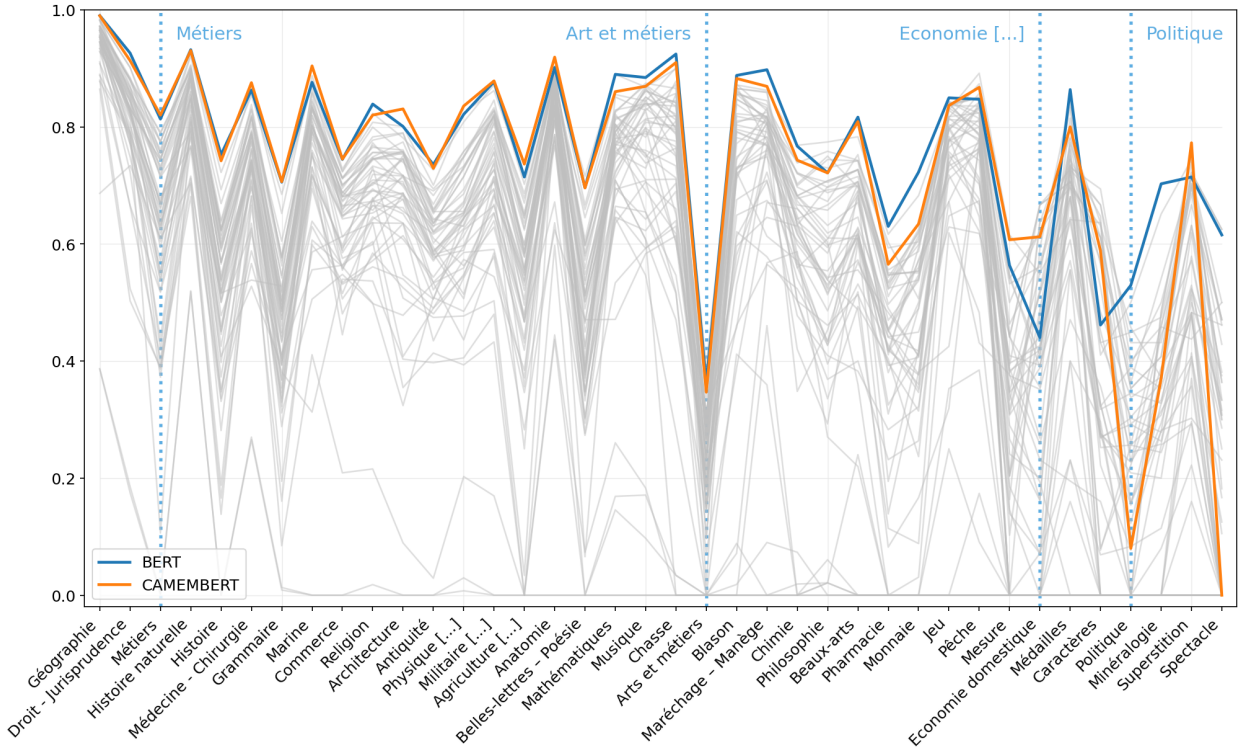



Figure 3: F-scores obtained with BERT Multilingual and CamemBERT for each class.

In general, increasing the sample size per class (e.g. the maximum number of articles), and thereby increasing the gap between the least and most populated classes, improves the global results but negatively impacts the results for underrepresented classes. This suggests that it is better to have more data even with unbalanced classes than less data with balanced classes, if the goal is to improve results across many different classes. Two methods have a different behavior: *Random Forest* and *Naive Bayes* + TF-IDF (see Fig. 4). In the case of Naive Bayes, we can clearly see on Figure 4 that the experiment without sampling (green

line) has poor results and that the sampling strategy has a positive impact. For example, Figure 4 shows that the unsampled model (green curve) trained with unbalanced classes only works for the most populated classes (towards the left). The model trained with a sample where there are no more than 1,500 articles per class (orange curve) works well, and only performs dramatically worse when a class contains fewer than 500 articles. Finally, only the most-balanced model correctly classifies a majority of the classes (excluding ones with very limited articles such as *Minéralogie*, *Superstition* and *Spectacle*). Among the ML methods, TF-IDF almost always has better results than bag-of-words and Doc2vec. This is the case for mean f-scores, but also those for each class. Figure 5 shows f-scores obtained with SGD on each class with three different vectorizers without sampling. These three models obtain similar results, but the TF-IDF model has better scores for most classes.

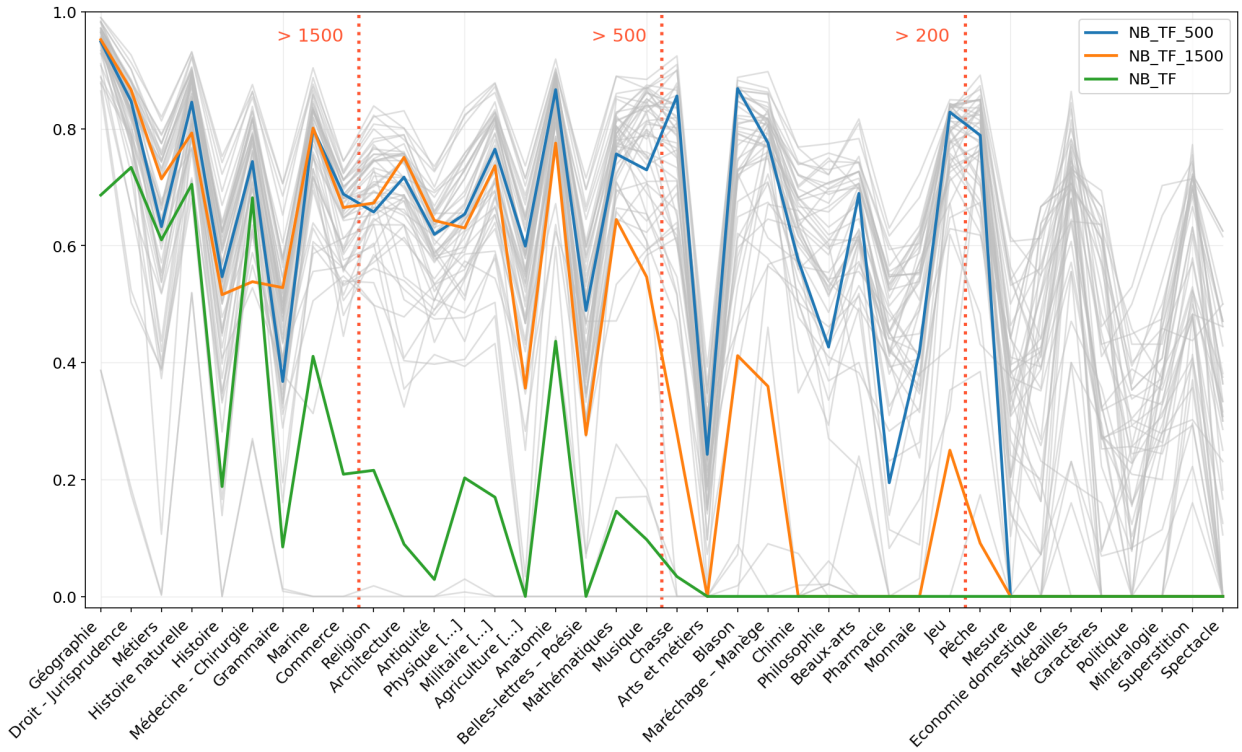



Figure 4: F-scores obtained with Naive Bayes + TF-IDF on each class with three different sampling.

Table 3 and Figure 6 show disparity among results for different classes based on f-scores for the test set (without sampling) for each class (sorted by the number of articles in each class) for (1) SGD + TF-IDF, (2) BiLSTM + FastText and (3) BERT Multilingual. Of the 38 classes, 31 obtain more than 70% with BERT (25 with SGD+TF-IDF and 19 with LSTM+FastText) while only 3 have less than 50% with BERT (5 with SGD+TF-IDF and 10 with LSTM+FastText). Generally, classes with the many articles (¿1,000) obtain very good scores. *Géographie*, for example, has a score of 99%. For the classes with the least data (¡500 articles), there is a notable drop in performance. There are, however, exceptions, such as *Pêche* (Fishing) and *Médailles* (Medals). Both have very few articles in the training set (168 and 94, respectively) but are surprisingly well classified (at 85% and 86% with BERT).

Beyond underlining the importance of the number of articles per class, these results highlight the difficulty of distinguishing between classes due to lexical or semantic similarity. This is clearly visible in Figures 3 to 6 for the class *Arts et métiers* (a broad term in French that refers to professions, but especially mechanical and technical ones). This domain is badly classified by every method when compared to similarly-sized classes. It is likely that this domain is confused with *Métiers* (a more generic term for professions, which here is a composite of different professions that were originally independent ENCCRE domain groups), which is among the most well-represented domains. This hypothesis seems confirmed by the confusion matrix shown in Figure 7.

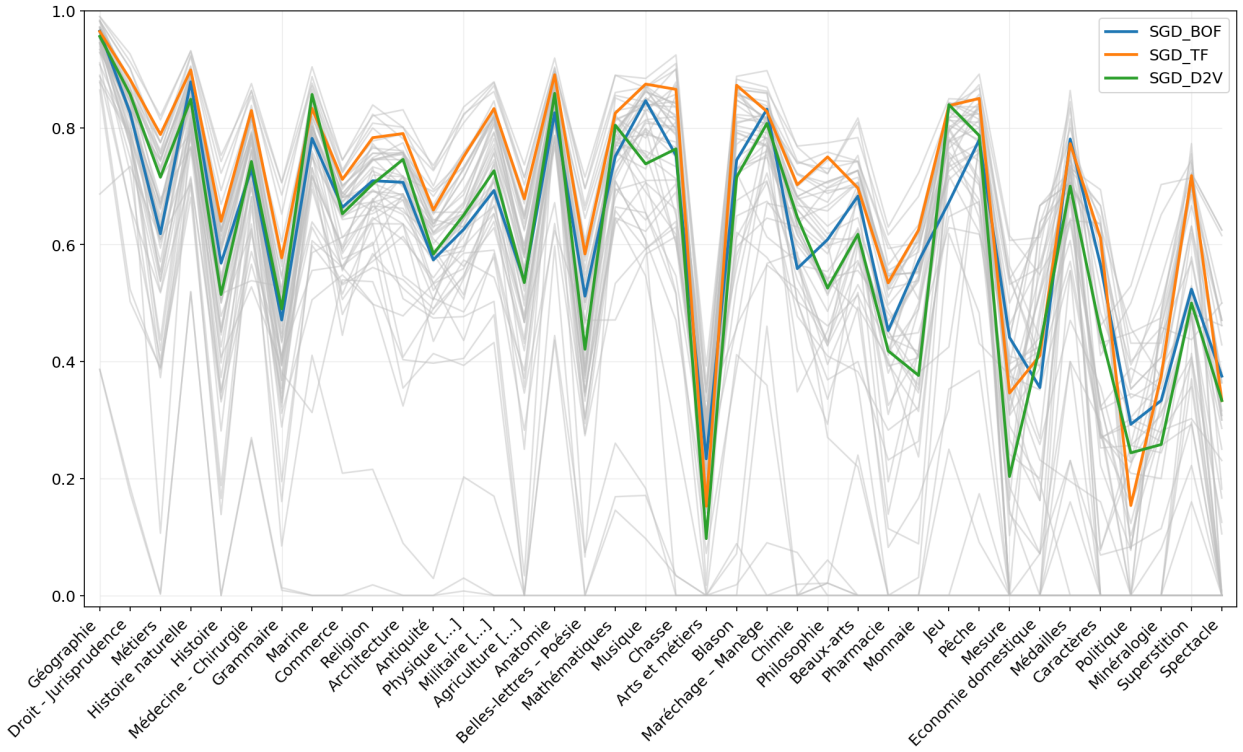



Figure 5: F-scores obtained with SGD on each class with three different vectorizers without sampling.

| Domain groups (classes) | # | (1) | (2) | (3) | Domain groups (classes) | # | (1) | (2) | (3) |
|---|---|---|---|---|---|---|---|---|---|
| Géographie | 2 621 | 0.96 | 0.98 | 0.99 | Chasse | 116 | 0.87 | 0.87 | 0.92 |
| Droit - Jurisprudence | 1 284 | 0.88 | 0.90 | 0.93 | Arts et métiers | 112 | 0.15 | 0.27 | 0.36 |
| Métiers | 1 051 | 0.79 | 0.76 | 0.81 | Blason | 108 | 0.87 | 0.86 | 0.89 |
| Histoire naturelle | 963 | 0.90 | 0.87 | 0.93 | Maréchage [...] | 105 | 0.83 | 0.86 | 0.90 |
| Histoire | 616 | 0.64 | 0.64 | 0.75 | Chimie | 104 | 0.70 | 0.58 | 0.77 |
| Médecine [...] | 455 | 0.83 | 0.80 | 0.86 | Philosophie | 94 | 0.75 | 0.49 | 0.72 |
| Grammaire | 452 | 0.58 | 0.54 | 0.71 | Beaux-arts | 86 | 0.70 | 0.62 | 0.82 |
| Marine | 415 | 0.83 | 0.86 | 0.88 | Pharmacie | 65 | 0.53 | 0.38 | 0.63 |
| Commerce | 376 | 0.71 | 0.69 | 0.74 | Monnaie | 63 | 0.63 | 0.50 | 0.72 |
| Religion | 328 | 0.78 | 0.77 | 0.84 | Jeu | 56 | 0.84 | 0.74 | 0.85 |
| Architecture | 278 | 0.79 | 0.74 | 0.80 | Pêche | 42 | 0.85 | 0.84 | 0.85 |
| Antiquité | 272 | 0.66 | 0.68 | 0.74 | Mesure | 37 | 0.35 | 0.10 | 0.56 |
| Physique | 265 | 0.75 | 0.76 | 0.82 | Economie domestique | 27 | 0.41 | 0.48 | 0.44 |
| Militaire [...] | 258 | 0.83 | 0.82 | 0.88 | Caractères | 23 | 0.61 | 0.08 | 0.46 |
| Agriculture [...] | 233 | 0.68 | 0.58 | 0.71 | Médailles | 23 | 0.77 | 0.70 | 0.86 |
| Anatomie | 215 | 0.89 | 0.84 | 0.90 | Politique | 23 | 0.15 | 0.22 | 0.53 |
| Belles-lettres - Poésie | 206 | 0.58 | 0.41 | 0.70 | Minéralogie | 22 | 0.38 | 0.39 | 0.70 |
| Mathématiques | 140 | 0.82 | 0.85 | 0.89 | Superstition | 22 | 0.72 | 0.48 | 0.71 |
| Musique | 137 | 0.87 | 0.83 | 0.88 | Spectacle | 9 | 0.33 | 0.46 | 0.61 |

Table 3: F-scores for classes on the test set obtained with SGD + TF-IDF (1), BiLSTM + FastText (2) and BERT Multilingual (3).

Figure 7 presents the confusion matrix for the SGD+TF-IDF model on the test set. We see that most articles in the classes *Arts et métiers* and *Economie domestique* (Domestic economy) were classified as *Métiers*. In the same manner *Mesure* (Measurement), *Minéralogie* (Mineralology), *Pharmacie* (Pharmacy) and *Politique* (Politics) were confused with *Commerce*, *Histoire naturelle* (Natural history), *Médecine - Chirurgie* (Medicine - Surgery) and *Droit - Jurisprudence* (Law), respectively. The semantic similarity between these classes illustrates the difficulty a model has when choosing a "best match." The results confirm that when there is great semantic similarity, the model chooses the best represented class in the

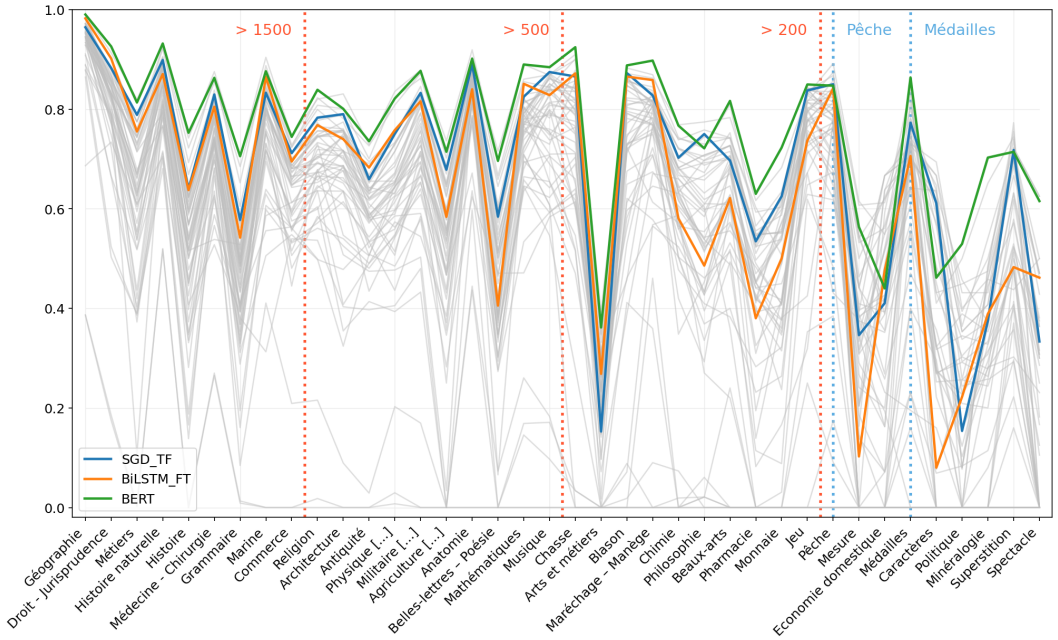



Figure 6: F-scores for classes on the test set obtained with SGD + TF-IDF, BiLSTM + FastText and BERT Multilingual.

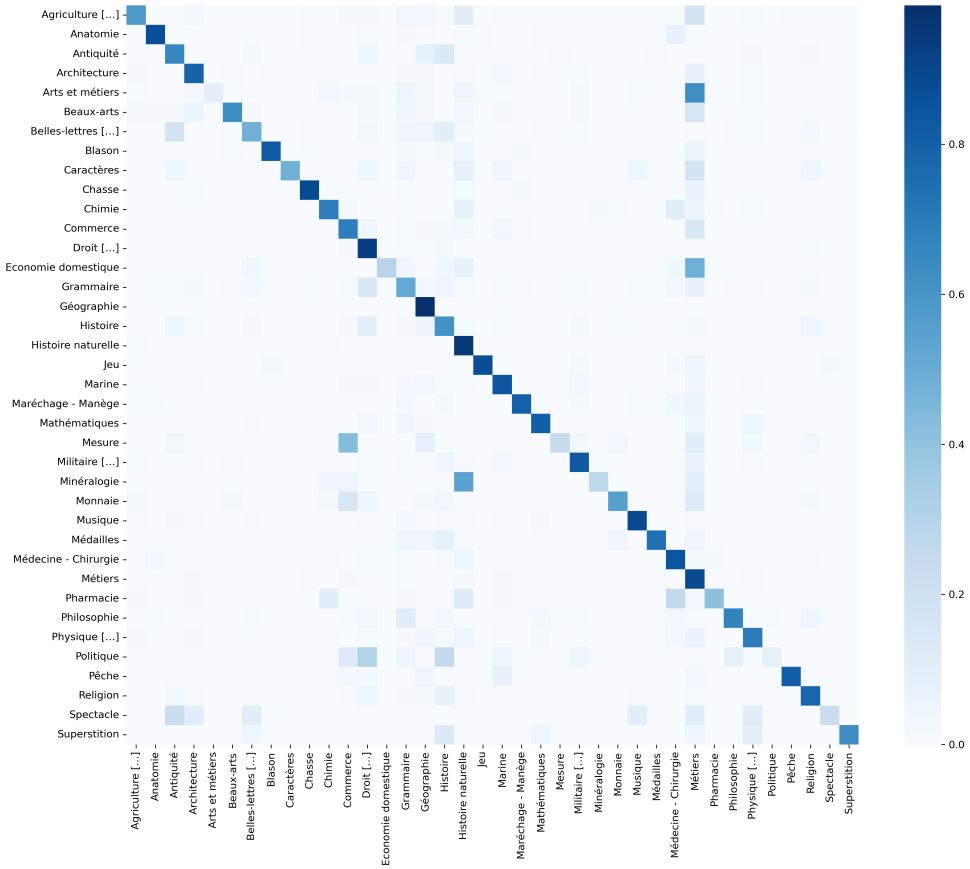



Figure 7: Confusion matrix for SGD+TF-IDF on the test set.

dataset, thereby privileging certain ENCCRE domains that contain more articles.

### *4.3. Exploration of predictions and analysis of similarities between domains*

As we can see, the concept of a "best match" is not not necessarily a relevant measure of success when faced with a complicated text like *EDdA*. In what follows, we explore the ways that probability measures allows us to dig deeper into the results, particularly around the idea that a given article can be classified in multiple domains. It bears remembering that the editors sometimes originally assigned articles to more than one class of knowledge (and ENCCRE duplicated this), but in our experiments we have thus far only applied one domain per article. We can nonetheless explore the ways that the models find great similarity between domains and often rank them very closely for articles that were originally multiclass.

#### *4.3.1. Probability estimates and misclassification evaluation*

We first investigate misclassification by examining false negatives, false positives, and the probability estimates assigned to each class. In this context, misclassification means that the model did not predict the single ENCCRE domain group: our analysis of these results allows us to understand similarities and differences between the domain groups from the perspective of the models. Using the model trained with SGD+TF-IDF, 9,624 out of 11,702 articles (82%) in the test set are well-classified. Of the 2,078 misclassified articles, 1,082 articles have the ENCCRE class associated with the second highest probability (9%). Only 996 articles do not have the correct ENCCRE class in the top two probabilities (8.5%). While there are quantitative results that indicate such "errors" in the model performance, these results actually point to important areas of further investigation. "Misclassified" articles could have 1) originally belonged to multiple classes or 2) logically could belong to one or multiple classes that do or do not match either the original or ENCCRE classes. Therefore, certain misclassification errors can be considered true positives. This should allow us to consolidate the evaluation of the model before using it on other data sets for which we do not have labels, and before conducting further investigations and qualitative analysis.

As a case study, and to limit the number of errors to manually review, we focus on specific classes. First, we are interested in *Géographie*, which is the class with the highest number of articles (see figure 2). According to the f-score (96%, see table 3) and the confusion matrix (see figure 7), this class is among those which are very well classified by the model.

*False negatives: Geography.* Only 39 articles labeled as *Géographie* (out of 2,621 in the test set) were classified with a domain other than *Géographie*. These are considered false negatives. However, 25 of those 39 false negatives do have *Géographie* as the second highest prediction probability.

The prediction probabilities tell us a great deal about the interwoven nature of *EDdA*'s knowledge organization. Indeed, they point to ways that we can "move beyond the editors' original classification scheme and begin to draw out the multi-layered discursive practices that contribute to the rich dialogical texture" (10) of *EDdA*[9]. For a simple example, we turn to the article *Indoustan* (described as part of the Mughal Empire). This was classified originally as *Géog.* and is part of the Géographie ENCCRE domain. *Histoire* is the best prediction (50%), with the "correct" domain prediction of *Géographie* coming next with 47.6%. Articles about toponyms are notoriously filled with a multitude of content that is not necessarily "geographical." The History and Geography domains are particularly close, and it is therefore no surprise that the models confuse these two. If History and Geography were nearly neck-to-neck in the prediction for *Indoustan*, in Diderot's article about the Azores islands Geography has a very low probability (2.7%). Originally *Açores* was unclassified, and was grouped in the *Géographie* domain by ENCCRE. Its most probable classification is *Commerce* (45%), followed by *Métiers* (43%), *Arts et métiers* (8.6%), and lastly *Géographie*.

*Golgotha* was originally classified as *Géographie* and *Théologie* (Theology). We used the ENCCRE domain group for Ggeography as the sole class for this article. However, the results not only reflect the short distance between Geography and History, they also pick up on the theological nature of the article: the top probabilities are 60% *Religion* and 31% *Histoire*. *Géographie* only comes in at 3%. Golgotha refers to the name of place where Jesus Christ was crucified and the article is, true to typical *EDdA* form, a comparison of different early Christian writings about the place which concludes by rejecting more imaginative origins for the name of this place. The author, abbé Mallet, writes in favor of Saint Jerome's proposal which explained that Golgotha was thus named because it was a traditional site of execution where skulls were

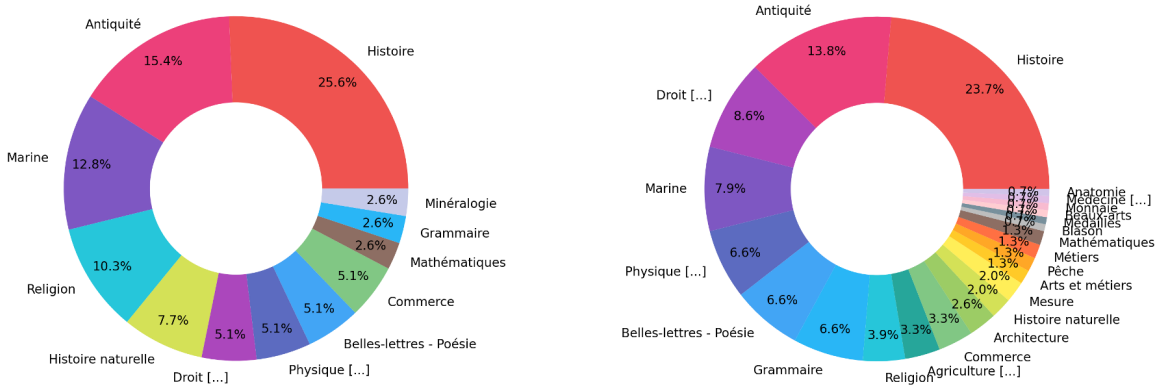
left to decompose. An excellent example of a multi-class article, the model results pointing to History and Religion are useful for understanding the multi-faceted purpose of Geography articles in *EDdA*. Like the topic modeling approach taken by ARTFL, the probabilities that suggest an article covers multiple classes allow us to find content about specific themes in likely as well as unexpected places. Overall, the classification probabilities point to the challenge of working at the level of ENCCRE domains. This choice was made to reduce the number of classes we wished a model to predict, assuming this would improve the results. We review this in more detail in Section 5.

*False positives: Geography.* 152 articles not labeled as *Géographie* by ENCCRE, but for which the model predicts *Géographie* as the most probable class. Starting with a very simple example, we can attempt to explain why the model would classify an article as Geography. The article *Rocher*, for example, is extremely short. It says simply that *un rocher* is the same thing as *un roc* or *une roche*, and includes a cross-reference (*Voyez*) to *Roc*.

> ROCHER, s. m. (Gram.) c'est la même chose que roc & roche. Voyez Roc.

After pre-processing, the model sees "rocher s. m. chose roc roche roc." We hypothesize that the abundance of words related to physical features of the earth (rocks) drives Geography as the most probable classification. The original classification is nevertheless not Geography, it is Grammar! Once again, the model is useful in highlighting thematic content that would otherwise be hidden if we were dependent on the original classifications. Perhaps even more interesting both from a historical and a technical perspective is the fact that there are actually 16 different non-null probabilities for *Rocher*: the lack of information in the article makes classification difficult, even if the top class makes sense. After Geography, the top probabilities are Grammar (17%), Marine (14%), and Law (13%). Based on this manual review, we propose that very short articles are more likely to have a larger number of predicted probabilities than longer articles.

Now we can look at this from a broader perspective. Figure 8 shows the overall class distribution for misclassified articles classified as *Géographie* (false negatives and false positives).




(a) False negatives

(b) False positives

Figure 8: Class distribution for misclassified articles labeled as *Géographie* by ENCCRE

Overall, 12 domains are predicted among 39 false negatives. Figure 8a shows the proportion of each class. Over half of the false negatives are classified as *Histoire*, *Antiquité* or *Marine*. Among the 152 false positives there are 23 possible domains (figure 8b). There are almost twice as many domains for false positives as for false negatives. One way of understanding this is to say that *Géographie* is easy to find across *EDdA*: geographical content is prevalent outside of articles identified as such by ENCCRE. In general, it is difficult for the model to distinguish between *Géographie* and other domains. *Histoire*, *Antiquité*, and, to a lesser extent *Marine* and *Droit* (law) are the top probabilities for *Géographie* articles as frequently as the other way around.

*4.3.2. A graph representation of our classical model*

Matrices and graphs are complementary representations of information. A graph can be transformed into an adjacency matrix and, conversely, a graph can show the dynamics implied by a matrix. Indeed, the weights in a confusion matrix can operate as attraction coefficients between nodes because this is how the confusion matrix was generated: by counting the number of articles from a source class which the model predicted to be in a destination class. In a confusion matrix, a strong identity diagonal is a sign of "stability": an ideal model would map each node to only itself, meaning that all articles in the corresponding class are tagged as belonging to this class. The colored cells outside this diagonal show a proportion of articles incorrectly classified by the model.

Considering the confusion matrix in Figure 7, which we will call $C$ in what follows, most classes are misclassified to some extent by the SGD+TF-IDF model. The graph represented by this matrix would therefore be extremely dense, with most nodes connected to most others. We limit data in the graph therefore to only the top "misclassified" domain for every domain. To do this we derive a transition matrix from the previous confusion matrix. The first step is to empty the cell on the diagonal because we are not interested in the accurate predictions of the model. Then, we compute the maximum value on each row and, for that row, set all cells to zero except the ones where the maximum was reached.

The corresponding graph is displayed in figure 9. It already shows meaningful patterns based on our qualitative knowledge of *EDdA*. For instance, for us, it is logical that *Mathématiques* is most often mistaken with *Physique - [Sciences physico-mathématiques]* (Physics - Physical-mathematical Sciences), and that *Mesure* and *Monnaie* (Currency) both point to *Commerce*. *Médecine - Chirurgie* is linked *Anatomie*, *Chimie* (Chemistry) and *Pharmacie*. And *Métiers* attracts articles from many other classes describing a trade or professional activity such as *Agriculture - Économie rustique* (Agriculture - Rural Economy), *Architecture*, *Arts et métiers*, *Commerce* and *Militaire (Art) - Guerre - Arme* (Military Arts - War - Arms). Representing only the most common misclassifications, the graph is a useful visual tool for understanding the most closely connected ENCCRE domains.

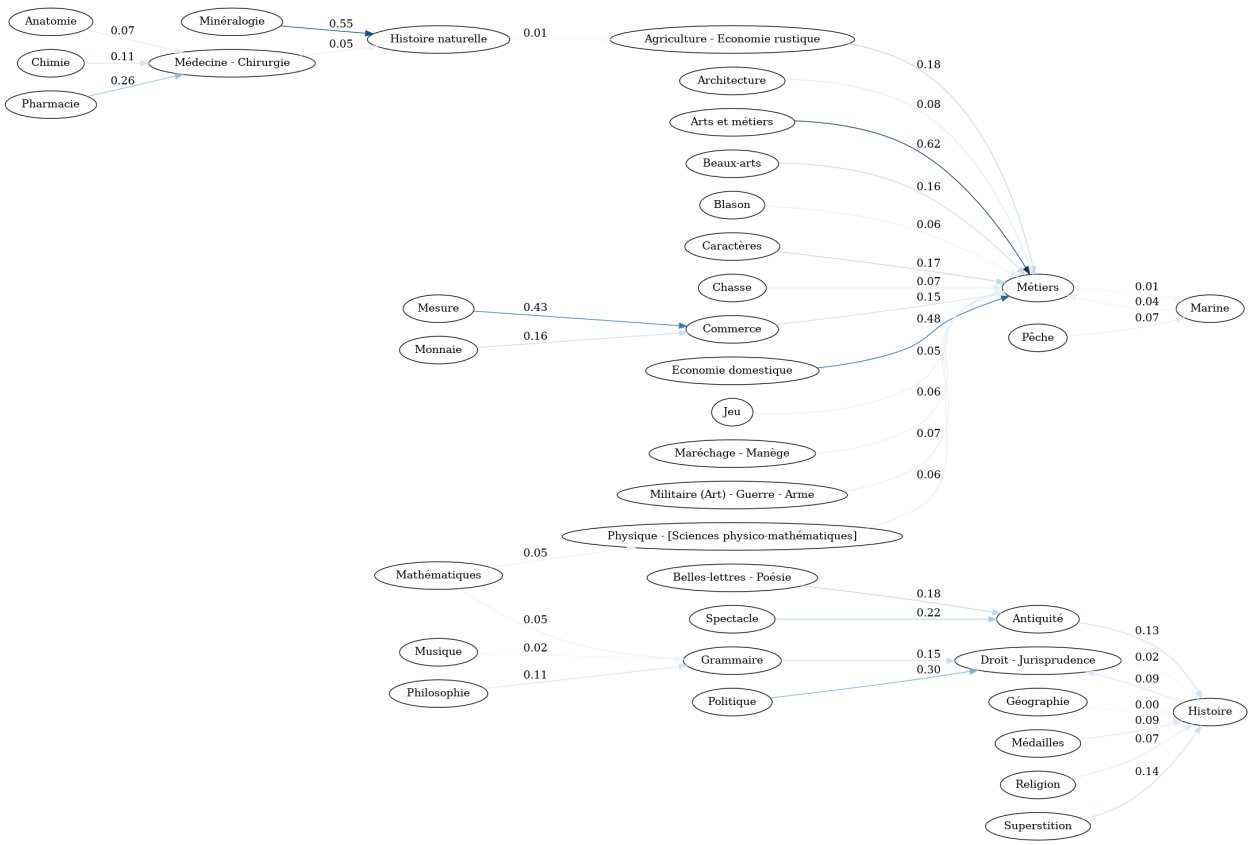



Figure 9: Confusion graph for SGD+TF-IDF on test set.

*4.3.3. Lexical similarities between classes*

Because our data set does not include punctuation and stop words, the model's predictions can only be based on the remaining, not necessarily contiguous words which are mostly nouns, adjectives, adverbs and verbs (as in the *Rocher* example above). Here we investigate the extent to which the lexical similarities between classes can explain the difficulties faced by the model to classify some articles. For this purpose, we study the similarities between the most frequent n-grams for each class and compare them to the confusion graph we have just produced for the model (see Figure 9).

First, we call $\mathcal{V}$ *EDdA*'s lexicon, or, the set of all words occurring in all articles. For any given integer $n \in \mathbb{N}$, an n-gram (the name given to a sequence of $n$ contiguous words) is an element of $\mathcal{V}^{\mathrm{n}}$ which can also be written informally $\mathcal{V} \times \mathcal{V} \times \ldots \mathcal{V}$, $n$ times. To simplify calculations and avoid the disparity caused by the uneven length of text and number of articles in each class, we retain a fixed number of n-grams for each domain (10, 50, or 100) and express this with a new parameter $r$. Using multiple values allows us observe the impact of $r$ and ensure this simplification does not distort the results.

For each of the $r$ most frequent n-grams, we record the words' sequence and the associated number of occurrences. For each class $c$, let us write $\mathcal{T}_{c,n,r}$ the set of the $r$ most frequent n-grams found in the articles labeled with class $c$. Let us also call $|\ldots|_{c,n}$ the function $\mathcal{V}^n \to \mathbb{N}$ which returns the total number of occurrences of a given n-gram within all the articles of class $c$. With this notation, we associate each class $c$ with a "vector" $V_{c,n,r}$:

$$V_{c,n,r} = \{(W, |W|_{c,n}), W \in \mathcal{T}_{c,n,r}\} \tag{1}$$

Despite this compact notation where each $V_{c,n,r}$ has only $r$ components, the vectors they represent inhabit a space of much-higher – but still finite[13] – dimensions. With the classic representation of vectors where only the position of a coefficient within them is relevant and corresponds to a unique vector in a reference base of the vector space considered, each vector would contain $\|\mathcal{V}\|^n$ components, and most of them would be null. Indexing the counts by the corresponding n-grams frees us from this positional notation. As long as we keep this in mind, the vectors still live in a regular Euclidean space where the usual inner products and derived notions of norm and distances are defined. An added benefit of our notation is that by considering all the "classic" components of $V_{c,n,r}$ in $\mathcal{V}^n \setminus \mathcal{T}_{c,n,r}$ to be null, the inner product can still be written as the usual dot product:

$$\langle V_{c_i,n,r}, V_{c_j,n,r} \rangle = \sum_{W \in \mathcal{T}_{c_i,n,r} \cap \mathcal{T}_{c_j,n,r}} |W|_{c_i,n} \times |W|_{c_j,n} \tag{2}$$

To compute this similarity between the class vectors, we consider two metrics: 1) counting keys (in our case the n-grams) in common, and 2) computing a normalized dot product by summing the products of the number of n-grams occurrences they have in common, and dividing by the product of their norm.

Assuming $c_i$ and $c_j$ to be two classes, their *keys* similarity $\langle c_i, c_j \rangle_{n,r,keys}$ can be defined as follows:

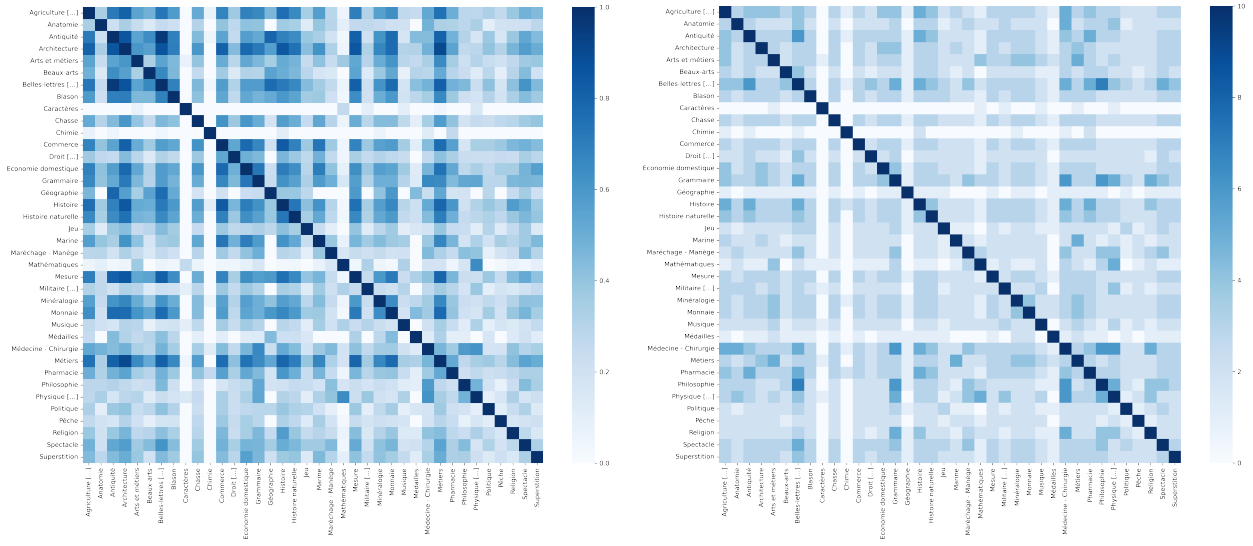
$$\langle c_i, c_j \rangle_{n,r,keys} = \|\mathcal{T}_{c_i,n,r} \cap \mathcal{T}_{c_j,n,r}\| \tag{3}$$

With the same notation, the other metric can be written:

$$\langle c_i, c_j \rangle_{n,r,dot} = \frac{\langle V_{c_i,n,r}, V_{c_j,n,r} \rangle}{\|V_{c_i,n,r}\| \times \|V_{c_j,n,r}\|} \tag{4}$$

This lets us generate similarity matrices to visualize distances implied by those metrics on the space of classes. As expected, the matrices become emptier as the $n$ and $r$ parameters increase, because the probability of two classes sharing a top-ranking n-gram decreases (longer n-grams tend to be more unique, and extending the list of top-ranking n-grams makes it more likely for different n-grams to occur). Their first diagonal (from the upper-left to the bottom right corner) is a line of symmetry because the relations they represent are symmetric: for any two classes $c_i$ and $c_j$, $\langle c_i, c_j \rangle = \langle c_j, c_i \rangle$. Both matrices for a given value of $n$ and $r$ have similar features such as particularly light or dark rows and columns, but the background noise

[13] since there are finitely many n-grams using a finite number of words



(a) Top-10 2-grams (words) with dot product similarity

(b) Top-10 2-grams (words) with keys similarity

Figure 10: Similarity matrices of top-10 2-grams for the classes

is usually more intense in matrices with the *dot* metrics than in ones with the *keys* metrics as can be seen of Figure 10.

This is counter-intuitive, considering the fact that computing the *keys* metrics can be viewed as simply a particular case of the *dot* metrics, where all coefficients in the vectors are set to 1. Indeed one could expect the *dot* product to be systematically lower than or equal to *keys*, in the sense that having different coefficient can only make the vectors less co-linear. But this is actually defeated by the "mass" distribution in the vectors: if two vectors have only few components in common, but all their highest coefficients belong to those components, projecting on their common subspace of $\mathcal{V}^n$ produces vectors of very similar sizes. If, in addition, the distribution of components on these projected vectors has relatively the same shape (the same order when ranked by coefficient and the same ratios between them), then they may have a high-enough dot-product in this subspace to be close to the product of their norms. Therefore, the *dot* metric will yield a similarity close to 1, whereas they share only a few n-grams. In other words, they do not have much in common, but what they do have in common is what mostly defines them. The components they do not share play only a minor part. Of course when they have absolutely no keys in common, their *dot* product will be null.

This phenomenon tends to be more linked to noise for small values of $r$, as many frequent n-grams come from common patterns *EDdA* such as author signatures ("D.J.") or grammatical information at the beginning of the articles ("s. f." for "substantif féminin", or feminine noun). There are nevertheless meaningful occurrences. For example, the similarity between *Histoire* and *Religion* can be seen in Table 4. They share only 16 of their top 100 3-grams (corresponding to a similarity score of 16%), but have a normalized *dot* product of 0.45 because they remain very similar to their projections on the space defined by these 16 common 3-grams (respectively 81% and 78% of their original norm). We notice in particular the occurrences of the 3-grams "depuis long tems" and "avant jesus christ" which one can indeed expect to find in articles dealing with History as well as those dealing with Religion. The 3-gram "trévoux chambers gramme" points to common sources (the *Dictionnaire universel de Trévoux*, printed in multiple editions between 1704-71, and Chambers' *Cyclopaedia* also in multiple editions between 1728-53). We plan further research based on n-gram similarity measures.

As previously with the confusion matrix for the model results, we generate graphs for each of the similarity matrices by selecting only the most important link from one node to another node. The resulting graphs show patterns similar to the ones present in the model's graph (see Figure 9). The graphs for the *keys* metric on top-10 n-grams are too dense to show useful information because the *keys* metric yields an integer between 0 and 10. This necessarily causes collisions on the outputs of the 38 classes, by a simple application

| n-gram | Histoire | Religion |
|---|---|---|
| ('m.', 'pl', 'hist') | 11 | 53 |
| ('s.', 'm.', 'pl') | 119 | 167 |
| ('s.', 'm.', 'hist') | 21 | 51 |
| ('a', '-t', '-il') | 21 | 24 |
| ('depuis', 'long', 'tems') | 24 | 22 |
| ('a', 'donné', 'lieu') | 10 | 15 |
| ('a', 'donné', 'nom') | 15 | 14 |
| ('s.', 'm.', 'terme') | 36 | 12 |
| ('trévoux', 'chambers', 'gramme') | 11 | 10 |
| ('depuis', 'tems', '-là') | 18 | 10 |
| ('m.', 'chevalier', 'jaucourt' | 64 | 11 |
| ('avant', 'jesus', 'christ') | 72 | 10 |
| ('m.', 'pl', 'nom') | 25 | 30 |
| ('s.', 'f.', 'pl') | 25 | 10 |
| ('s.', 'm.', 'nom') | 162 | 40 |
| ('connu', 'sou', 'nom') | 9 | 12 |

Table 4: The most frequent 3-grams shared by *Histoire* and *Religion* classes, and the number of occurrences in both

of the pigeonhole principle. In addition, whereas the model's graph was connected, most similarity graphs have several disconnected components, some even featuring isolated nodes (classes which have absolutely no similarity with any other class).

In contrast to the similarity matrices, which were symmetric by construction, keeping only the strongest connection from one node to another introduces asymmetry because a class $a$ may have its strongest similarity with another class $b$, while class $b$ actually has an even stronger similarity with a third class $c$. This makes the graphs directed: edges have a well-defined source and destination and the opposite edge with reversed destination and source does not exist in general in these graph. However, when they do, they have the same weight since it is the value of their inner product.

With these remarks in mind, most of the interesting features found in the model's graph are also found in one or more of the similarity graphs, such as Figure 11. Apart from the graphs for both top-10 and top-50 1-grams, the nodes for *Mathématiques* and *Physique* are always connected, either uni- or bi-directional. The triangle formed by *Mesure*, *Monnaie*, and *Commerce* is found in all graphs except 2-grams for the *dot* metric. The closeness between *Médecine*, *Anatomie*, *Pharmacie* and *Chimie* is also found in most graphs. Finally, node *Métiers* has at least 5 incoming edges, except for top-10 3-grams, and for top-50 and top-10 2-grams with the *dot* metric, but in these two cases it is directly connected to *Architecture* which was one of the source nodes to *Métiers* in the model's graph, and has itself even more incoming edges so the resemblance is still striking.

These patterns tend to confirm the intuition that our model's predictions would be largely based on n-grams. And in particular, it seems likely that the model is "looking" more at 2- and 3-grams rather than uni-grams.

*4.3.4. Using centrality measures to analyze predicted classifications*

Looking at the simplified graphs of both the model and the classes' n-grams similarities, however, edge weights have very different values which hints at a limit we have envisioned earlier: by choosing to consider only the strongest outgoing edge of each node, we lose information about the fine structure of the graphs. Despite the number of edges which defeats an exhaustive study, other methods such as spectral analysis can give an overall view of the dynamics at stake between the classes by studying the eigenvalues of the model's confusion matrix.

The graph's adjacency matrix coefficients represent a proportion of articles misclassified by the model. If we take, for example, 1,000 Geography articles, the model will predict 985 of them to be indeed *Géographie*,

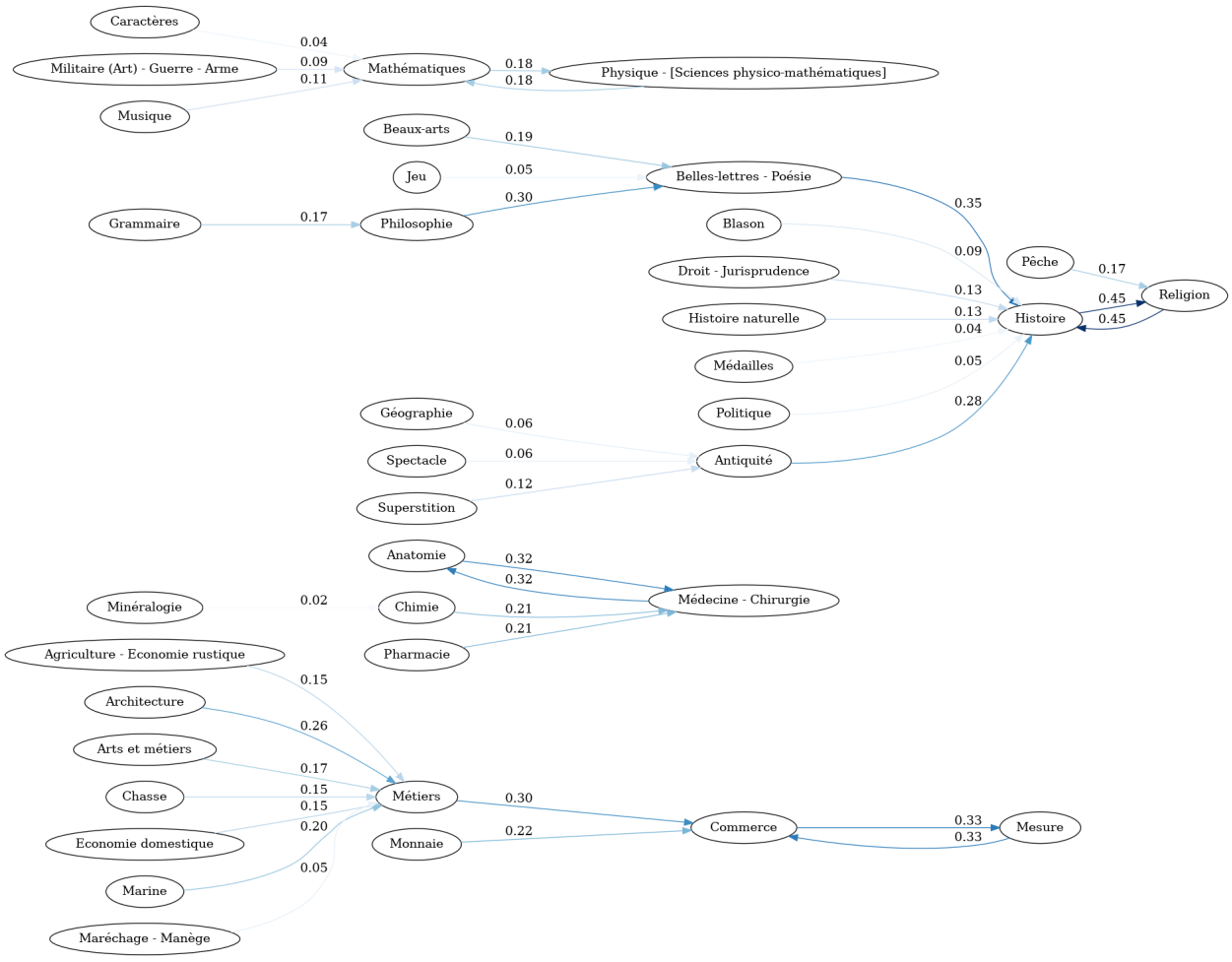



Figure 11: The similarity graph between classes for top-100 3-grams

nearly 4 as *Histoire*[14], 2 of them as *Antiquité*, etc.

This behaviour is a trace of the imperfections in our model, but, again, suggests useful pathways into understanding the complexities of *EDdA*'s content. The graph of a perfectly accurate model would be 38 independent nodes, each pointing to themselves only. The fact that our 1,000 hypothetical articles end up distributed on several nodes other that *Géographie* is a sign that our model is flawed from the perspective of predicting ENCCRE domains. But which nodes "attract" articles from *Géographie*, and in what proportion? The answers provide interesting information about the model's bias. Continuing with the previous thought experiment, one can then consider the fact that if, instead of taking all 1,000 articles from the *Géographie* class, some had been taken – still at random – from the *Histoire* class, then some of these would statistically be mislabelled as *Géographie* by the model. According to our model's confusion matrix, this happens to 5.84% of articles from class *Histoire* so had we taken 900 *Géographie* articles and 100 *Histoire* articles, about 6 would have been subject to that mistake and would have made the opposite trip from *Histoire* to *Géographie*. This is the kind of information that was lost in the previous analysis because *Géographie* is not the class with which articles from class *Histoire* are most frequently mistaken. Of course, there are not only two classes, and the model may likewise make the other classes attract or repel articles to or from both *Géographie* and *Histoire*. The next question is to determine whether adjusting the number of articles in each class would make it possible to find a distribution which is left unchanged by our model: a sort of "preferred" distribution which would resonate with its own bias. The model would still, statistically, make errors on such a distribution. But, while it would be wrong at the level of individual articles, our model would predict the right number of articles for each class.

Making such an adjustment is in fact one way to define a centrality measure for a graph. We represent article distributions using vectors from $\mathbb{R}^{38}$, allowing real numbers as coefficients because we are handling statistical objects, not a particular sample of articles. With this more formal algebraic notation[15] the previous problem is to find

$$\boldsymbol{v} \in \mathbb{R}^{38} \text{ such that } \boldsymbol{v} \cdot C = \boldsymbol{v} \tag{5}$$

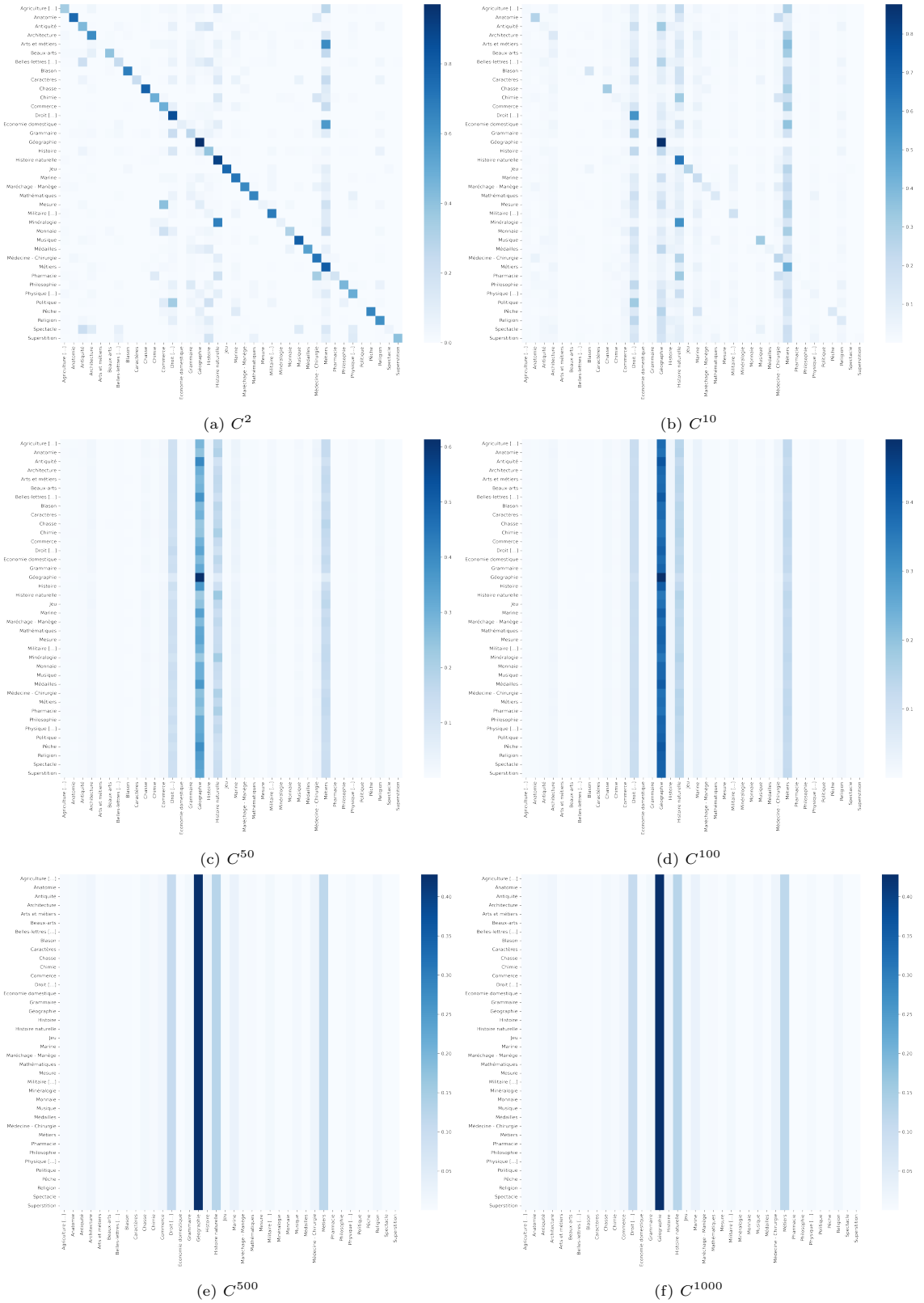
This is the definition of a (left) eigenvector associated to eigenvalue 1. By definition of the confusion matrix, each row represents a probability distribution (the probability for a given article of the corresponding class to be predicted by the model to be in each possible output class): $C$ is right-stochastic. As such, the sum of its coefficients must be 1. Given the rules of matrix products, and calling $\boldsymbol{v_1}$ the vector of $\mathbb{R}^{38}$ where coefficients are equal 1, the property of being right-stochastic can be written

$$C \cdot \boldsymbol{v_1} = \boldsymbol{v_1} \tag{6}$$

This vector full of 1s is hence an eigenvector for matrix $C$, but on the right. The eigenvalues of a matrix $M$ are the roots of its characteristic polynomial $P_M[\lambda] = det(\lambda Id - M)$. Since by definition of the determinant $det(M) = det(M^T)$, we get that $M$ and $M^T$ have the same characteristic polynomial and hence the same eigenvalues. Having found one right eigenvector of $C$ for eigenvalue 1, we know there must exist one left eigenvector for this same eigenvalue, which proves that 5 must admit a solution. A convenient way to compute a value for this solution is provided by applying Gershgorin's circle theorem which, on a stochastic matrix, implies that all eigenvalues must be lower than or equal to 1. Therefore, by iterating the matrix $C$ over and over, values $< 1$ will all tend towards 0 except those equal to 1 (because $\lim_{n\to\infty} \lambda^n = 0$ if $\lambda < 1$). In itself this is not enough to guarantee that the iterates of $C$ will converge because there could be several eigenvectors associated to eigenvalue 1 if it had a multiplicity $> 1$. On the other hand, if it converges, then it is necessarily to the (now only) eigenvector. We can then compute the iterates of $C$, for powers of 2, 10, 100 and 1000 are displayed in Figure 12.

[14] After a particular run of the model, each class can only have a natural number of articles predicted to be in it, but the real numbers in our model describe statistical behaviours, the actual 3.8 for this figure can be interpreted as the fact that if the experiment is repeated and each time 1000 geography articles are taken at random, on average we would get 3.8 predicted in class *Histoire*

[15] Please note that, with our convention of placing true labels (our inputs) in rows and predicted labels (our outputs) in columns, applying the model corresponds to a matrix product with the vector on the left of the matrix.



(a) $C^2$ (b) $C^{10}$

(c) $C^{50}$ (d) $C^{100}$

(e) $C^{500}$ (f) $C^{1000}$

Figure 12: The iterates of the model's confusion matrix $C$

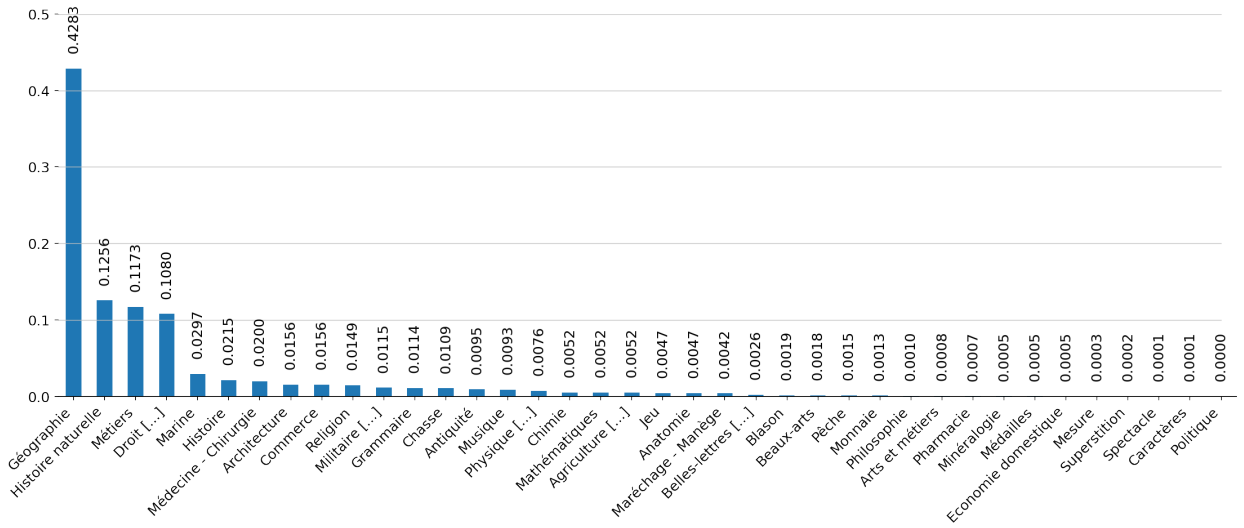



Figure 13: The distribution of centrality measures in the model's graph for each class

A stable distribution appears neatly after a few hundred iterations. This distribution is the eigenvector for $C$ we were looking for and its coefficients correspond to the centrality of each node in the graph, shown in Figure 13.

*Géographie* has highest score: 0.42. This finding suggests that far from playing a peripheral role in the graph, it attracts relatively more articles than other classes. This is true despite the fact that it is not the class with which articles from other classes are most often mistaken. This relatively strong attraction is counterbalanced by the sheer number of articles correctly predicted in the class, which is visible in the excellent f-score reached by the model on this class (0.96). Given the fact previously detailed in Figure 8b that this class receives false positives from 23 other classes, a rather high number but ranking only $5^{\text{th}}$ – far behind *Métiers* attracting articles from 36 classes – it follows that those links must be of some importance to account for such a high centrality. In other words, the *Géographie* class is not a "catch-all" class that any other class can be mistaken with, but, when classes are mistaken with it (e.g. History mistaken for Geography), this occurs frequently. This suggests that the content of Geography articles had a relatively strong connection with the content of these other classes. This depicts geography not as an isolated and very specific domain but instead as a domain adjacent to many others in the scientific and cultural discourse of the Enlightenment. We know that Geography, as the most common class overall in *EDdA* has many roles: it describes places, often giving complete histories; it sneaks in biographies based on the place a person was born or lived in, which these were otherwise banned as article types; it describes peoples (demonyms); and it describes the scientific jargon related to the emerging professional practices of mapmakers, naturalists, and historians. Geography is indeed central to the publication, and to say that other classes are sometimes indistinguishable emphasizes that geographical content infuses a great deal of *EDdA* beyond what was initially classified as *Géographie*.

The most immediately visible feature of the distribution of centrality measures shown in Figure 13 is that its shape differs from that of the distribution of the number of articles per class, with three almost flat groups: 1) the most central class, 2) the three secondary classes next, and 3) all the others in a long tail. Moreover, while the classes ranked by centrality tend to globally have the same positions as when ranked by number of articles as displayed in Figure 2, with the most populated classes usually having higher centrality ranks and the least populated ones ranking lower, the order is nonetheless locally altered. For instance, *Histoire naturelle* ranks $2^{\text{nd}}$ in centrality, but only $4^{\text{th}}$ in terms of its number of articles. Similarly, *Politique* is the least central class, but it has more than twice as many articles as *Spectacle*, the least populated class. Although the class centrality measure exhibits a loose dependency on sample size, the differences we have underlined show that size cannot explain their distribution in detail. This hints at the fact that the model identifies some classes more easily because of distinctive lexical patterns.

## 5. Discussion

### *5.1. Clustering experiment*

As a complement to the supervised classification methods, we have performed a preliminary experiment with unsupervised learning. This will allow us to further study the relationship between article content, original classifications, ENCCRE domains, ARTFL classifications, expert-determined classes, and machine-generated predictions. We focus on clustering methods in order to automatically group articles based on their similarity, with no regard for any kind of label. The similarity is based on a distance calculation between article vectors. For this preliminary experiment we tested KMeans clustering with the TF-IDF vector representation. The first experiment consisted of training a clustering model to build 38 clusters (corresponding to the number of ENCCRE domains). Poor results led us to look for the optimal number of clusters according to the Silhouette method [28]. The results suggest 36 clusters, however re-running this model produced similar results.

Figure 14 shows a heatmap of the normalized distribution of clusters per domain. Many clusters contain articles from several domains, and cluster 0 regroups articles from every domain. Moreover, for many clusters, the proportion is very high. Likewise, many classes are spread across several clusters. This is particularly true for *Géographie*, *Histoire naturelle* and *Arts et métiers*. While the results are complex, the analysis of clusters is still useful. For example *cluster 22* groups together articles labeled as *Belles-lettres - Poésie*, *Histoire*, *Médailles*, *Religion* and *Superstition*. These categories have topic similarities which often belie their original and ENCCRE classifications. Similarly, *cluster 7* groups articles labeled as *Médecine - Chirurgie*, *Anatomie*, *Physique - [Sciences physico-mathématiques]* and *Pharmacie*, which also reflects a "super group" of medical articles that span these ENCCRE domains. *Cluster 30* groups together *Commerce* and *Mesure*, likely based on writing about money and measurement using numbers. However, some clusters seem to be more homogeneous in comparison to ENCCRE's classification, for example, *cluster 10* contains 96.46% articles labeled as *Géographie* and *cluster 34* is 99.05% *Histoire naturelle*.

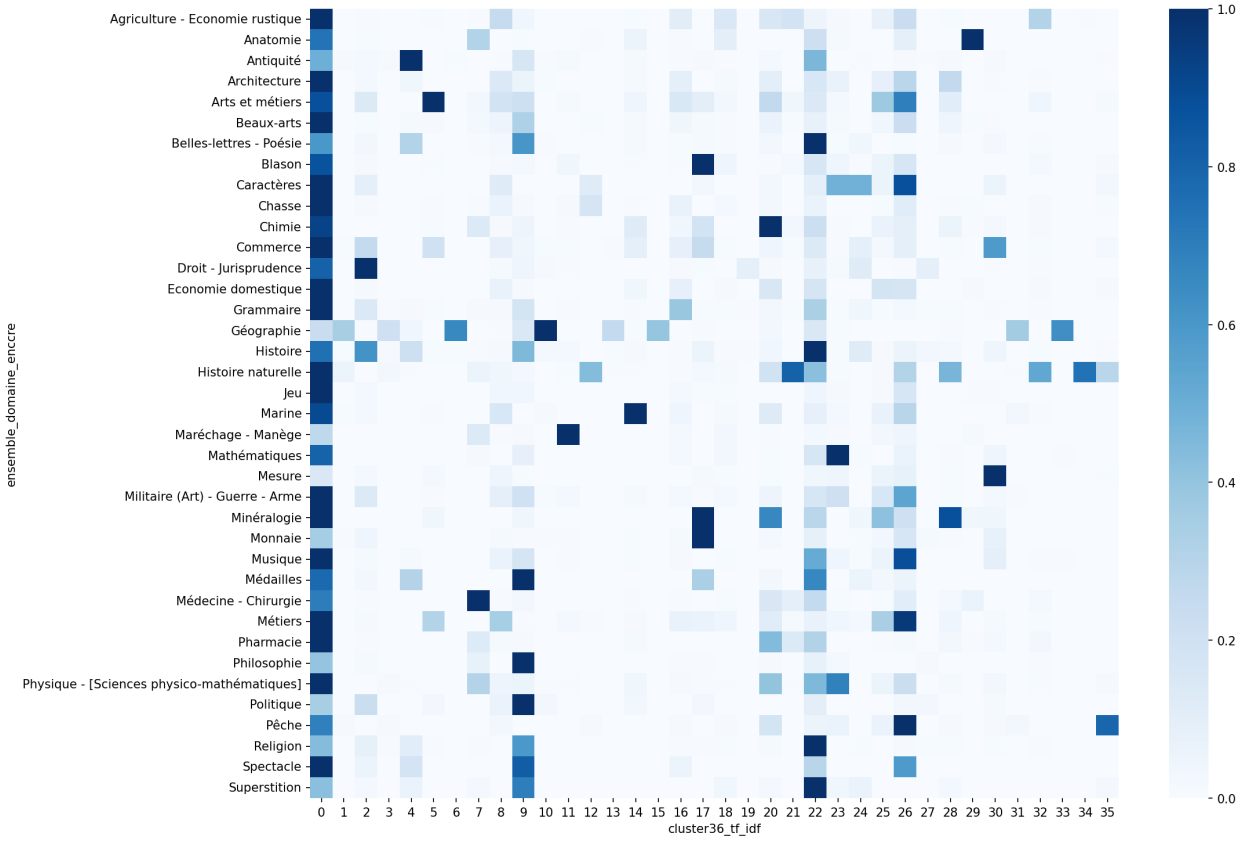



Figure 14: Normalized distribution of clusters per class.

Figure 15 shows the distribution of classes per cluster. In addition to the heatmap (see Figure 14), this diagram also shows the number of articles per cluster. *Cluster 0* has more than 14 000 articles while almost all the other clusters are below 2,000 articles. This huge gap between one cluster over the others highlights the fact that the clustering task is difficult on our dataset. While the results are difficult to interpret at this stage, we see the heterogeneity of classes in almost every cluster (with the exceptions described above).

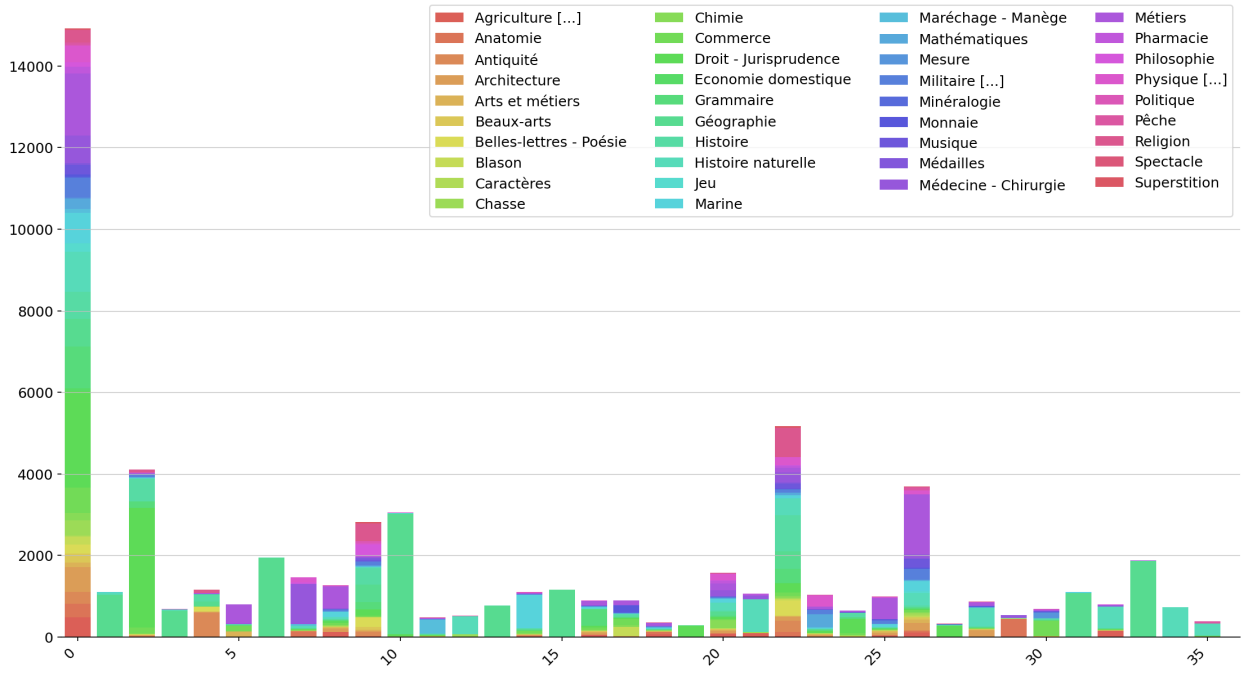



Figure 15: Classes per cluster

In the same manner, Figure 16 shows the distribution of clusters per class. This highlights the same observations made with the heatmap chart (see Figure 14) and leads us to the same conclusion, these preliminary results of unsupervised clustering are not easy to read according to the existing classification system.

Furthermore, among the 38 classes, only 15 appear as the main class of at least one cluster (see Table 5). 9 clusters have a main class above 90% (clusters 1, 6, 10, 13, 15, 19, 31, 33 et 34) and for those, only three classes are listed (7 clusters for *Géographie*, 1 for *Droit - Jurisprudence* and 1 for *Histoire naturelle*) In addition to the 7 clusters grouping together *Géographie* articles, the *Géographie* class is spread on 31 different clusters.

| | Main class | Pct | | Main class | Pct | | Main class | Pct |
|---|---|---|---|---|---|---|---|---|
| 0 | Droit - Jurisprudence | 15.85 % | 12 | Histoire naturelle | 83.24 % | 24 | Droit - Jurisprudence | 52.67 % |
| 1 | Géographie | 91.95 % | 13 | Géographie | 96.53 % | 25 | Métiers | 51.76 % |
| 2 | Droit - Jurisprudence | 71.54 % | 14 | Marine | 74.93 % | 26 | Métiers | 39.85 % |
| 3 | Géographie | 89.03 % | 15 | Géographie | 100.00 % | 27 | Droit - Jurisprudence | 82.77 % |
| 4 | Antiquité | 51.38 % | 16 | Grammaire | 44.65 % | 28 | Histoire naturelle | 52.82 % |
| 5 | Métiers | 58.78 % | 17 | Blason | 24.06 % | 29 | Anatomie | 82.93 % |
| 6 | Géographie | 99.28 % | 18 | Métiers | 20.51 % | 30 | Commerce | 52.82 % |
| 7 | Médecine - Chirurgie | 66.10 % | 19 | Droit - Jurisprudence | 95.50 % | 31 | Géographie | 95.50 % |
| 8 | Métiers | 42.17 % | 20 | Chimie | 12.33 % | 32 | Histoire naturelle | 65.09 % |
| 9 | Géographie | 15.25 % | 21 | Histoire naturelle | 74.48 % | 33 | Géographie | 99.47 % |
| 10 | Géographie | 96.46 % | 22 | Histoire | 16.99 % | 34 | Histoire naturelle | 99.05 % |
| 11 | Maréchage - Manège | 70.80 % | 23 | Mathématiques | 30.95 % | 35 | Histoire naturelle | 71.61 % |

Table 5: Percentage of the top domain for each cluster.

It is difficult to draw conclusions from these results, and before going further a more extensive qualitative

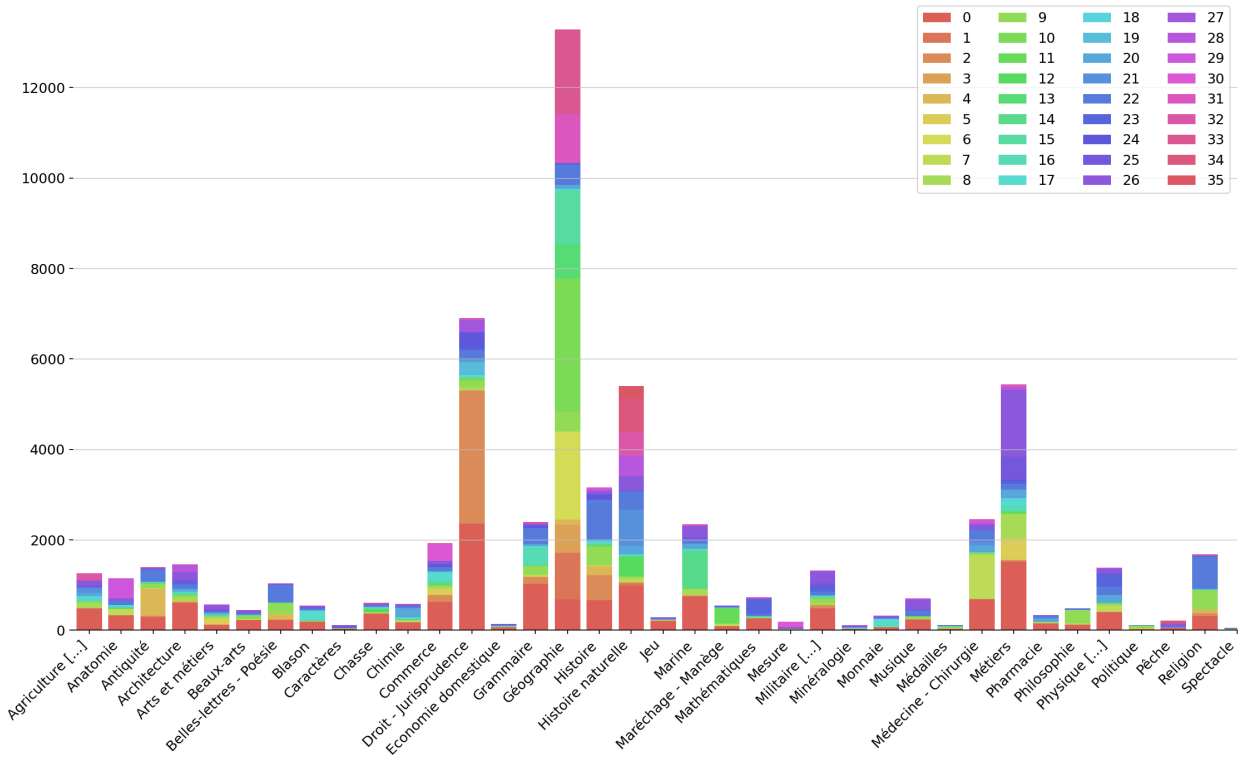



Figure 16: Clusters per class

analysis is needed to see if unsupervised learning may be useful as a complementary method for new automatic classification of *EDdA* and other encyclopedia articles. However, at this stage we can say a little about the promise of these methods. Clustering is, like LDA topic modeling, a way of unpacking the contents of our data without assuming that a particular set of classes is the "correct" way of organizing knowledge. Our choice to mirror the number of ENCCRE classes we predicted with the name number of clusters is only a first step in exploring potential uses of clustering. Next, we plan to test working up from very small numbers of clusters, and to manually evaluate the contents of the clusters.

Looking back at d'Alembert's *Système figuré* (see Figure 17), and considering our knowledge about the proximity of domains like History and Geography, or History and Religion, it does seem incongruous to see Geography alongside *Uranographie* (celestial mapping) and *Hydrographie* (marine mapping) under the broader class of *Cosmographie*. This mathematically-inspired ordering positions Geography as mapping, rather than Geography as knowledge of the the peoples and places of the world. And yet, it is the latter that clearly shows through in the content that was in practice labeled as geographic. The opportunity to scrutinize the content of encyclopedias, to compare different knowledge organization systems–both human- and machine-generated–gives us a chance to move back and forth between the qualitative, close reading of the articles and the quantitative understanding of their relationships. Using prediction probabilities, measurements of lexical similarity, network metrics, and clustering, we have put forward a range of methods for classifying and re-classifying the content of encyclopedias.

## 6. Conclusion

In this article we have presented a comparison of different supervised classification methods in order to automatically classify encyclopedia articles. Using *EDdA* and the domain groups proposed by ENCCRE, we have trained a variety of models. We tested methods for creating word vectors (bag-of-words, tf-idf, doc2vec and FastText) combined with ML methods (Naive Bayes, logistic regression, random forest, SGD, and SVM) as well as deep learning ones (CNN and LSTM). Finally, we fine-tuned BERT and CamemBERT pre-trained

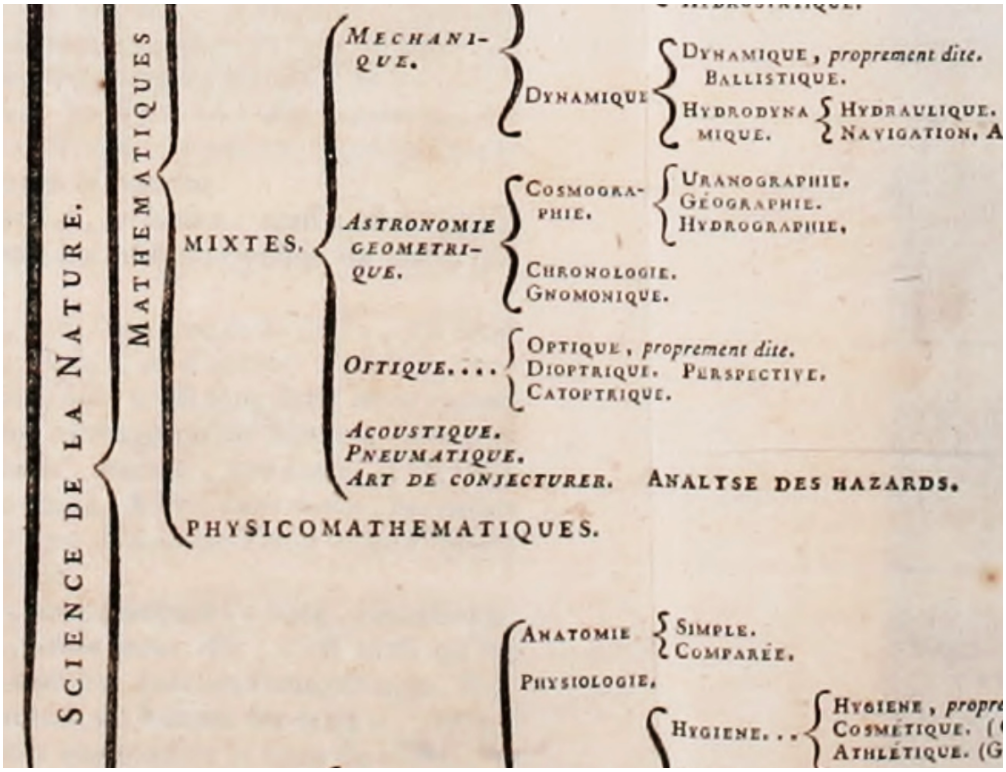



Figure 17: Detail showing *Géographie* in the 1751 *Système figuré des connoissances humaines* (Figurative system of human knowledge). Source: Wikipedia

transformer language models for the classification task. We evaluated these experiments using test sets of different configurations: sampling the number of articles in order to establish an equilibrium among the domain groups and selecting hyperparameters for the classification algorithms. For our classification experiment, the results are encouraging: 31/38 domain groups have an f-score of over 70%. The contextual pre-trained language models obtained comparable results to SGD and logistic regression methods based on TF-IDF word vectors (86% mean f-score against 81%). Our results confirm that BERT models obtain better or comparable results with less training data. Finally, our results show the difficulty of distinguishing between certain domain groups. This is partly because of their semantic similarity, but mostly because of the small number of articles in those domain groups. BERT models obtain the best results for these domain groups, but they remain difficult to isolate from more dominant domain groups. However, it is important to remember that poor performance is sometimes a useful signal and points to meaningful areas of qualitative research. For example, we used misclassifications of domain groups to examine ENCCRE domain groupings and the original classifications by the authors and editors of *EDdA*. This kind of investigation suggests future experiments to test different groupings of classifications, where automatic methods may perform better. An iterative approach like this benefits researchers from multiple disciplines, and allows us to learn from one another.

Our objective is now to apply this method of predicting classes for other historical encyclopedia articles in French. One of the next texts we will work with, *La Grande Encyclopédie* (1886-1902), does not include article classifications. Others do include classifications, but using different classes. For the former, we seek to predict classes in order to use that as metadata about articles for further research. As with *EDdA*, for encyclopedias that already employ a classification system, we can use this labeled data to continue to improve our models. One of the challenges will be to evaluate how models perform on encyclopedias from different time periods: e.g. given a model that performs well on articles written in the mid-eighteenth century, how does it perform on articles written in the late nineteenth century? Another challenge we face is accounting for articles that have multiple classifications (e.g. Geography and History). Collecting more training data from additional encyclopedias that contain multi-class articles will help to improve results in the future. Once

we have a corpus of encyclopedias with article classifications (both predicted and original) we will examine changes in the organization of knowledge over time. This allows us to better understand the evolution of the encyclopedic genre in France (and beyond), and to compare encyclopedic representations of knowledge to the ways those subjects are discussed in other genres like novels, newspapers, political texts, or non-fiction books.

**Acknowledgements**

The authors are grateful to the ASLAN project (ANR-10-LABX-0081) of the Université de Lyon, for its financial support within the French program "Investments for the Future" operated by the National Research Agency (ANR).

**References**

[1] M. Foucault, The order of things: An archaeology of the human sciences, Tavistock.

[2] G. C. Bowker, S. L. Star, Sorting Things Out: Classification and Its Consequences, MIT Press.

[3] A. Blair, Too much to know: managing scholarly information before the modern age, Yale University Press.

[4] C. Wellmon, Organizing Enlightenment: Information Overload and the Invention of the Modern Research University, JHU Press.

[5] M. Groult (Ed.), L'encyclopédie ou la création des disciplines, CNRS Éditions.

[6] L. Holmberg, The Maurists' unfinished encyclopedia, Voltaire Foundation.

[7] J. d'Alembert, Preliminary discourse, in: The Encyclopedia of Diderot & d'Alembert Collaborative Translation Project, Michigan Publishing, University of Michigan Library.
URL http://hdl.handle.net/2027/spo.did2222.0001.083

[8] G. Recchia, E. Jones, P. Nulty, J. Regan, P. de Bolla, Tracing shifting conceptual vocabularies through time, in: P. Ciancarini, F. Poggi, M. Horridge, J. Zhao, T. Groza, M. C. Suarez-Figueroa, M. d'Aquin, V. Presutti (Eds.), Knowledge Engineering and Knowledge Management, Springer International Publishing, pp. 19–28. doi:10.1007/978-3-319-58694-6_2.

[9] G. Roe, C. Gladstone, R. Morrissey, Discourses and disciplines in the enlightenment: Topic modeling the french encyclopédie, Frontiers in Digital Humanities 2 (2016) 8.

[10] R. Horton, R. Morrissey, M. Olsen, G. Roe, R. Voyer, et al., Mining eighteenth century ontologies: machine learning and knowledge classification in the encyclopédie 3 (2).

[11] S. Grabus, J. Greenberg, P. Logan, J. Boone, Representing aboutness: Automatically indexing 19th-century encyclopedia britannica entries 7 (1) 138–148. doi:10.7152/nasko.v7i1.15635.
URL https://journals.lib.washington.edu/index.php/nasko/article/view/15635

[12] A. Peterson, A. Spirling, Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems 26 (1) 120–128. doi:10.1017/pan.2017.39.

[13] J. Guldi, Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change 60 (1) 1–33. doi:10.1353/tech.2019.0000.
URL https://muse.jhu.edu/article/719944

[14] A. T. J. Barron, J. Huang, R. L. Spang, S. DeDeo, Individuals, institutions, and innovation in the debates of the french revolution 115 (18) 4607–4612. doi:10.1073/pnas.1717729115.
URL https://www.pnas.org/content/115/18/4607

[15] T. Underwood, The historical significance of textual distances (2018). doi:10.48550/ARXIV.1807.00181.
URL https://arxiv.org/abs/1807.00181

[16] T. Underwood, The life cycles of genres 1 (1) 11061. doi:10.22148/16.005.
URL https://culturalanalytics.org/article/11061-the-life-cycles-of-genres

[17] T. Underwood, Machine learning and human perspective, PMLA/Publications of the Modern Language Association of America 135 (1) (2020) 92–109. doi:10.1632/pmla.2020.135.1.92.

[18] E. Bender, The #benderrule: On naming the languages we study and why it matters, The Gradient.

[19] I. Galina Russell, Geographical and linguistic diversity in the digital humanities 29 (3) 307–316. doi:10.1093/llc/fqu005.
URL https://doi.org/10.1093/llc/fqu005

[20] P. J. Spence, R. Brandao, Towards language sensitivity and diversity in the digital humanities 11 (1). doi:10.16995/dscn.8098.
URL https://www.digitalstudies.org/article/id/8098/

[21] A. Guilbaud, L'enccre, édition numérique collaborative et critique de l'encyclopédie, Recherches sur Diderot et sur l'Encyclopédie (52) (2017) 5–22.

[22] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., USA, 1986.

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[24] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.

[26] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de La Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, arXiv preprint arXiv:1911.03894.

[27] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[28] K. R. Shahapure, C. Nicholas, Cluster quality analysis using silhouette score, in: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), 2020, pp. 747–748. doi:10.1109/DSAA49011.2020.00096.