Collège Sciences et Technologie

Text classification of data citations in Open access papers from EuropePMC

Author :

THOUVENIN Arthur

Bioinformatic master of Bordeaux (from genome to ecosytems)

Supervisors :

Dr. McENTYRE Johanna (Team leader)

Dr. YANG Xiao (Text miner)

Dr. VENKATESAN Aravind (Senior Data Scientist)

Hinxton, Wellcome Genome Campus, EMBL-EBI
31-08-2019

Summary

Acknowledgements

For giving me the opportunity to be a part of the great team of EuropePMC, I would like to thank Dr Johanna McEntyre for her advises and her accompaniment throughout this internship.

For his precious advises concerning the annotation part and his accompaniment I would like to thank Aravind Venkatesan.

I would like to thank warmly Xiao Yang, indeed he is my direct supervisor during this internship, he gives me really good advice. He teaches me a lot and also I would like to thank him for his availability at every moment.

Those three persons were really important for me during this project, indeed they gave me guidelines and their precious helps whenever I needed it. They have been really nice with me and pushing me to gave my best on this project. Therefore, one more time I would like to thank them really much.

Obviously I would like to thank the EuropePMC team, Lynne Faulk, Yogmatee Roochun, Mariia Levchenko and all other members for all the goods moments that they gave, without those this internship would not have been the same.

For his precious help I would like to thank Awais Athar, he gave me his experienced thoughts about this project and gave me really good advice.

In the end I would like to thank my family and especially Claire Pistien for her support and advises during the whole duration of this project.

Lexicon

- *AI* : Artificial Intelligence
- *DL* : Deep Learning
- *CNN* : Convolutionnal neural network
- *ComplementNB* : Complement Naive Bayes
- *EPMC or EuropePMC* : Europe PubMed Central
- *FN* : False negative
- *FP* : False positive
- *GaussianNB* : Gaussian Naive Bayes
- *LR* : Logistic Regression
- *LSTM* : Long Short Term Memory
- *ML* : Machine learning
- *MultinomialNB* : Multinomial Naive Bayes
- *NLP* : Natural Language Processing
- *OA* : Open Access
- *PMCID* : Identifier for an article from EuropePMC
- *RF* : Random Forest
- *SimpleNN* : Simple Neural network
- *SVM* : Support Vector Machines
- *Tfidf* : Term frequency-inverse document frequency
- *TN* : True negative
- *TP* : True positive

Chapter 1

Introduction

> *“ **Given my background, I would start an AI company whose goal would be to teach computers how to read, so that they can absorb and understand all the written knowledge of the world,** ”*
> *Bill Gates told David Rubenstein on Monday 24th June 2019 (CNBC).*

Artificial intelligence (AI) is a very popular field nowadays. It is found more and more in our daily lives, between our smart-phones, airport surveillance, weather predictions, and even in the art. It is the same for the study of the texts. Indeed, natural language processing (NLP) is a well studied field. It brings together three main areas: linguistics, computer science and especially artificial intelligence. The purpose of NLP's methods is to make a machine understand the meaning of a text. Thus, as Bill Gates says, it would be possible for a machine to accumulate all the knowledge produced by man.

These methods are nowadays used in many cases such as : automatic translation, text generation, automatic text summarization but especially in text classification and sentiment analysis. Some examples of these applications have been observed, including tweet analysis on *Twitter* regarding the USA elections in 2017, but also the analysis of reviews on *Amazon* products.

To succeed in analyzing these texts with such good results, it is necessary to look at the artificial intelligence part of the NLP. First of all, it is important to understand that the AI is a very vast field, this one grouping together a set of automatic learning methods called machine learning (ML). It is a set of algorithms and statistical models that allow a machine to learn from data. These methods can be used in two ways. First, a set of unlabelled data can be provided to the model for training, this is called unsupervised learning. The other method, however, is to annotate the dataset by an expert, the model will try to reproduce what it has learned from this labelled dataset, this is called supervised learning.

There is also another set of methods in the machine learning area called Deep Learning (DL). This one is inspired by the functioning of the brain by working with networks of neurons. Each neuron processes small amounts of information, then another will take its conclusion and associate it with another's conclusion to draw its own conclusion and so on. The model will summarize the information and learn only the essentials.

Those technologies occupies a very large place in the biomedical field. First of all, this allows rapid and automatic analysis of bio-medical documents as well as their auto-completion. Thus, retrieval of relevant information from clinical examination reports can be done automatically and associated with other data, for example scanner images for more accurate diagnostics. But one of the most studied applications in the biomedical field is the exploration and extraction of information from scientific articles. Indeed, these documents contain a lot of information. In addition, they form a network of knowledge through multiple citations to support facts.

In the field of life science, Europe PubMed Central (EPMC or EuropePMC) [1,2] has established itself. It is an open access repository that contains more than 35 millions Abstracts, 5 millions Full text articles, 4 millions patents and lots of other open access documents.

One of the goals of EuropePMC is to promote the reuse of data cited in scientific documents. It's why EPMC publicly provides its document annotation tool called SciLite [3]. This tool was created to solve an important problem which is the need to link literature and underlying data. The platform integrates text-mined annotations from different sources and overlays those outputs on research articles. When open an article in EuropePMC platform, it results in highlighted terms. Various bio-medical concepts are text-mined at EuropePMC, such as genes, proteins, diseases, chemicals, etc.

Regarding the need of linking data it is interesting to look at data citations text mined by this platform.

Data citations are accession numbers that correspond to unique identifier given to any molecules stored in database related to biology. It can be a way to follow different molecules versions. For example, a SNP can have multiple accession numbers because the sequence has been modified through years but these can't be associated with another molecule. This is why it is relevant to study those. This is one of the links between scientific literature and bio-medical data. In scientific papers, there are a lot of knowledge about those data, it is therefore important to understand how they are cited and for which reasons. Indeed, it is commonly known in scientific community that data or paper should be cited to support findings. But there is not that much knowledge on how data citations are used in a scientific paper.

First of all, a sentiment analysis on these citations seemed relevant, because this kind of analysis is common for paper citations analysis. However, it is very difficult to find an opinion in data citation because it is only data and it is therefore difficult to express a positive or negative opinion. Another way to study those citations is to assign categories. This could be achieve with text classification and as a good analysis need a lot of data to perform, it should be done with automatization.

Nowadays, to the best of our knowledge, there is no studies about text classification of data citations although there is some about data citations. Most of those are focused on data reuse but there is no clue that data could be cited in another way. H. Mooney and MP. Newton have published an interesting paper named : "*The anatomy of a data citation: Discovery, reuse, and credit*" [4], however, this one mainly focuses on the data reuse. This paper is at least relevant because of its title, it suggests that data citation are not always made because they are reused. Indeed, if data citation are not made only because they are used, it will induce a bias in the study of data reuse. In these studies, data citations are important to give credit to findings, but those citations needs some rules [5] and also it showed with multiple features how data are reused [6]. It is therefore important to know how data citations are made and for which reason.

The previously mentioned study brings another interesting feature for analysis of data reuse, the category of data citation, as this has not been studied before. It is a brand new approach of data citations. However, this approach was designed to paper citations and some categories has been established for those. In the article of Jurgens et al. [7], six categories has been proposed for paper citations (”Paper” corresponds to the citation of a scientific article) :

- Background *("Paper" provides relevant information for this domain)*
- Motivation *("Paper" illustrates need for data, goals, methods, etc.)*
- Uses *(Uses data, methods, etc. from "Paper")*
- Extension *(Extends "Paper"'s data, methods, etc.)*
- Comparison or Contrasts *(Express similarity/differences to "Paper")*
- Future *("Paper" is a potential avenue for future work)*

Additionaly, training a classifier requires some data to perform an automatic classification. It is critical to develop a labelled data set. There are some similar and existing data-sets like :

- *Sentiment140* [8] : This dataset is made of tweets tagged with sentiments.

- *Stanford Sentiment Treebank* [9] : This dataset is made of movie reviews tagged with sentiments.

- *Multi-domain sentiment analysis dataset (version 2.0)* [10] : This dataset is made of products reviews from Amazon tagged with sentiments.

- *Paper Reviews Data Set (PeerRead)* [11] : This dataset is made of scientific articles peer reviews tagged with sentiments.

- *Awais Athar - Citation Sentiment Corpus* [12] : This dataset is made of citations of scientific articles tagged with sentiments. This one is the most closer of this studies

This project is based on those datasets and papers, the paper of Jurgens et al. [7] was especially a great inspiration.

In the end this study will help analysts understand data reuse in a better way, help annotators and curators, but also librarians and publishers to understand in quicker way how authors have performed their studies.
Indeed, here, thanks to established categories and a good classifier, we shown that each data citation is not necessary a reuse of data. But also that those categories can describe a database. If it is a recent one or curated or even a database containing knowledge and not necessary biological data.

Then, in the next chapter, it is explained how a dataset was created, how data citation categories were established and how a model was formed and tested. Finally, an analysis of model predictions is presented.

Chapter 2

Material and methods

2.1 Dataset determination

2.1.1 Existing datasets

For an automatic classification, a training dataset is needed to establish a predictive model. It is common to seek to use existing datasets, moreover this reuse avoids the duplication of data and improve the reproducibility of research. Therefore datasets presented in the introduction have been analyzed.

Thus *Sentiment140* [8], *Stanford Sentiment Treebank* [9] and *Multi-domain sentiment analysis dataset* [10] concern respectively a social network language, a cinematographic language and a set of language from different areas, this does not correspond specifically to the scientific language studied here. Indeed, the type of language learned by the model is really important. It has been shown that it plays a key role in model training [13]. Certain words or expressions can have a different meaning from one language to another, and sometimes even within the same language.

The *Paper Reviews Data Set* [11] dataset is a little more in line with the expectations of this study as this dataset contains reviews of scientific papers around the mathematics and computing area, however EPMC focuses on biomedical documentation, making it irrelevant.

The most interesting thing seemed to be *Awais Athar - Quote Sense Corpus* [12], however, being annotated on the basis of sentiment analysis, it was not sufficiently adapted to the set objectives.

Therefore, a dataset is required for the purposed of this project. The following sections describe the creation of our dataset.

2.1.2 Dataset creation

Here, the data were collected from EuropePMC, and correspond to *Open Access* articles containing at least one citation referring to a dataset (text mined by EPMC).

Features

To build this data structure, it is first necessary to easily find each citation, and this was possible with two elements : the PMCID of the corresponding article (an unique EPMC identifier for each biomedical document) and the accession number of interest.
Subsequently, the citation itself was necessary. There are two major approaches to NLP (natural language processing). First, the *document-level*, in this case the data citation will be studied in the whole scientific paper. It's a more general approach indeed it will learn from a complete scientific paper and all mention of the citation of interest. However this relatively long and complicated approach was not retained. The second, called *sentence-level*, is easier and more specific. In this case only the sentence containing the data citation of interest will be studied. It means that conclusions are made mention by mention.

This work was focused on *sentence-level.* Indeed, the context of citations can significantly improve results of classification [14]. Thus, the sentence preceding the one containing the accession number has been extracted (*PreCitation*), as well as the two sentences following that of interest (*PostCitation*). However, the section change may represent a bias. Indeed, providing the model with a sentence from a different section can create unnecessary noise. Therefore, sentences that precede or follow the sentence of interest have not always been used. Thus, during the feature engineering, the *PreCitation* feature can contain between 0 and 1 sentence while the *PostCitation* feature can be up to 2 sentences.

The section(*Section*) and the sub-type(*SubType*) of the citation have also been extracted to provide additional information to the model. The sub-type corresponds to the original database of the cited data then involving the data type (example: *ENA* = nucleotide sequences). Sometimes the section for some citations was "*Article*", meaning EuropePMC could not determine the section because it was not provided by the article, in the end "*Article*" is assigned as the default section name.

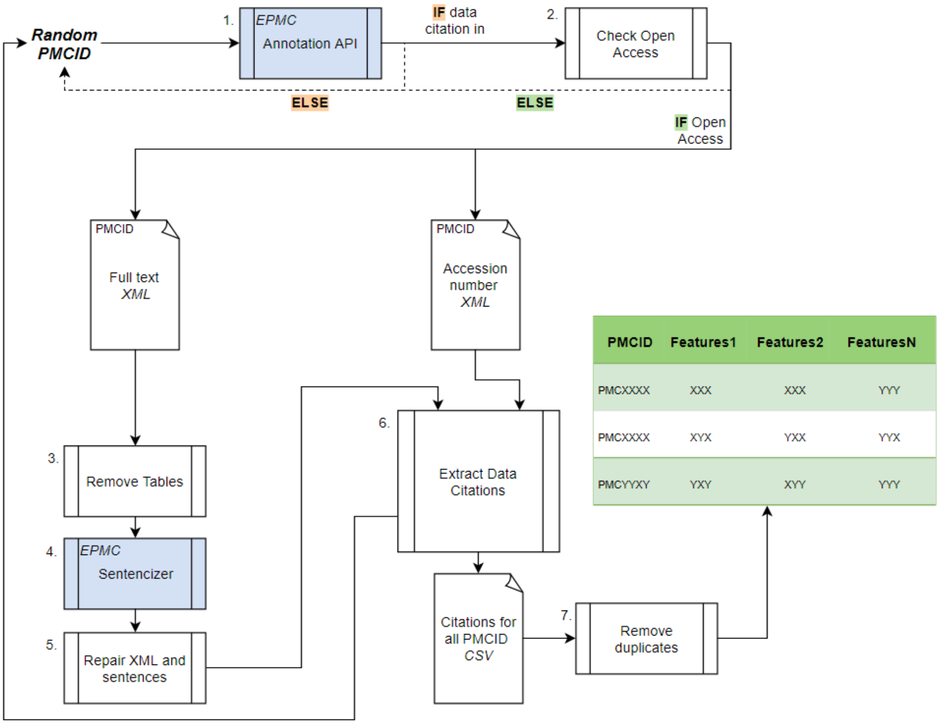
Creation of an automatic data extraction tool

In order to build this dataset semi-automatically, a pipeline was created, allowing the extraction of data citations and their meta data. This tool uses a directory and a number of scientific articles (containing data citations) to work with. An average number of 4.5 data citations by paper was observed (when paper contains data citation). This information allows to have an approximation for the number of validated papers (papers containing at least one data citation text mined) required to work with.

A diagram of the proposed the pipeline is shown below (Figure 2.1).

The pipeline generates a random PMCID represented as PMCXXXXXXX, where each X is a number between 0 and 9. These random identifiers thus made it possible to obtain random bio-medical documents stored in EPMC. Then the *Annotation API* from EPMC is used to check if there is some accession numbers in the paper. The *Annotation API* is an API that provides text mining annotations contained in abstracts and open access full text articles, using the *W3C Open Annotation Data Model.*

Here the pipeline requests accession numbers from a specific PMCID, if there is at least one in the paper (**1**), then it verify that the full text of this PMCID is *Open Access* (**2**).



PMCID	Features1	Features2	FeaturesN
PMCXXXX	XXX	XXX	YYY
PMCXXXX	XYX	YXX	YYX
PMCYYXY	YXY	XYY	YYY

Figure 2.1: Work-flow of pipeline for dataset creation

If these two conditions are satisfied, then two files are saved, one containing the full text of the document and the other one containing the accession number annotations and their meta data from *Annotation API*. Both files are downloaded in the XML format. Then the pipeline removes tables that are not useful in this study (**3**) (Tables can't be use for a text classification task and there is also some problems with the *Sentencizer*).

After this step, the full text document is split by the Sentencizer from EPMC into a list of sentences saved in another temporary XML (**4**).
This file was then repaired (**5**), indeed, mistakes can be made by the *Sentencizer*, but some of these mistakes can be corrected. The most frequent mistake occurs when a paper contains citation to another scientific article : *et al.* (example from [15]).

> The two main model families for learning word vectors are: 1) global matrix factorization methods, such as latent semantic analysis (LSA) (Deerwester *et al.***SPLIT**, 1990) and 2) local context window methods, such as the skip-gram model of Mikolov *et al.***SPLIT** (2013c).

At each **SPLIT** the Sentencizer splits the sentence because of the dot in *et al.*, instead of leaving it as one sentence.

The resulting XML sentencized file and the XML file containing annotations were used to extract data citations (**6**).

Here the goal is to work only with data citations that are text mined by EPMC and provides by *AnnotationAPI*. Text mined accession numbers, or a term that pass each validation steps, is not necessarily retrieved in the full text paper. Therefore, if an accession number is mentioned more than once in a document and one of the entries is recognized as an accession number by the *Annotation API*, it will not always be found in the article due to validation steps?. The pipeline therefore needs to find the context of the accession number text mined by EPMC to match exactly.

For each annotation, *AnnotationAPI* gives a *pre-tag* (some characters before the accession number), *exact match* (accession number) and *post-tag* (some characters after the accession number) where *pre-tag* and *post-tag* corresponds to the context of the data citation (repectively almost ten charcacters before and after the exact match). However, because it exists some differences between the full text and the resulted string : *pre-tag + exact-match + post-tag*, the pipeline only extracts annotations using the context of the exact match. Then, white spaces of those two resulted strings are deleted and ratio of similarity calculated thanks to the *SequenceMatcher* of the *difflib* library. If the similarity is greater than 85% then the sentence and all meta data corresponding are saved in a CSV file.

Finally, duplicate citations are deleted (**7**). Indeed, sometimes there is a lot of data citations in the same sentence, it results in multiples lines with the same sentence. But the goal here, is to train a model and give it the same sentence several times, could lead to an over-fit of the model.

2.2 Categories definition

2.2.1 Inspiration from literature

Some categories presented by [7] were not suitable for this study. Indeed, the *Motivation* category could be removed : data citations can not illustrate a need for data because it is already data. For the *Extension* category, it seems that extended data are really rare events in data citations and could be removed from categorizations for our machine learning models. As mentioned before, the *Comparison or Contrasts* category can not be used here because it is also very rare in data citations. First, this category was kept, but since there was not enough data to train a model correctly, it was removed. *Future* category has been removed too because data could not be a potential avenue for future work.

The “**Background**” and “**Use**” categories seemed relevant to the study and were therefore retained. Indeed, the “**Use**” category seemed obvious for the data, because it makes it possible to show the results of the research. This has been more complex for the “**Background**” category. Indeed, it can be very difficult to see how data can be used as a context, but there are many cases, for example (from *Lu et al. 2016* [16]):

> A2M is an evolutionarily conserved element of the innate immune system and a non-specific protease inhibitor involved in host defense, and it has been revealed that A2M is relative to immunity in L. vannamei [65]. *The F11 gene (**GenBank: AFW98990.1**) was reported to play a role in immunity [1].* Recent studies revealed the importance of KLKB1 in shrimp immune response, particularly towards protect animals from the microbial pathogens [66]. Aquatic animals metabolize foreign toxicity of chemicals mainly by oxidation, reduction, hydrolysis and conjugation reactions catalyzed by various enzymes, and the metabolic activation is primarily catalyzed by the cytochrome P450-dependent oxygenase system in the endoplasmic reticulum [67].

It is not easy to say here that the *F11 gene* was used. Authors try to give more information about what they try to show and even more they made a reference to a past study about this gene to support their purpose. Therefore "**Background**" category has been kept even if sometimes for an human it could be difficult to see a difference between those two categories.

2.2.2 New category : Creation

Previously selected categories were not enough to correctly describe data citations. Indeed, in some cases, scientists cite data because they have created those. That is why there is a need of a new category : "**Creation**". These citations were made to show who published those data and made those available for all other scientists.
This category can be described as ("Accession number" corresponds to the accession number of data of interest) :

- *Creation* (The authors have published, deposited "Accession number")

In the end, three categories were selected for labelling the data citation data set, which could be used in training machine learning or deep learning models for automatic classification :

- **Background** ("Accession Number" provides a context for the purpose)
- **Use** (Uses data from "Accession Number")
- **Creation** (Authors published "Accession Number")

1507 data citations from 535 papers, extracted by the created pipeline, have been manually annotated with those categories.

2.3 Approaches, models and predictions

2.3.1 Scores and Approaches

Scores

To estimates the quality of a model, different scores were used :

TP:True positive - TN:True negative - FP:False positive - FN:False negative

- *Accuracy* : It is simply a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- *Precision* : Ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{TP}{TP + FP}$$

- *Recall* : Ratio of correctly predicted positive observations to the all observations in actual class.

$$Recall = \frac{TP}{TP + FN}$$

- *F1-score* : This score is the weighted average of Precision and Recall.

$$F1 - score = \frac{2(Recall \cdot Precision)}{Recall + Precision}$$

It is important not just use the accuracy, especially when the data is unbalanced, indeed, this score could indicate that a model is great, but hide that the precision is really bad for example : In cancer detection, 99% data are negative and only 1% data are positive. If all the data are predict as negative, accuracy will be 99% but it is not useful at all because positives examples are the most important to detect accurately. It also results in zero precision and recall because there is no true positive.

The labelled data-set was split with a 5-fold cross validation, meaning that it is splitted in five subsets. Each subset is used as a test in turn, while the rest is used as training. Then an average score is calculated for each measure (precision, recall, etc.). This step is really important, indeed it allows to avoid over-fitting but also some misunderstanding about predicting on a test set that is really similar of the training set. Training and testing on different parts of the dataset gives a good overview of the model quality.

Approaches

To use machine learning or deep learning models, data should be numeric and not alphabetic, it is therfore important to change words to numeric values. First there is some pre-processing steps. Here for the first step two approaches has been used :

- *Tokenization* : In this case words are consider as tokens, it means that for a sentence each word will be consider as a unique piece and the punctuation is throw away.
  The sentence : "*p53 is a protein.*", become $\boxed{\text{p53}}$, $\boxed{\text{is}}$, $\boxed{\text{a}}$, $\boxed{\text{protein}}$.

- *N-gram* : Here it is a little bit different indeed will not consider each word as a token but this method will take a sequence of N words and consider this as a sequence.
  A 2-gram approach of the sentence : "*p53 is a protein.*", become $\boxed{\text{p53 is}}$, $\boxed{\text{is a}}$, $\boxed{\text{a protein}}$.

Then each token or piece has been pre-processed, here following methods has been used :

- *Raw*, no modification was made on tokens
- *Stemming* [17], when we stem a word it is reduce to its stem, base or root for example *"fishing", "fisher", "fished"* become *"fish"*. The stem needs to be really close to the morphology of the root word.
- *Lemmatization* [18], is different from stemming. First, link was made between a word and a lemma. A lemma is a base form that one might look up in a dictionary.

Those three approaches were used here, the simple one where words were not modified, the stemming and then the lemmatization approach.

In the end tokens or pieces has been vectorized thanks to two methods :

- A *Term frequency-inverse document frequency*(*Tfidf*) Vectorizer [19] was used. This method takes a document and computes the frequency of each term in the document and the frequency in the corpus (to balance over-represented words). It results in a score that corresponds to each word. But sometimes a simple *count vectorizer* is used, this one will not balance the frequency of a term in a document by its frequency in the corpus (it is why the *Tfidf* is generally preferred).
- *Embedding* represents each word by a numerical vector. If two different words have the same or similar context, they should have really close vector, for example we can expect that *"cat"* and *"dog"* should be really close. Here a pre-trained embedding was used. *GloVe* [15] has provided pre-trained word vectors from *Wikipedia* and *Gigaword 5* that could be represented between (50 dimensions , 100 dimensions, 200 dimensions or 300 dimensions) each of these is used here.

2.3.2 Models and Predictions

In text classification and sentiment analysis there are a lot of different models that are used, there is therefore a need to select some of them to test those on data.

- In the first time it seems that *Naive Bayes* models are working well with text classification [20] as it gives good results. In this study following Naive Bayes models used here are :
  - Complement Naive Bayes (*ComplementNB*)
  - Gaussian Naive Bayes (*GaussianNB*)
  - Multinomial Naive Bayes (*MultinomialNB*)
- Some neural networks :
  - Recently Convolutionnal neural network (*CNN*) gives really good results in text classification tasks [21–23], it is an interesting model to study.
  - Long Short Term Memory (*LSTM*) gives also good results [23, 24] recently, it can therfore be used here to compare to other models.
  - Simple Neural network (*SimpleNN*) was also selected to have an overview (One input layer, two dense layers, One output layer).
- Three machine learning models seems also interesting for the task :
  - Support Vector Machines (*SVM*) are well known to give good results [25, 26].
  - Random Forest (*RF*) sometimes are pretty good too [27].
  - Logistic Regression models (*LR*) [28] seems pretty fast for good performances.

These are all models studied here, at least it's nine models have been studied.

First of all, a first selection was performed using those nine models, then using the scores described before and especially the F1-score. Models were trained and tested on manually labelled dataset. Thereafter, the most powerful models were selected and then further optimized thanks to a grid search. Finally, the best model was selected according to another test on manually labelled set.

Once the best model and the best approach were chosen, 2.532 unlabelled citations from 847 papers and 49.894 others from 10.406 papers were extracted with the automatic tool for dataset creation for classification. The distribution of citations was observed, analyzed and compared.

Chapter 3

Results

3.1 Created dataset and analysis

3.1.1 Data citation dataset

The data extraction tool combined with a manual annotation of 1507 citations from 535 *Open Access* biomedical documents stored by EPMC, therefore result in our manually annotated dataset. Here is an overview of this one Table 3.1.

As described in section 2.1.2, the dataset contains all features for more than 1500 data citations and has been manually annotated.

PMCID	AccessionNb	Section	SubType	Figure	Categories	PreCitation	Citation	PostCitation
PMC5520553	rs10052999	Abstract	RefSNP	False	Use	Both donor and recipient CYP3A5 rs776746 allele A were correlated with decreased concentration/dose (C/D) ratios.	Recipient C6 rs9200 allele G and donor C6 rs10052999 homozygotes were correlated with lower C/D ratios.	
PMC6256655	SRP158285	Article	ENA	False	Creation	This whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession number PREU00000000.	Raw sequences were deposited in the NCBI SRA database under accession number SRP158285.	
PMC3386246	3M0E	Results	PDBe	True	Use	Western blot analyses demonstrate that the wild type and all mutants produce comparable amounts of LuxO protein.	(B) The locations of the resistance-conferring mutations are inferred from the ATP-bound Aquifex aeolicus NtrC1 structure (3M0E).	Two monomers of NtrC1 are shown (cyan and green). The residues predicted to form the Walker B motif are shown in blue.

Table 3.1: Extract from the training dataset

This dataset was used here to train and test our models to select an appropriate approach and model. The same principle was used to produce an unlabelled dataset of 2.532 citations for 847 articles and another unalebelled dataset of 49.894 citations from 10406 papers. Those were used after the selection of the model to predict categories of these citations and thus analyze its results.

An initial analysis of the annotated model has been carried out.

3.1.2 Analysis of the labelled dataset

Sections

At first, the proportion of the total citation in each section was studied :

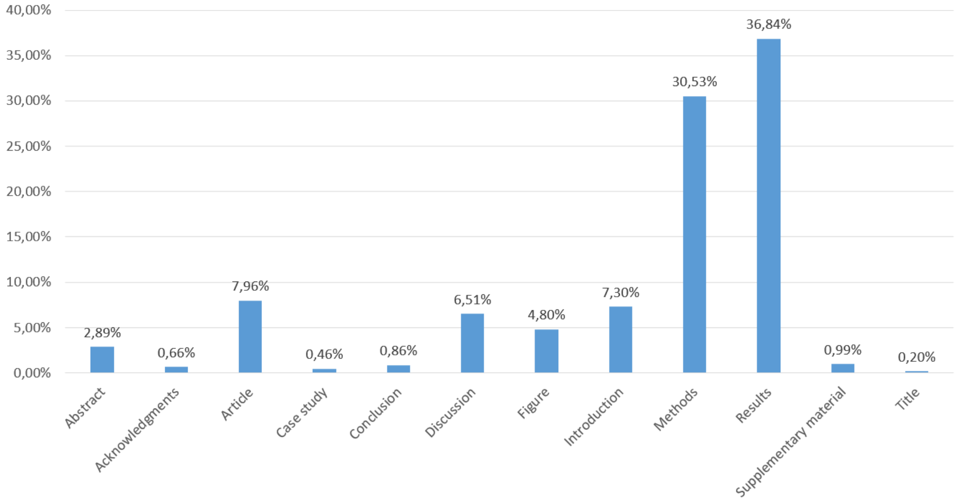


Figure 3.1: Data citation distribution in sections for the total of citation (*Manually labelled dataset*)

We can notice that two sections stand out clearly from the others: *Methods* and *Results* (Figure 3.1), with more citations in the *Results* section. However, a problem persists : indeed, for example, during SNP studies, it happens that many data citations are made in the same paragraph, it was therefore interesting to report this number of citations to the number of citations per article.

Thus for each paper a citation report by section has been established and the total of these reports has been studied.

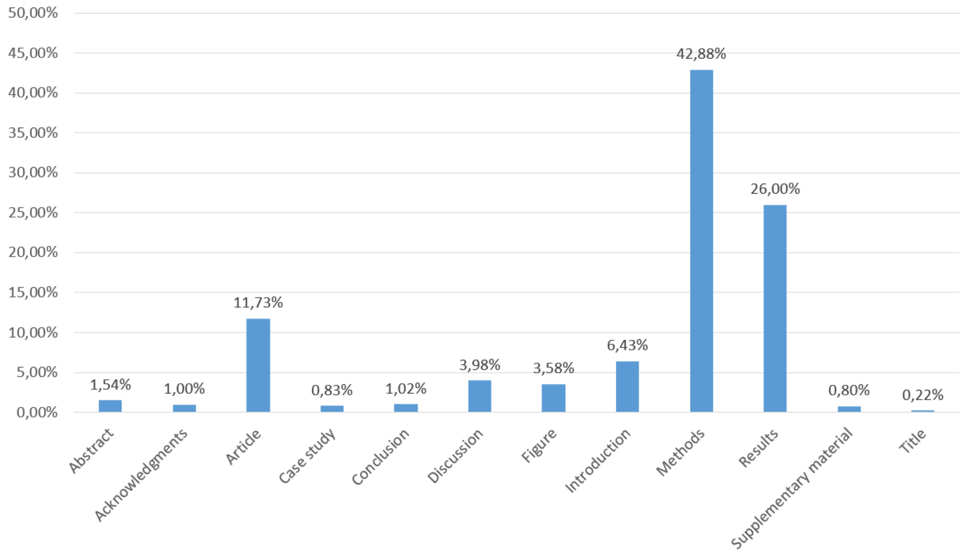


Figure 3.2: Data citation distribution in sections by paper (*Manually labelled dataset*)

It can thus be observed that in reality the frequency of data citations is higher in the *Methods* section than *Results* (Figure 3.2). It also shows that in the *Results* section there are more data citations. It can therefore be assumed that it is more common to find in the *Results* section lists of data citations, such as a list of SNPs proving to be interesting for the study.

Sub-Type

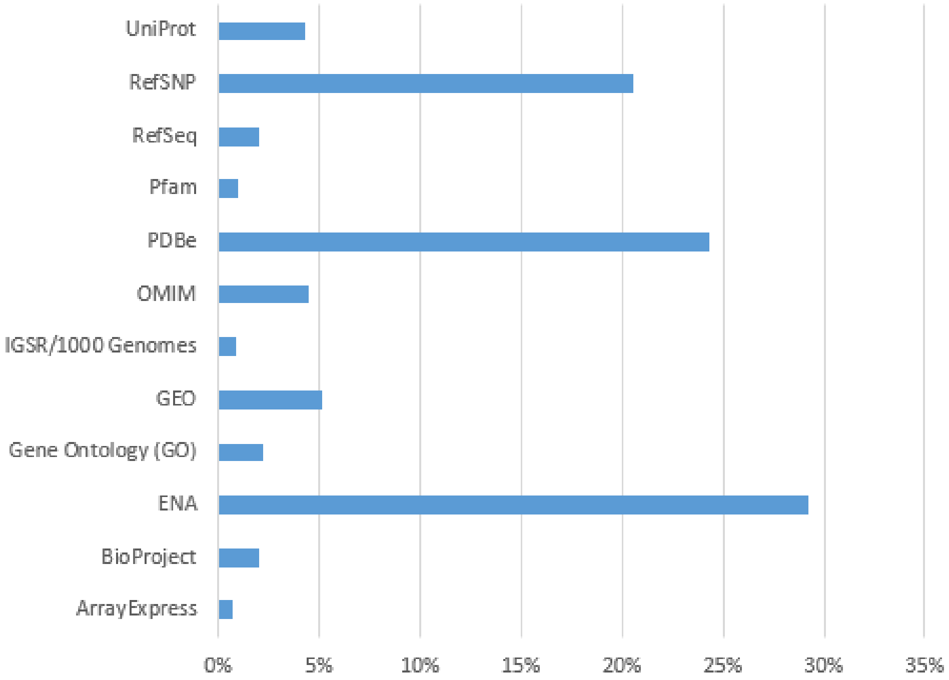


Figure 3.3: Sub-Type of data citations distribution for the total of citation (*Manually labelled dataset / 12/24 under-represented sub-types are not shown*)

Here (Figure 3.3), one can observe that a large number of databases provide annotations to EuropePMC, however only a few are relatively frequently cited : *ENA* [29] almost 40%, *PDBe* [30] almost 20%, *RefSNPs* [31] and *GEO* [32] almost 10% each. About 80% of citations come from four databases, which represents a significant difference.

Categories

Results of the manual annotation are also interesting to study (Figure 3.4).

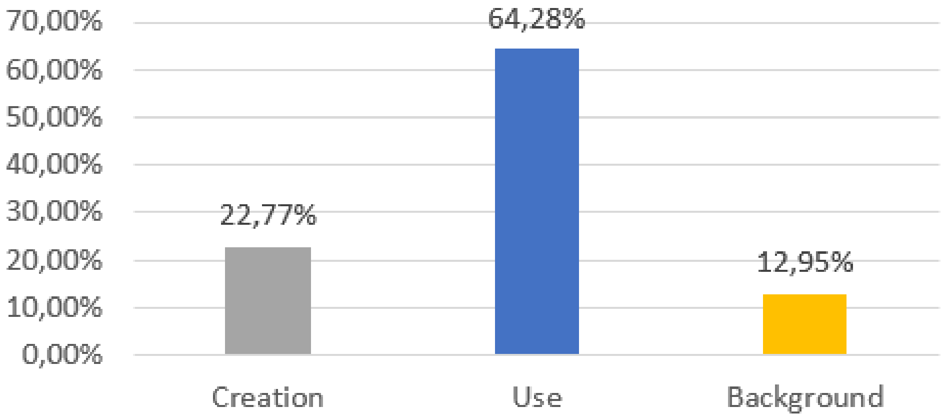


Figure 3.4: Categories manually annotated of data citations distribution by paper (*Manually labelled dataset*)

One interesting thing here is to note that the frequency of category "**Use**" is high, which may indicate a high re-use of the data. However this point will be discussed later. This chart also tells us something to consider. Indeed our data are unbalanced, it will therefore be important to assign weights to categories to compensate for this imbalance.

Subsequently, the distribution of these classes was studied according to the section.

Categories by Section

In the Figure 3.5 it is possible to notice an unequal distribution of categories according to sections. In addition, some categories like *Title* are restricted by the number of citations representing them, explaining such amazing results for these sections. However, these sections remain exceptional (*Acknowledgments*, *Case study*, *Title* <10 citations each) as they are under represented (Figure 3.1).

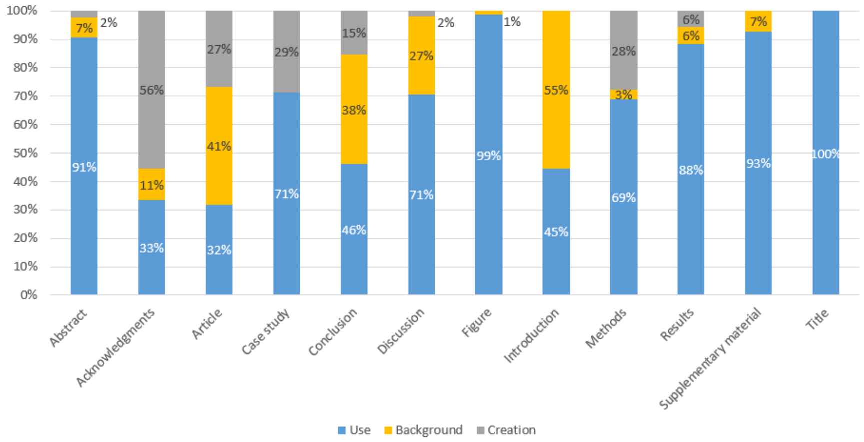


Figure 3.5: Categories distribution through sections
(*Manually labelled dataset*)

Categories by Sub-Type

In the following figure 3.6, some facts are interesting indeed, it seems that some databases produce only one or two categories and not the three defined before (section 2.2.1). There is only *PDBe* and *ENA* that follow the global distribution. "**Use**" category is still the one most represented and followed by "**Creation**". It is however important to note that this category is the one that appears less, in the 22 databases presented here only nine of them produce "**Creation**" data citations.

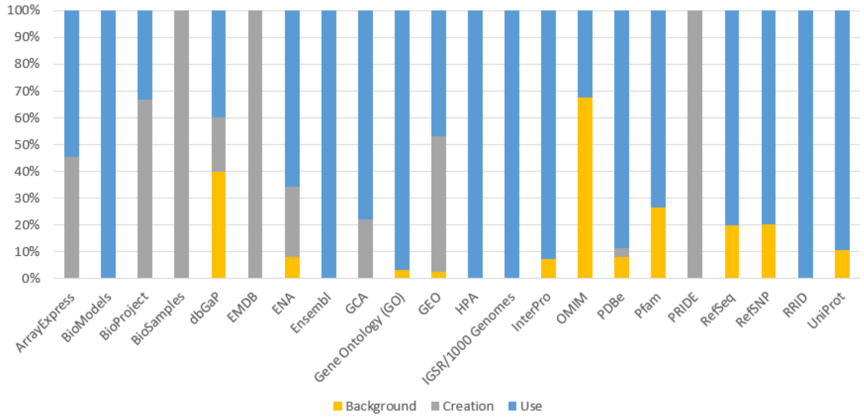


Figure 3.6: Categories distribution through sub-types
(*Manually labelled dataset*)

3.2 Model selection

3.2.1 Without optimization

A first selection among the nine non-optimized models was carried out and the following results were obtained : Table 4.1, in Supplementary material section.

This table represents the cross validation (CV) and test results on 20% of the data "not seen" by the model during the cross validation. Thus, the scores ending in CV represent the cross validation scores while those without this acronym match the scores on the prediction on unseen data during the cross-validation. Data separation was done randomly. However, in order to replicate the same random separation from one model to another, the random state is fixed.

The score that interests us the most here is the F1-score because it is a more reliable score when evaluating a model. A clear fracture of the F1-score is visible between 78.034% and 72.034% respectively *SimpleNN* and *ComplementNB*. On this criterion, models with F1-score greater than 75% were selected for optimization.

In Table 4.1 we can clearly see the interest of not showing part of the dataset during the cross validation step. We can note here that the first model (*simpleNN* + *Raw* approach) in bold has got the best results for cross validation and especially for the F1-score. But as soon as it test on unseen data during the cross validation, scores became lower than almost the first ten models. Indeed, even cross validation can sometimes lead to an overfitting of the model, it seems here that this model can't generalize from cross validation data.

Four models and their respective approaches were therefore selected :

- Logistic Regression
- SVMs
- SimpleNN
- CNN

3.2.2 After optimization

F1-score	*Precision*	*Recall*	*Accuracy*	*F1-scoreCV*	*PrecisionCV*	*RecallCV*	*AccuracyCV*	*Algorithm*	*Approach*
77.961	**79.816**	**76.822**	**87.368**	**83.572**	**87.179**	**81.209**	**89.384**	**Logistic Regression**	**Stemming**
76.017	79.317	74.648	86.579	83.081	87.568	80.45	89.474	Logistic Regression	N-gram, Lemmatization
76.624	78.857	75.373	86.842	82.543	86.767	79.8	88.947	Logistic Regression	Tokenization
75.821	78.958	74.648	86.579	82.41	87.526	79.733	89.121	Logistic Regression	N-gram, Stemming
73.66	75.685	72.353	85	82.217	86.207	79.434	88.592	SVM	Stemming
73.594	79.904	71.015	86.053	82.135	88.69	78.553	89.121	SVM	N-gram, Stemming
73.504	80.127	71.503	86.316	82.119	87.529	79.232	89.035	SVM	N-gram
75.162	76.27	74.288	85.263	81.899	85.831	79.492	88.42	Logistic Regression	Lemmatization
74.207	76.589	72.591	85.263	81.851	86.317	78.976	88.681	SVM	Lemmatization
73.466	80.729	70.529	86.053	81.732	88.205	78.247	88.946	SVM	N-gram, Lemmatization
73.535	78.841	71.988	86.053	81.727	85.948	80.233	88.773	Logistic Regression	N-gram
73.992	77.231	71.986	85.526	81.312	85.236	78.673	88.155	SVM	Tokenization
75.423	76.535	74.404	84.474	80.898	83.051	80.839	86.669	SimpleNN	Stemming
72.299	75.86	72.122	83.421	80.789	84.708	79.074	88.1	CNN	Lemmatization
74.24	80.556	71.133	85.789	80.776	84.167	78.741	87.9	CNN	Stemming
73.675	78.929	71.381	85.526	80.3	85.354	77.996	87.807	SimpleNN	N-gram
70.295	75.486	68.849	83.947	79.999	84.133	77.823	88	CNN	Tokenization
71.59	69.935	80.89	75.263	79.284	83.913	79.859	85.516	SimpleNN	N-gram, Stemming
73.926	76.074	72.231	84.211	78.844	81.731	79.187	84.738	SimpleNN	Lemmatization
74.028	79.088	70.64	83.947	77.524	86.961	74.053	86.574	SimpleNN	N-gram, Lemmatization
67.271	73.449	63.994	80	74.183	83.723	71.913	83.506	Dplearn	Tokenization

Table 3.2: Results with optimization

Table 3.2 presents the scores of the models and their different approaches after optimization.

Thus, thanks to these multiple stages of selection, the model and the approach with the most developed capacities were retained in order to predict later the 2.532 and 49.894 citations extracted previously.

The chosen combination corresponds to a **tokenization** followed by a token **stemming** step followed by a vectorization by **Tfidf vectorizer** as pre-processing, then the algorithm retained is a **Logistic Regression**.

3.3 Validation and Predictions

3.3.1 Validation

Chosen model was first train on ¾ of the labelled dataset and then test on the ¼ remaining (377 citations) this split has been made randomly. Thanks to this step it is possible to understand where are the weaknesses and strengths of the trained model.

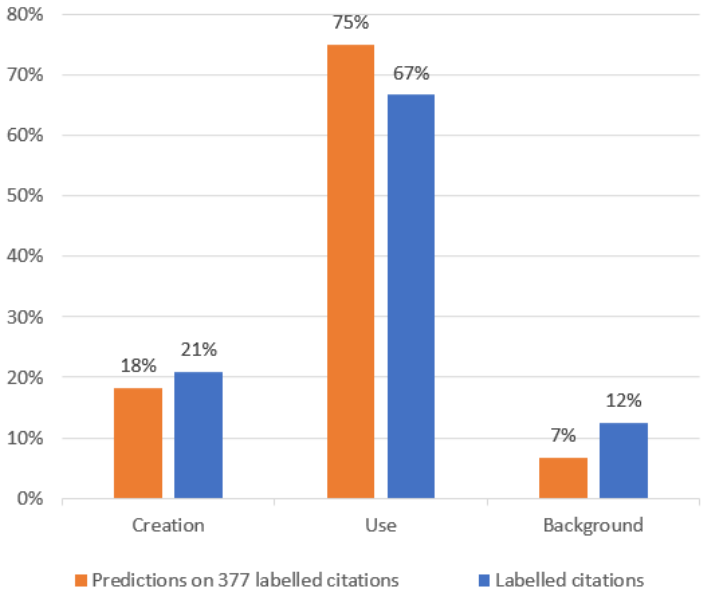


Figure 3.7: Categories distribution comparison between predictions and labelled citations (*Manualy labelled dataset*)

It is first important to note that this prediction produce 28% of True Positives and 61% of True Negatives and also 5% of False Negatives and 5% of False Positives, those scores gives a good overview of the model. Indeed it reach 89.21% of accuracy and a F1-score of 83.82%. Most of the wrong predictions comes from "**Background**" and "**Use**", indeed there is no mistake between "**Creation**" and "**Background**". It is also interesting to see that categories distribution are similar in Figure 3.7.

The Figure 4.2 in appendix show the same differences between "**Use**" and "**Background**". It is also interesting to note that the model seems to reproduce the categories distribution to section. This can confirm that this feature is definitely important for categorization. Model performance can also be verified by categories distribution through sub-types (Figure 4.3 in appendix), indeed it is possible to see that the model reproduce the sub-type distribution, indicating this feature is also important for categorization. (*Sections or sub-types of Figure 4.2 and 4.3 under-represented are not shown*).

3.3.2 Predictions

Chosen model was trained with the whole labelled dataset (1507 data citations from 535 scientific articles) and then it predicts each citation from the unlabelled dataset (49.894 citations from 10.406 papers). First of all, it is important to note that a threshold has been fixed, if a data citation has less than 95% to be one of the category it will be tagged as unknown.

Categories

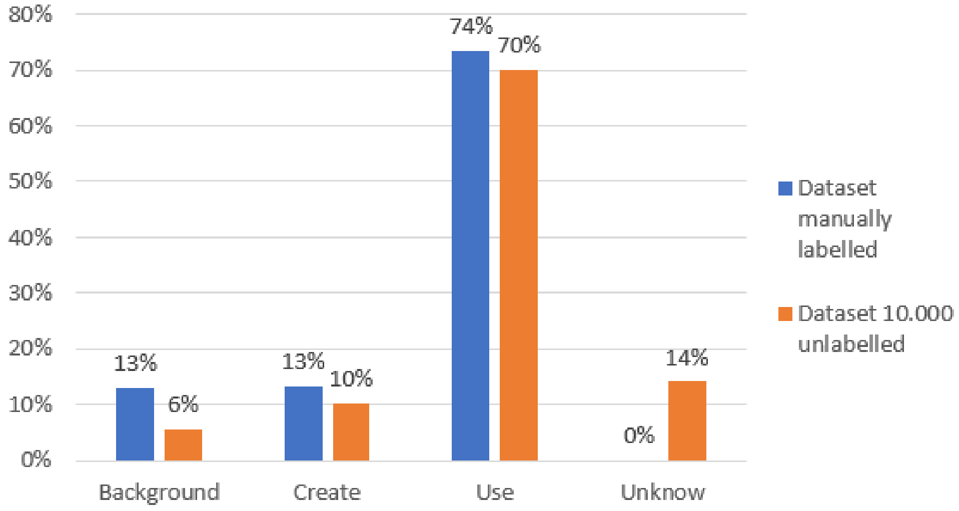


Figure 3.8: Categories distribution comparison between manually labelled dataset and unlabelled dataset (10.406 papers)

Figure 3.8 shows the comparison of categories distribution of unlabelled and labelled dataset. Here more than 85% of the unlabelled dataset has been tagged with the three established categories. The "**Background**" category distribution is the least similar to that of the labelled dataset but it could be explained by a difference between datasets. However it is interesting to observes that both distributions are similar.

In order to study the reuse of data, a single category has been assigned per accession number for each PMCID. Thus if an accession number is mentioned more than once in the same article, it will be identified with only one of the established categories. Categories distribution seems to be similar to observed distributions (Figure 3.9 and 4.1)

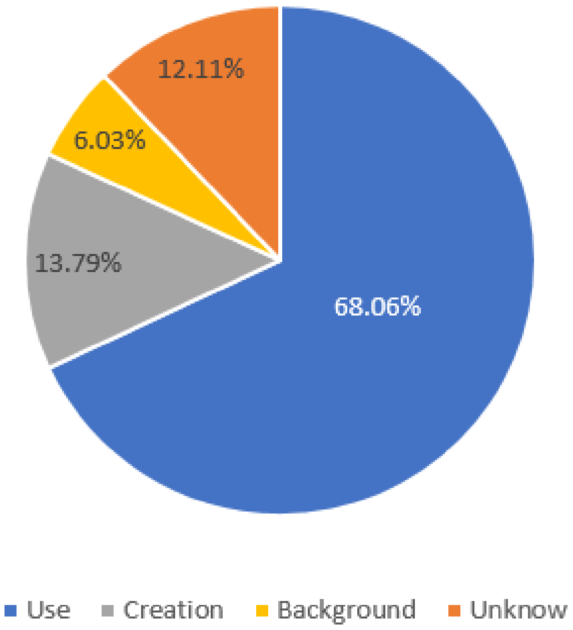


Figure 3.9: Predicted categories distribution of unlabelled dataset (10.406 papers / *1 category by accession number by PMCID*)

Section

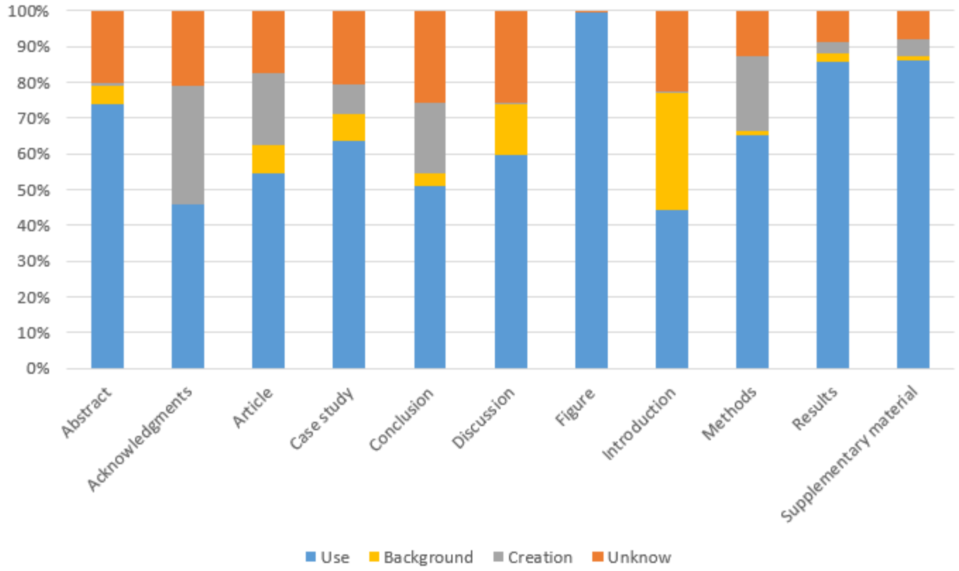


Figure 3.10: Predicted categories distribution of unlabelled dataset through sections (10.406 papers / *6/17 under-represented sections are not shown*)

Some sections are particularly interesting to study in the Figure 3.10, for example *Figure* section provides only data citations tagged as "**Use**". Also only a few sections contains "**Background**" citations (*Introduction* and *Discussion* especially). The *Acknowledgments* section provides also a lot of "**Creation**" comparing to other sections. As *Results* and *Methods* are the most represented sections it is interesting to note that they share almost the same proportion of "**Background**" but it seems that there is more "**Creation**" in *Methods* section.

Subtype

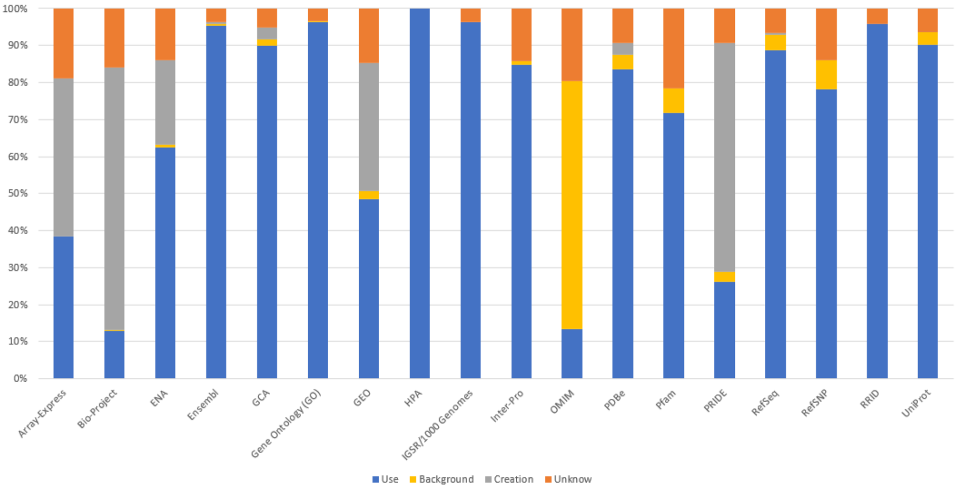


Figure 3.11: Predicted categories distribution of unlabelled dataset through sub-type (10.406 papers / *17 under-represented sub-types are not shown*)

In this figure (Figure 3.11), a lot of things are interesting to observe. First the *OMIM* [33] database is the most remarkable as there is a lot of data citations coming from this database that are tagged as "**Background**", it is because this database contains only text data, in the end it is more a database of biology's knowledge than a database of biological data.

Some databases seems to have almost zero "**Creation**", it could be because those databases are curated (for example *UniProt* [34]), it is therefore not possible to submit data.

However databases like *ArrayExpress* [35], *Bioproject* [36] or *PRIDE* [37] have a lot of "**Creation**", even some databases not shown produce only this category of data citations, it could because those are recent databases.

Chapter 4

Discussion

Nowadays, the automatic text study becomes imperative. Indeed, the amount of scientific knowledge is growing exponentially, especially in the field of biology. The human is no longer able to adapt to this growth, it becomes therefore imperative to try to teach machines to learn. As Bill Gates says, teaching a machine to understand our language is a sector of the future. The work done for this internship is anchored in this evolution.

The goal here is to automatically categorize data citations based on different categories. For this, automatic learning methods have been used.

A dataset was created using an automatic data extraction tool. It has extracted data from Open Access articles available through EPMC. The quotes corresponding to the phrase containing the accession number of interest were extracted but also the context of this quote in the form of sentences or metadata. This powerful tool has produced data quite close to what the human needs to categorize a quote. The sentencizer used here could be improved. Indeed, the separation into sentences of the articles used here sometimes results in separate sentences in the wrong place or unseparated sentences. However, this one generated more than satisfactory results, less than 5% of the resulting data were affected by this problem.

In the end, three datasets were generated. The first one was manually annotated thanks to the categories defined here. Two categories, pre-existing for another case, have been adapted and then one new category is created, respectively : “**Background**”, “**Use**”, “**Creation**”.

As a first step, this dataset reveals that, in general, data citations are done in two sections of scientific articles, *Results* and *Methods.* This result could be expected. Indeed, it is in these two sections where methods are described and thus data used but it's also in the result section where resulting data or related data are shown. However, multiple citations like the discovery of SNPs (often more than two or three) greatly influence the distribution by article (Figure 3.1). Surprisingly, when this factor is normalized, *Methods* section stands out with more than 40% of the data citations (Figure 3.2). This difference can indicate that there is a lot of paper containing only one data citation in *Methods.* Indeed papers containing only one data citation are common in almost 40% of the dataset of 10.000 papers (Figure 4.4). Moreover only 25% of those have their citations in the *Results* section while *Methods* section which represent 67% of those papers.
In the end it is possible to say that the section that have the most chances to contain data citations is the *Methods* section.

Subtype analysis reveals a great imbalance in the origin of data citations (Figure 3.3), almost 40 databases has been matched in this study.
On 56 databases providing accessions numbers to EuropePMC it is a great overview of data citations in EuropePMC. But there is still just a few databases that provides the majority of accessions numbers.

Finally, the manual annotation of this dataset also reveals something interesting and expected. Here, data citation is usually made because used by the authors (Figure 3.4 and 4.1). This is one element in favor of data reuse. Indeed, the creation of these data seems to be much less frequent. Moreover, citation of data as context seems to be an infrequent thing, which is also expected. However, the latter is sometimes even complicated to identify for a human.

The distribution of these categories in the sections indicates an effect of the section on the category (Figure 3.5). If this had not been the case, the proportion of citations in general would have been found in all these sections. Here, we find a great imbalance between sections especially for the *Figure* section which seems to have a significant effect on the distribution. Some categories are however difficult to analyze because of the lack of data concerning them (for example *Title*, Figure 3.1).

Categories distribution throughout sub-types is a little bit more complicated (Figure 3.6), as some sub-types are not well represented due to the lack of data (Figure 3.3). But it's something interesting to see some distributions like *PDBe* or *ENA* that follow the global categories distribution (65% "**Use**", 25% "**Creation**" and 10% "**Background**"). There are some others that don't follow this general distribution such as *UniProt*. Indeed it is not possible to submit data as it is a curated database, here it is well represented. The "**Creation**" category is produced by only 9 databases on 22 present in this dataset. For some databases like *EMDB* that are more recent comparing to the others, this difference can be associate to their ages. But this lack is still a problem, it could be great to run this categorization to all documents in EPMC. In the end this feature is consider here as important.

The annotated dataset allowed the selection of the ideal model and the most appropriate approach to obtain a powerful model. Thus, after the model selection without optimization (Table 4.1) and with optimization (Table 3.2), the accepted algorithm is *Logistic Regression* using the *stem* of each word in the citation. Surprisingly, the *N-gram* approach was not the one-providing the best results as expected [38]. Indeed, during the experiment, this one proved many times to be the best, on the other hand, other approaches seemed to have negligible effects. It would be interesting to continue on this path. Moreover, we can notice that certain sections or sub-types are under-represented. It would also be interesting to collect and annotate more data to see if this lack affects prediction results. However, the results of the selected model here are satisfying for the expected categorization.

Selected model and approach were used to predict the category of data citation of the two other unlabelled dataset generated. It was first necessary to check the similarity of distribution for the subtype and the section, thus making it possible to predict on a similar dataset to that of training. These appeared quite similar despite some differences in subtype distribution (Figure 3.3 and 4.5).

The predictions and manually annotated dataset distributions are very similar, suggesting that they have learned correctly. Only the "**Background**" category, is a little different (Figure 3.8). Categories distribution in sections or by subtypes also appear similar between datasets, especially the *Figure* section that seems to have an important effect, as expected, similarly for the *Introduction* or the *Abstract*. It is however important to remember that differences between datasets can come from the predicted dataset itself.

In this study, several things have been advanced.
First of all, it is the idea that it is possible to automatically categorize scientific data citations according to three categories ("**Use**", "**Background**", "**Creation**"). For this, an automatic extraction tool was created, and several models were tested. The *Logistic Regression* algorithm seems to be definitely the best for this task. Regarding approaches, the one selected here is to use the *stem* of each word in the sentence of interest. However, some additional studies would be needed. Here the selection was based on the level of the sentence. It would be interesting to look at the document level as it can provides more information [39]. If several citations with a different category are in the same sentence, they will be categorized with the same category. However the *dependency* approach has been studied and this didn't provides better results. Citations in tables are also interesting, here they are discarded but studying whether they are 100% used or not could be interesting.

In the end this work shows some good clues concerning the reuse of data in bio-medical science. It highlights some features for text classification of data citations in scientific papers (*Section, Subtype*). Those categories revealed a lot about data citations, database and how a scientific paper is written. It also bring a lot to data citations in scientific papers, some categories ("**Background**", "**Creation**", "**Use**") has been established and then used to train a model. These categories and their distributions can be a way to describe a database. For example the *OMIM* database describe, in 3.3.2, has a text and knowledge database. But also curated database like *UniProt*, *RRID* or *Inter-Pro* [40]. Or even databases with a lot of "**Creation**" could be associate with a more recent one. Established categories are also submit to a special scheme for sections distribution.
Several algorithms have been trained and then one has been selected (*Logistic Regression*) as some approaches has been tested and then selected (*Stemming*). This model and approach are able to reproduce this categorization with a 87.36% accuracy (and a F1-score of 89.4%).

For the future it could be interesting to manually annotate more data and train again this model, it would be also great to look further for different approaches.
It could be interesting to take a look at the variation of "**Creation**" through time for the sub-type feature it can maybe reveal important event like the discovery of a new technology like *CRISPR*.
And also as the "**Use**" category seems to be the most represented here it could be a potential avenue to see if there is any sub-categories that could established.

Bibliography

[1] Europe PMC Consortium, “Europe pmc: a full-text literature database for the life sciences and platform for innovation,” *Nucleic acids research*, vol. 43, p. D1042—8, January 2015.

[2] M. Levchenko, Y. Gou, F. Graef, A. Hamelers, Z. Huang, M. Ide-Smith, A. Iyer, O. Kilian, J. Katuri, J.-H. Kim, N. Marinos, R. Nambiar, M. Parkin, X. Pi, F. Rogers, F. Talo, V. Vartak, A. Venkatesan, and J. McEntyre, “Europe pmc in 2017,” *Nucleic acids research*, vol. 46, p. D1254—D1260, January 2018.

[3] A. Venkatesan, J.-H. Kim, F. Talo, M. Ide-Smith, J. Gobeill, J. Carter, R. Batista-Navarro, S. Ananiadou, P. Ruch, and J. McEntyre, “Scilite: a platform for displaying text-mined annotations as a means to link research articles with biological data,” *Wellcome open research*, vol. 1, p. 25, 2016.

[4] H. Mooney and M. P. Newton, “The anatomy of a data citation: Discovery, reuse, and credit,” 2012.

[5] M. A. Parsons, R. Duerr, and J.-B. Minster, “Data citation and peer review,” *Eos, Transactions American Geophysical Union*, vol. 91, no. 34, pp. 297–298, 2010.

[6] H. A. Piwowar and T. J. Vision, “Data reuse and the open data citation advantage,” *PeerJ*, vol. 1, p. e175, 2013.

[7] D. Jurgens, S. Kumar, R. Hoover, D. McFarland, and D. Jurafsky, “Measuring the evolution of a scientific field through citation frames,” *Transactions of the Association of Computational Linguistics*, vol. 6, pp. 391–406, 2018.

[8] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” *CS224N Project Report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[9] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

[10] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

[11] B. Keith, E. Fuentes, and C. Meneses, "A hybrid approach for sentiment analysis applied to paper reviews," 2017.

[12] A. Athar, "Sentiment analysis of citations using sentence structure-based features," in *Proceedings of the ACL 2011 Student Session*, (Portland, OR, USA), pp. 81–87, Association for Computational Linguistics, June 2011.

[13] B. Liu *et al.*, "Sentiment analysis and subjectivity.," *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.

[14] A. Athar, "Sentiment analysis of scientific citations," tech. rep., University of Cambridge, Computer Laboratory, 2014.

[15] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[16] X. Lu, J. Kong, S. Luan, P. Dai, X. Meng, B. Cao, and K. Luo, "Transcriptome analysis of the hepatopancreas in the pacific white shrimp (litopenaeus vannamei) under acute ammonia stress," *PloS one*, vol. 11, no. 10, p. e0164396, 2016.

[17] J. B. Lovins, "Development of a stemming algorithm," *Mech. Translat. & Comp. Linguistics*, vol. 11, no. 1-2, pp. 22–31, 1968.

[18] J. Plisson, N. Lavrac, D. Mladenic, *et al.*, "A rule based approach to word lemmatization," *Proceedings of IS-2004*, pp. 83–86, 2004.

[19] J. Ramos *et al.*, “Using tf-idf to determine word relevance in document queries,” in *Proceedings of the first instructional conference on machine learning*, vol. 242, pp. 133–142, Piscataway, NJ, 2003.

[20] A. McCallum, K. Nigam, *et al.*, “A comparison of event models for naive bayes text classification,” in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[21] Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.

[22] Y.-Y. Hsu, M. Clyne, C.-H. Wei, M. J. Khoury, and Z. Lu, “Using deep learning to identify translational research in genomic medicine beyond bench to bedside,” *Database*, vol. 2019, 02 2019.

[23] E. Vilar, “Word embedding, neural networks and text classification: What is the state-of-the-art?,” *JUNIOR MANAGEMENT SCIENCE*, vol. 4, no. 1, pp. 35–62, 2019.

[24] G. Liu and J. Guo, “Bidirectional lstm with attention mechanism and convolutional layer for text classification,” *Neurocomputing*, 2019.

[25] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *Machine Learning: ECML-98* (C. Nédellec and C. Rouveirol, eds.), (Berlin, Heidelberg), pp. 137–142, Springer Berlin Heidelberg, 1998.

[26] S. Tong and D. Koller, “Support vector machine active learning with applications to text classification,” *Journal of machine learning research*, vol. 2, no. Nov, pp. 45–66, 2001.

[27] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, and A. Gupte, “Comparative study of classification algorithms used in sentiment analysis,” *International Journal of Computer Science and Information Technologies*, vol. 5, no. 5, pp. 6261–6264, 2014.

[28] G. Ifrim, G. Bakir, and G. Weikum, “Fast logistic regression for text categorization with variable-length n-grams,” in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 354–362, ACM, 2008.

[29] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, *et al.*, “The european nucleotide archive,” *Nucleic acids research*, vol. 39, no. suppl_1, pp. D28–D31, 2010.

[30] S. Velankar, C. Best, B. Beuth, C. Boutselakis, N. Cobley, A. Sousa Da Silva, D. Dimitropoulos, A. Golovin, M. Hirshberg, M. John, *et al.*, “Pdbe: protein data bank in europe,” *Nucleic acids research*, vol. 38, no. suppl_1, pp. D308–D317, 2009.

[31] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, “dbsnp: the ncbi database of genetic variation,” *Nucleic acids research*, vol. 29, no. 1, pp. 308–311, 2001.

[32] R. Edgar, M. Domrachev, and A. E. Lash, “Gene expression omnibus: Ncbi gene expression and hybridization array data repository,” *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[33] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders,” *Nucleic acids research*, vol. 33, no. suppl_1, pp. D514–D517, 2005.

[34] U. Consortium, “Uniprot: a hub for protein information,” *Nucleic acids research*, vol. 43, no. D1, pp. D204–D212, 2014.

[35] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, *et al.*, “Arrayexpress—a public repository for microarray gene expression data at the ebi,” *Nucleic acids research*, vol. 31, no. 1, pp. 68–71, 2003.

[36] T. Barrett, K. Clark, R. Gevorgyan, V. Gorelenkov, E. Gribov, I. Karsch-Mizrachi, M. Kimelman, K. D. Pruitt, S. Resenchuk, T. Tatusova, *et al.*, "Bioproject and biosample databases at ncbi: facilitating capture and organization of metadata," *Nucleic acids research*, vol. 40, no. D1, pp. D57–D63, 2011.

[37] J. A. Vizcaíno, A. Csordas, N. Del-Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, *et al.*, "2016 update of the pride database and its related tools," *Nucleic acids research*, vol. 44, no. D1, pp. D447–D456, 2015.

[38] W. B. Cavnar, J. M. Trenkle, *et al.*, "N-gram-based text categorization," in *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, vol. 161175, Citeseer, 1994.

[39] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[40] A. L. Mitchell, T. K. Attwood, P. C. Babbitt, M. Blum, P. Bork, A. Bridge, S. D. Brown, H.-Y. Chang, S. El-Gebali, M. I. Fraser, *et al.*, "Interpro in 2019: improving coverage, classification and access to protein sequence annotations," *Nucleic acids research*, vol. 47, no. D1, pp. D351–D360, 2018.

Supplementary Material

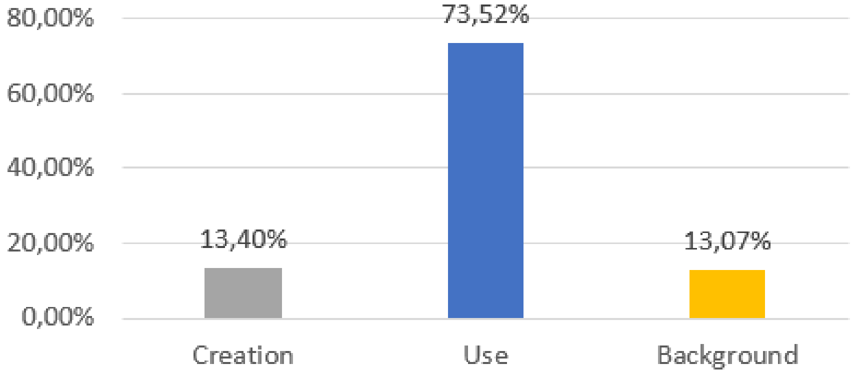


Figure 4.1: Categories of data citations, manually labelled, distribution for the total of citation (*Manually annotated dataset*)

Figure 4.1 show the general distribution of categories in the manually labelled dataset. It is important to compare this distribution with the distribution of categories by paper, indeed some differences are possible between those. For example it is possible that the "**Use**" category could be over-represented by paper containing only data citation categorized as a "**Use**".

F1-score	*Precision*	*Recall*	*Accuracy*	*F1-scoreCV*	*PrecisionCV*	*RecallCV*	*AccuracyCV*	*Algorithm*	*Approach*
69,932	**75,212**	**69,702**	**83,421**	**83,645**	**84,809**	**82,966**	**88,595**	**SimpleNN**	**Raw**
75,561	77,875	74,041	86,053	83,612	88,014	80,587	89,471	Logistic-Regression	Stemming
77,379	82,991	74,640	87,895	82,855	88,076	79,563	89,208	Logistic-Regression	N-gram, Stemming
75,756	79,359	73,676	86,842	82,770	87,591	79,685	89,034	Logistic-Regression	Raw
73,914	76,233	72,472	85,263	82,675	87,074	79,602	88,856	SVM	Stemming
76,622	82,482	73,915	87,632	82,626	87,776	79,341	89,033	Logistic-Regression	N-gram, Lemmatization
73,594	79,904	71,015	86,053	82,461	89,125	78,820	89,297	SVM	N-gram, Stemming
74,207	76,589	72,591	85,263	82,158	87,181	78,959	88,944	SVM	Lemmatization
73,829	80,652	71,622	86,579	82,119	87,529	79,232	89,035	SVM	N-gram, Raw
74,707	73,121	76,845	83,421	82,092	83,737	81,157	87,807	SimpleNN	Stemming
75,673	80,492	73,557	86,842	82,076	86,677	79,515	88,948	Logistic-Regression	N-gram, Raw
77,172	79,437	75,610	86,842	81,916	86,374	78,970	88,507	Logistic-Regression	Lemmatization
73,466	80,729	70,529	86,053	81,732	88,206	78,247	88,946	SVM	N-gram, Lemmatization
74,249	77,854	72,106	85,789	81,343	85,647	78,530	88,243	SVM	Raw
74,028	75,755	73,206	84,737	81,335	85,553	78,989	88,500	CNN	Lemmatization
68,754	73,382	66,913	83,421	81,145	88,584	77,664	89,000	CNN	Lemmatization
72,612	75,317	70,412	83,158	81,109	85,624	78,355	87,717	SimpleNN	Lemmatization
70,709	74,496	69,334	83,421	80,779	84,805	78,872	88,300	CNN	Lemmatization
72,105	74,538	71,150	83,684	80,286	83,335	78,769	87,500	CNN	Raw
71,182	77,142	69,573	83,947	80,281	85,499	77,580	88,200	CNN	Lemmatization
74,706	80,194	71,738	85,789	80,007	85,551	77,239	88,300	CNN	Raw
73,233	81,689	70,897	86,053	79,753	86,837	76,767	88,700	CNN	Raw
77,113	82,604	73,175	86,316	79,720	85,308	76,769	86,753	SimpleNN	N-gram, Stemming
75,143	81,318	72,831	86,053	79,470	87,527	76,550	88,600	CNN	Stemming
76,713	80,136	74,766	86,316	79,408	84,929	77,178	88,100	CNN	Stemming
72,710	81,771	68,830	85,526	79,346	83,285	78,085	86,403	SimpleNN	N-gram, Lemmatization
74,389	77,627	72,959	85,263	78,899	86,115	75,782	88,000	CNN	Raw
75,110	78,537	72,710	85,526	78,895	85,454	76,121	87,900	CNN	Stemming
71,903	82,294	68,592	85,263	78,809	85,020	75,766	87,400	CNN	Stemming
68,675	84,924	65,566	85,263	78,034	79,813	79,774	84,385	SimpleNN	N-gram, Raw
66,158	74,839	64,491	82,895	72,034	87,841	67,319	84,823	ComplementNB	Stemming
69,018	82,443	67,146	85,789	72,006	88,981	68,800	85,875	ComplementNB	N-gram, Raw
66,384	76,284	64,610	83,158	71,746	87,838	67,266	85,085	ComplementNB	Raw

Table 4.1: Extract of the first results without optimization

Table 4.1 present different resulting scores for each combination of approaches and models, those are sorted by *F1-scoreCV*, it corresponds to the corss validation f1-score. The first combination in bold represent the best result by cross-validation however it seems not to be the best regarding other scores.

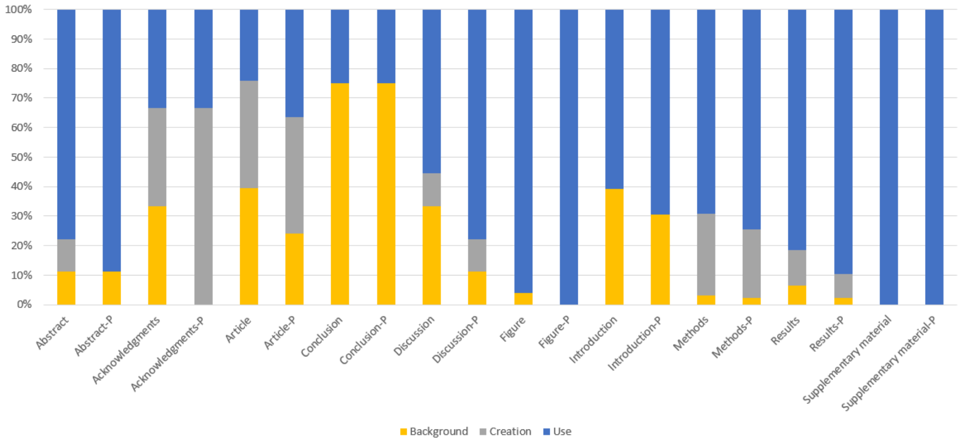


Figure 4.2: Categories distribution through sections comparison between predictions and labelled citations (*under-represented sections are not shown*)

Figure 4.2 show the comparison of categories distributions in papers sections between prediction (made by the trained model) and labelled citations. Sections ending with a *-P* corresponds to predictions. Here it is interesting to note that most of the differences come from the "**Background**" and "**Use**" categories. However it seems that the model succeed in its prediction task.

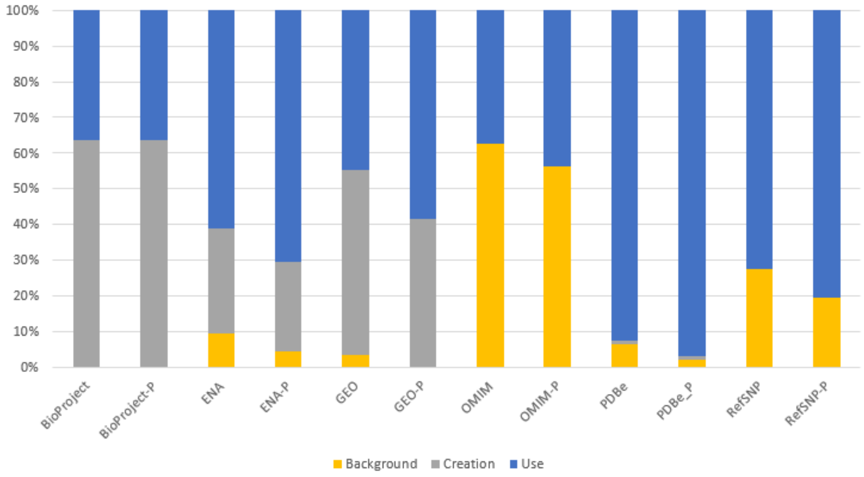


Figure 4.3: Categories distribution through sub-types comparison between predictions and labelled citations (*under-represented sub-types are not shown*)

Figure 4.3 show the comparison of categories distributions through data citations subtype between prediction (made by the trained model) and labelled citations. Subtypes ending with a *-P* corresponds to predictions. This figure show the same differences and similarity of figure 4.2 except that this figure is about subtypes.

It is however important to note that those distributions indicates an effect of both data citation section and subtype.

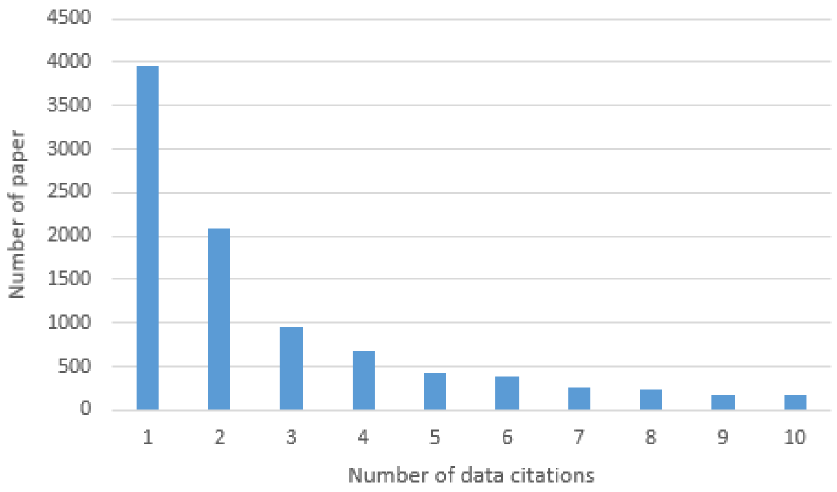


Figure 4.4: Number of data citations by paper distribution in the dataset
(*dataset of 10.000 papers / under-represented numbers of data citations are not shown / maximum : 102 citations in a paper*)

Figure 4.4 show the distribution of data citation by paper, here in the 10.000 papers dataset almost 4.000 contains only one data citation. But this distribution is truncated by the number of citation. Indeed some articles contains sometimes a lot of data citations.

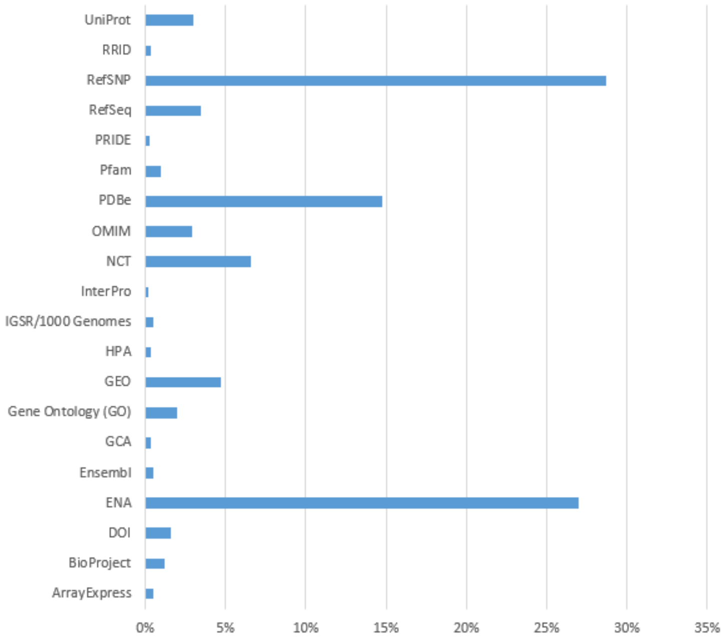


Figure 4.5: Sub-Type of data citations distribution for the total of citation (*10.000 papers dataset / under-represented subtype are not shown*)

Figure 4.5 show how databases are represented in the dataset thanks to data citation subtypes. It can be note that a lot of databases are represented and only a few are really well represented.