UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE, ECONOMICHE E SOCIALI

M.Sc. in Data Science and Economics

Perspectives on data sharing: a social media analysis

Supervisor:

Prof. Giancarlo MANZI

Co-Supervisor:

Prof. Alfio FERRARA

External Supervisor:

Prof. Dr. Christiane SCHWIEREN

Author:

Martina VIGGIANO

954603

Academic Year 2020/2021

Contents

List of Figures

List of Tables

Abstract

The goal of this work is to collect, study and analyze opinions about research data sharing.

The implementation of data sharing is being discussed recently both on the side of researchers, research funding agencies and universities as being desirable. The project will analyse incentives and perceptions of the actors concerned when it comes to data sharing, in order to derive actionable recommendations for policies.

The outcomes may be influenced by the academic culture in each field, the incentive structures for researchers, and individual characteristics. The objectives of the project are to understand how large the difference between perceived omission and description norms is and what can influence this difference.

In particular, the thesis project deals with several important questions, including: How do relevant groups of "actors" in data sharing field - i.e. scientists and data subjects - perceive the current standards? Is there a discrepancy between perceptions and opinions within years?

We also have a specific field of the subject we want to investigate: the view about sharing data underlying a given research, besides sharing the discovery itself.

Furthermore, we will study differences on tweets published before and after Covid19 era - namely, 2018 and 2019 compared to 2020 and 2021. To study this aspects, we will collect statements and discussions published on Twitter by users.

Introduction

Nowadays data represents a fundamental resource for every area of study, industry and innovation. The concept of data sharing is, then, essential for development, since spreading information fastens growth and innovation.
European Commission established the *Research and innovation strategy 2020-2024* which is based on letting knowledge and breakthrough innovation drive digital transformation. Sharing knowledge is a prerequisite for a faster convergence towards a sustainable and prosperous future, and also solidarity and respect are key pillars of European values. Specifically, research and innovation policies are have a significant role to play in addressing modern-day challenges.[34]
Focusing on academical field, data collected during researches does not only constitute an educational and innovative resource, but also an economical asset. While the entire industry would highly benefit from sharing data collected, some researchers and institutions are quite reluctant to open their research data to the outside. Among the reasons why this can happen, we can find: the economical and financial aspects, exclusive or non-disclosure contract, privacy and data protection protocols, and many more.
In this project we will consider tweets published between 2006 and 2021 for extracting opinions on this theme published by users. Data collection will be based on given keywords and hashtags presence. Then, we will filter and clean them before making an exploratory analysis and perform sentiment analysis and topic modelling on data, in order to detect relevant terminology and topics addressed.

Chapter 1

Spread of Knowledge

The concept of research data sharing lies in Open science ideal and runs in parallel with the general idea of Open data and Open Access. Open data and Open access to resources represent in turn underlying pillars for the concept of Open science.

While Open science describes specifically the on-going evolution of doing research and science organization, Open Data is a broad concept including different fields of data sharing - not strictly related to academic area.

These concepts are increasingly widespread and used in several fields; we owe the latest developments primarily to digital technologies and more recently to big data, but they are also is driven by the globalisation of scientific community, which itself keeps opening its boundaries and sharing knowledge worldwide.

Sharing research data has an impact on the entire life cycle of research: it influences the way research is performed, the collaborations within researchers, the way knowledge is shared, and how science is organised. This corresponds to openness of research data, methods, results and publications, always taking into consideration the limits of research agreements and research integrity.[2]

In practice, all these concepts - Open science, Open science, Open access, Open research - form one the subset of the other, but still intersecting each other: it is quite hard to clearly distinguish and separate each of them. In fact, also in

literature and speeches they tend sometimes to overlap and be even confused.

1.1 Open Science Definition

Open science can be defined as the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods.[1] In other words, it consists of a transparent and accessible knowledge that is shared and developed through collaborative networks.
Open science is about increasing rigour, accountability, and reproducibility for research. It is based on principles of inclusion, fairness, equity, and sharing, and ultimately seeks to change the way research is done, who is involved and how it is valued. It aims to make research more open to participation, review, improvement and re-use for the world to benefit.[3]
Though, it is hard to clearly and uniquely determine the concept and the definition of "openness", also with regards to various practices of science, namely: Open access to publications, Open research data, open source software/tools, open workflows, citizen science, open educational resources, and alternative methods for research evaluation including open peer review.
Fecher and Friesike[4] identified five Open science schools of thought, based on the underlying aims and assumptions to implement the practices:

- Infrastructure school - concerned with the technological architecture - motivated by the assumption that efficient research requires readily available platforms, tools and services for dissemination and collaboration;

- Public school - concerned with the accessibility of knowledge creation - which claims that science needs to be accessible for a wider audience. Moreover, in order to obtain a true impact on society a societal engagement in research and readily understandable communication of scientific results are

required;

- Measurement school - concerned with alternative impact measurement - motivated by the fact that traditional metrics (i.e. number of publications per researcher) proved to be problematic, some alternative metrics linked to network digitization of the field may be used to replace older metrics;
- Democratic school - concerned with access to knowledge - following the principle that everyone should have the same right to access knowledge, especially when its state funded. This concerns mostly research publications and scientific data, but also source materials, graphical materials, etcetera.
- Pragmatic school - concerned with collaborative research - based on the fact that the creation of knowledge is made more efficient through collaboration, this school looks for fostering network effects by connecting scholars and making scholarly methods transparent.[4]

Unlike the pure concept behind Open science, academic sphere and culture can often be hierarchical and conservative, thus even if researchers happen to be sympathetic to the concept of Open science, they might not yet see the worth in taking them up, as there is no existing mechanism encouraging the culture of openness and collaboration. Persuading researchers of the need to change their practices would require a good understanding not only of the ethical, social and academic benefits, but also of the ways in which embracing Open science practices will actually help them succeed in their work.[3]

Besides, Open science represents also a policy priority for the European Commission, which designs it as the standard method of working under its research and innovation funding programmes, since this approach improves the quality, efficiency and responsiveness of research. As a matter of fact, the Commission requires beneficiaries of research and innovation funding to make their publications available in Open access and make their data as open as possible and as closed as necessary, in order to encourage data sharing and Open access to

School of thought	**Central assumption**	**Involved groups**	**Central Aim**	**Tools and Methods**
Democratic	The access to knowledge is unequally distributed.	Scientists, politicians, citizens	Making knowledge freely available for everyone.	Open access, intellectual property rights, Open data, Open code.
Pragmatic	Knowledge-creation could be more efficient if scientists collaborated.	Scientists	Opening up the process of knowledge creation.	Wisdom of the crowds, network effects, Open Data, Open Code.
Infrastructure	Efficient research depends on the available tools and applications.	Scientists and platform providers	Creating openly available platforms, tools and services for scientists.	Collaboration platforms and tools.
Public	Science needs to be made accessible to the public.	Scientists and citizens	Making science accessible for citizens.	Citizen Science, Science PR, Science Blogging.
Measurement	Scientific contributions today need alternative impact measurements.	Scientists and politicians	Developing an alternative metric system for scientific impact.	Altmetrics, peer review, citation, impact factors.

Table 1.1: Five Schools of Thought by Fecher and Friesike[4]

researches.[6]

Our research question is about a subset of the wide idea of Open science: we are interested in studying the perception on Open research data world, which does not represent a precise area of Open science, but is the result of a mixture of more branches (see Figure 1.1), namely: Open research, Open data, Open Access and Open science tools.

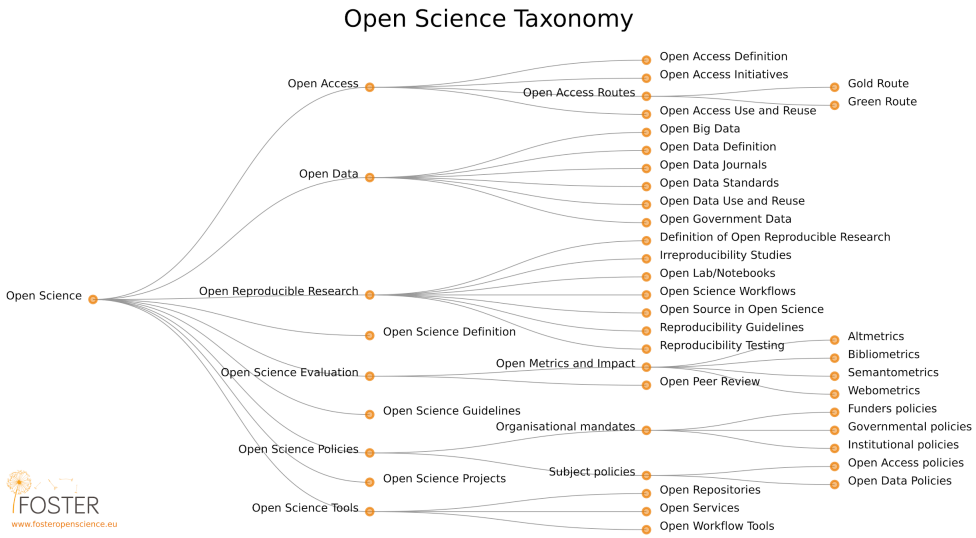

Figure 1.1: Open science taxonomy by Foster[1]

1.2 Open Research Definition

Open research is a concept that includes the openness throughout the entire research cycle: it goes through collaborative working, sharing and making research methodology, software, code and equipment freely available online, along with instructions for using it. Another main aspect is to publish and make studies freely available online - concept corresponding to Open access - in addition to the underlying research data - to which sometimes literature refer to as Open data, even if it is not precise.[13]

While, Open research data is the kind of data that is shared to be freely accessed,

reused, reshuffled and redistributed, for teaching research and academic purposes and beyond. Technically, Open research data should not have any restriction on reuse or redistribution, and be appropriately licensed as such. In exceptional cases - e.g. to protect the identity of human subjects - special or limited restrictions of access are set. Openly sharing data exposes it to inspection, forming the basis for research verification and reproducibility, and opens up a pathway to wider collaboration. At most, Open data may be subject to the requirement to attribute and share alike.[3]

1.3 Open Data Definition

Among the concepts already presented, Open data is the one shared among the higher number of environments. In fact, as it can be seen from Foster Open Science Taxonomy graph (see Figure 1.1), inside Open data idea there are not only research related content, but also journals publication and reference to data, government data releasing and big data sharing.
European Commission defines Open data as:

> "data that anyone can access, use and share. Governments, businesses and individuals can use open data to bring about social, economic and environmental benefits."[33]

This type of data must be also licensed: its licence enables people to use, transform, combine data, including also the possibility to further share it with anyone, even commercially.
An intrinsic feature behind Open data is to be freely to use, but it does not imply the access to be free. In fact, research may be really costly, in terms of data creation, maintenance and publishing. Thus, require a fee to have access to data is allowed - most of the time the fee to pay is quite negligible. Moreover, the cost should not exceed the reasonable reproduction cost of data extracted.
The rationale behind Open data lies in data contributing to information building, and without information there is no new knowledge. Data is the "raw material"

from which information and knowledge can be derived. Information originates from data when we give a given scenario to data. Knowledge in turn is built starting from information, and can be even personalised for specific contexts.
The reuse and distribution of data must go together with its format: in fact, without having a give structure and readability, even if data is shared with the outside, users couldn't really profit from it.[33]

1.4 Open Research Data Definition

The research question of this thesis project deals with Open research data, which can be defined as that type of data at the base of scientific research results, not having any access restrictions, and that enables anyone to reach information they may need.[35]
The main characteristic we can use to research data need to distinguish between open and not open research data are:

- Availability of data repositories;
- Policies of research funders and journals;
- Researchers' attitude towards data sharing.[35]

European Commission included Open research data as one of the objectives and pillars underpinning *Research and innovation strategy 2020-2024*[34]. As a matter of fact, the international organization keeps track of the evolution and numbers of research environment to see if industry meets requirements established.
European Commission released information collected on Open research data on the official website. From that data, we get to know that, among different subjects of study, Life Sciences and Natural Sciences are the one with the highest number of open data repositories (see Figure 1.2).

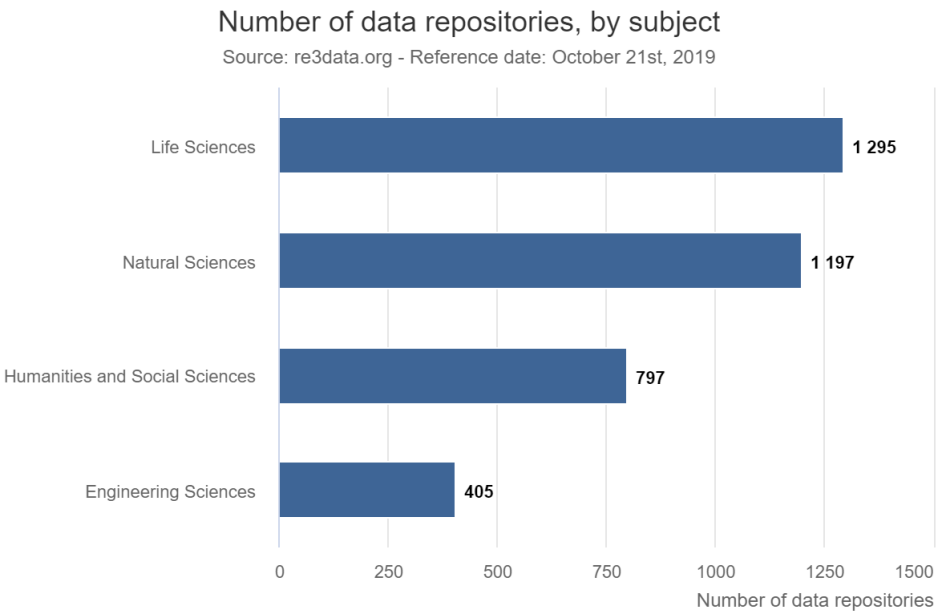

Figure 1.2: Number of Data Repositories by Subject[35]

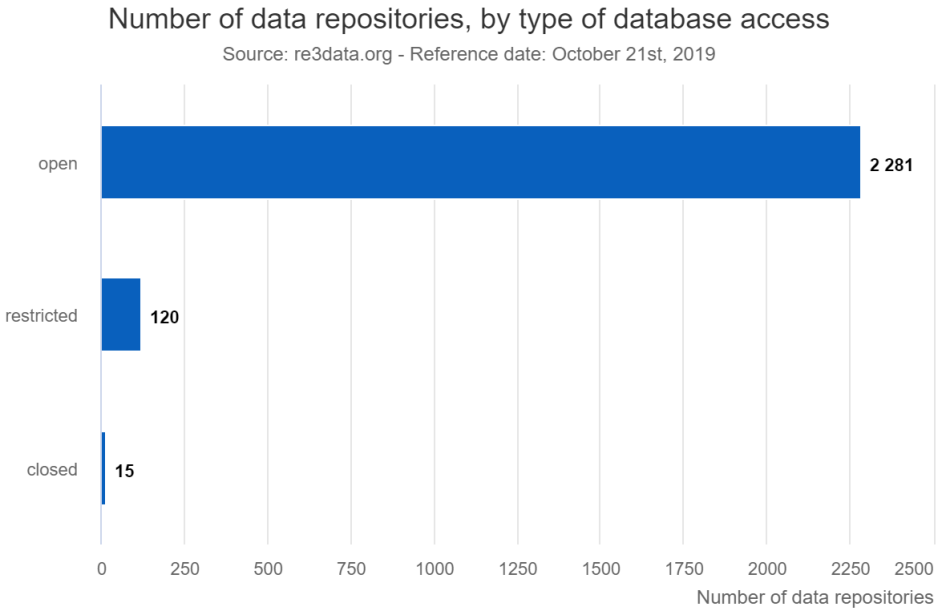

Figure 1.3: Number of Data Repositories by Type of Access[35]

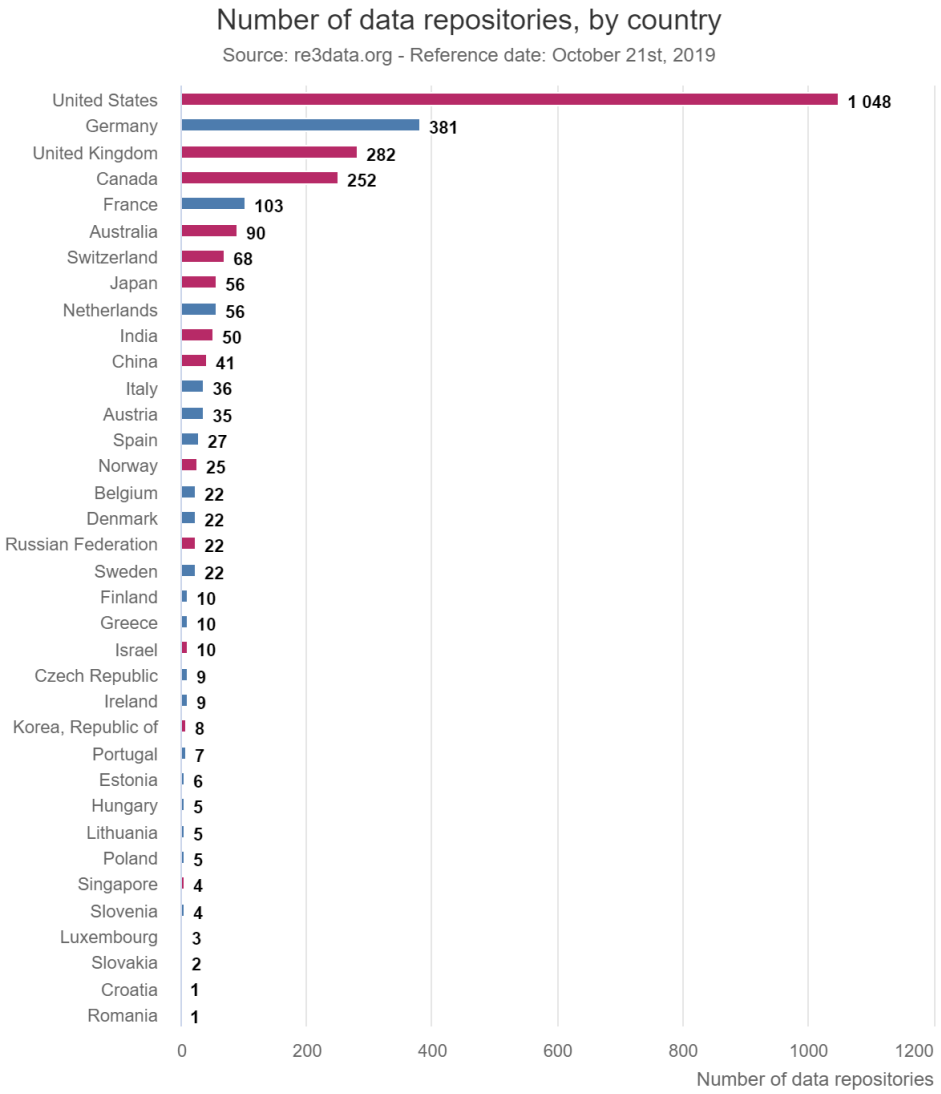

Figure 1.4: Number of Data Repositories by Country[35]

While, as regards type of access to the database, when we deal with open data repository, mostly there is no accessibility restriction (see Figure 1.3).
As for open data policies, for the vast majority standards do not require any policy nor imperative to publish data collected doing research (see Figure 1.5 and 1.6.

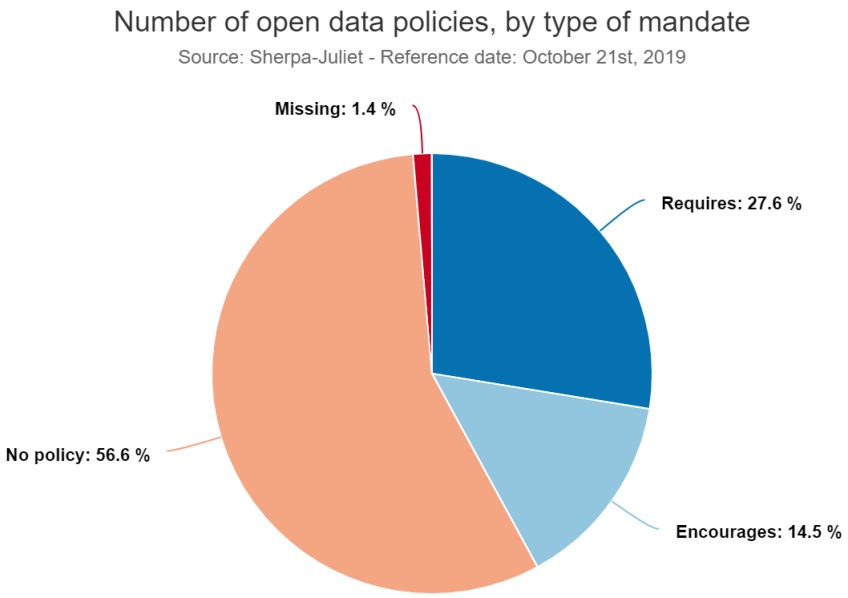

Figure 1.5: Number of Data Policies[35]

1.5 Historical Journey of Finding Sharing

In ancient Greece, science was used as the foundation for competitive public debates, operated to solidify knowledge into separate schools of thought and hinder the collaboration among scientists.

Similarly, medieval science was shaped by a political and - mostly - religious outlook that encouraged withholding the Secrets of Nature from the "vulgar multitude". The idea and practice of Open science emerged during the late sixteenth and early seventeenth centuries; it was a distinctive and vital organizational aspect of the Scientific Revolution. It represented a break from the previously practice of secrecy in the pursuit of Nature's Secrets, to a new set of norms, incentives, and organizational structures that reinforced scientific researchers' commitments to rapid disclosure of new knowledge. The rise of "cooperative rivalries" in the revelation of new knowledge, represented a functional response to asymmetric information problems built by the Renaissance system of court-

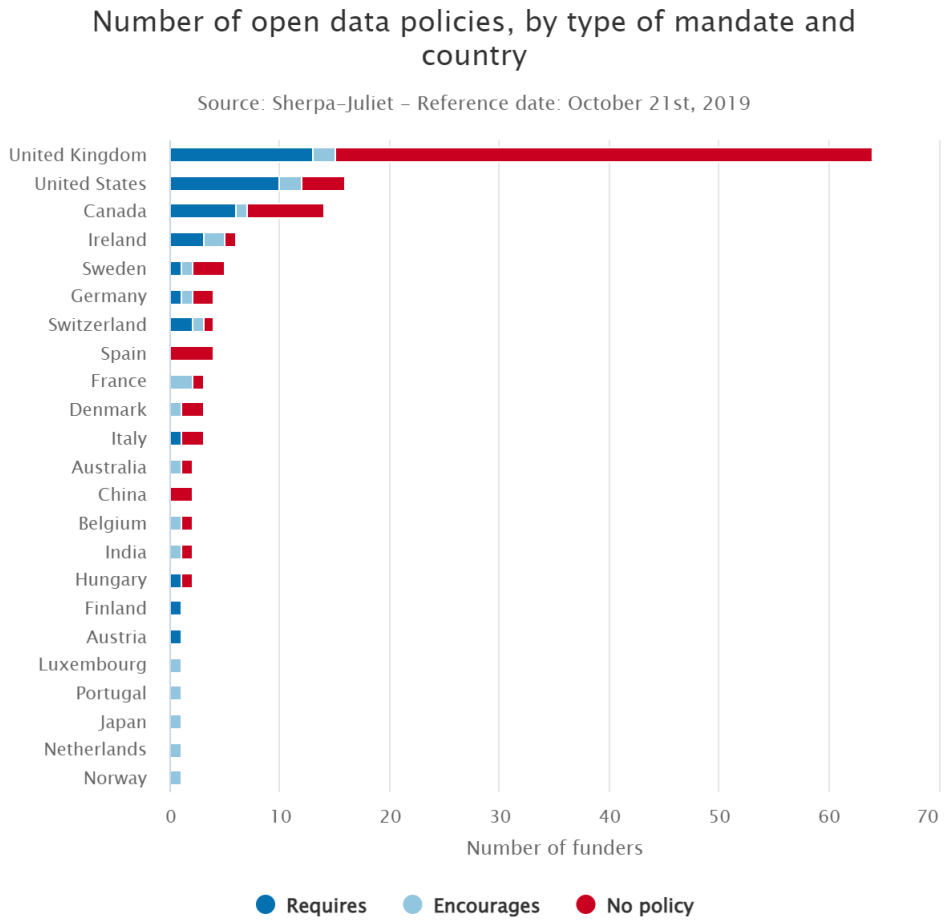

Figure 1.6: Number of Data Policies by Country[35]

patronage of the arts and sciences.

Competition among Europe's noble patrons motivated much of their efforts to attract to their courts the most prestigious natural philosophers, was no less crucial in the workings of that system than was the concern among their would-be clients to raise their peer-based reputational status.

An example of a particularly strong secrecy imperative, is the medieval and Renaissance tradition of Alchemy: this practice was considered as a form of personal knowledge, a "divine science" rather than a natural science. It survived from the seventeenth century to the eighteenth - side by side with the emergent

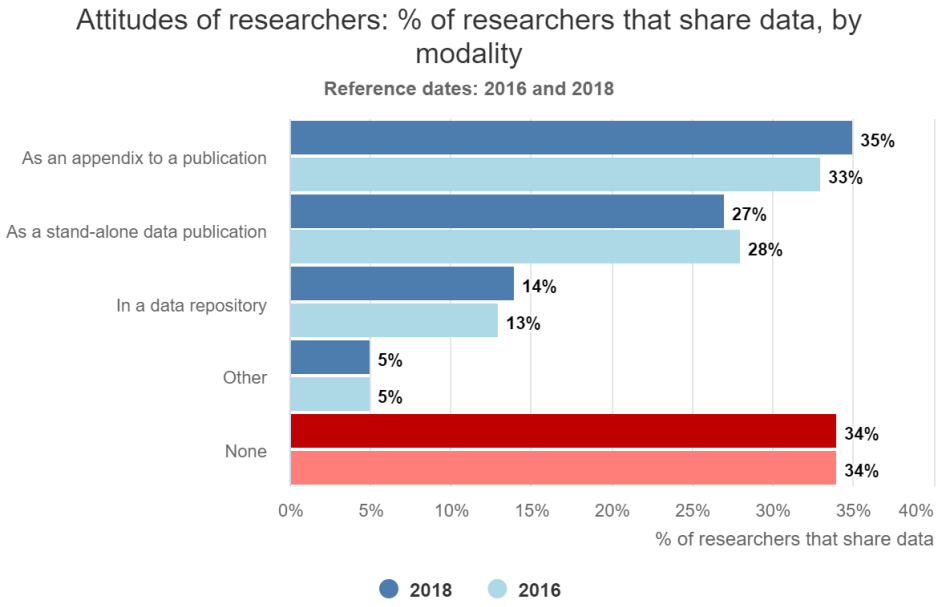

Figure 1.7: Attitude of Researchers Towards Data Sharing[35]

institutions of Open science.[7]

The world of research dissemination, copyright and collaboration has changed significantly since the 16^{th} and 17^{th} centuries, when scientific studies were kept secret to build the prestige of the sponsors.

Before the advent of scientific journals, researcher had not much to gain but much to lose by publicizing their scientific discoveries. [9] In the 1600s, scientists used to share their discoveries to competitors in an encrypted form, waiting for the time when their findings would become profitable before claiming their research. In fact, initially, the industry was controlled by patrons, who financed researchers and discouraged them from being open and share their results, in order to be the only ones to take profit a from that investment.

Thus, scientists were funded just to develop either immediately useful things or to entertain: this process gave prestige to the patron in the same way funding of artists, writers, architects, and philosophers did. That led to scientists under pressure to satisfy the desires of their patrons, and discouraged from being

open with research which would bring prestige to persons other than their own patrons.[11]

We can find an example of this procedure in some Galileo's and Isaac Newton's manuscripts: for instance the latter scientist maintained the tradition of secrecy and used some cryptographic characters in his voluminous writings, claiming their discoveries by writing papers coded in anagrams or cyphers and distributing the coded texts.

This system caused several problem: firstly it discouraged publications and therefore slowing down the entire scientific and research sector. Another issue produced by this situation was the difficulty for discoverers to prove priority: since they rarely published their researches - and when they did it, data had to be coded - scientists who wanted to benefit from priority struggled to profit from their discoveries.

Eventually, when aristocratic patronage could not meet demand for scientific knowledge anymore - not being able to sufficiently fund scientists, who needed consistent resources to have stable and proficient careers - academies began to form. By 1699, 30 scientific journals were founded and the number increases more than thirty times within 100 years. Academies enabled researches to share notions and researches to pursue a common purpose.

In the meantime, in 1710 in Great Britain passed the first statute regulating copyright - Statute of Anne, also known as the Copyright Act 1710. Initially copyright law only applied to the copying of books; over time other uses such as translations and derivative works were made subject to copyright and nowadays copyright covers a wide range of works.[10]

Pushed by the advent of academies, scientific journals started to be founded: the first popular science periodical was published in 1872, the *Popular Science*. In particular, it documented the invention of the telephone, the phonograph, the electric light and many other very relevant discoveries for modern life. But, one of the many innovations scientific journals brought was the normalizing data

sharing and Open science.

Lately, academies came across a trade off between research sharing and profit: some research products could potentially generate commercial revenues, so hoping to capitalize on these products, many research institutions withhold information and technology, even though those discoveries would clearly lead to overall scientific advancement by discolouring them to other research institutions. On the other hand, predicting payouts of technology and assess the costs of withholding it happens to be very difficult. Also, it is of general belief that the benefit one single institution gains holding technology is not greater than the potential benefit the entire sector would achieve by sharing it to all other research institutions.

1.6 Open Science Today

The term "Open Science" was coined quite recently: in 1998 Canadian engineer Steve Mann referred to Open Science as ideal science practices and scientific modes of communication. Even if Steve Mann technically invented the phrase, the Open science movements did not depend on him and his definition, since he did not take the initiative to found any movement to push the Open science model.

The actual change took place at the beginning of the 21st century. With the advent of "Internet age" the idea of open knowledge became more feasible and closer to reality. The main pillar was to make access to publications completely free. In fact, in 2001 emerged the concept of Open Access, a kind of free and unrestricted online availability, which should give readers the power to find and make use of relevant literature, and give researchers and their works vast and measurable new visibility, impact and readership. To achieve this, during the 2001 Budapest conference organised by the Open Society Foundations this issue was discussed, in order to introduce the Open Access Initiative on the political landscape. In the resulting statement they claimed the use of digital tools such

as open archives and Open access journals, free of charge for the reader, and invited individuals and all interested institutions to help open up access and encourage barriers removal. [12] One of the main barriers in Open science history was the price and profit barrier. Further on this point, the declaration states:

> "[...] (While) journal literature should be accessible online without cost to readers, it is not costless to produce. However, experiments show that the overall costs of providing Open access to this literature are far lower than the costs of traditional forms of dissemination. With such an opportunity to save money and expand the scope of dissemination at the same time, there is today a strong incentive for professional associations, universities, libraries, foundations, and others to embrace Open access as a means of advancing their missions. Achieving Open access will require new cost recovery models and financing mechanisms, but the significantly lower overall cost of dissemination is a reason to be confident that the goal is attainable and not merely preferable or utopian".[12]

More each year, this idea of openness gained traction in scientific community, and became the best way for researchers to gather the range of observations needed to speed discoveries or to identify large-scale trends.[5]

Chapter 2

Text Mining Introduction

Text Mining - to which we will also refer to by using the abbreviation TM - corresponds to a set of statistical and computer science techniques specifically developed to analyse text data. Its overarching goal is to turn text into data so that it is suitable for analysis[14]

In particular, the processes are based on transformation of unstructured textual data into a structured format to identify meaningful patterns and insights. The exploration and discoveries of hidden relationships within their unstructured data are achieved by applying computationally-intensive artificial intelligence algorithms and advanced statistical techniques, for example Naïve Bayes, Support Vector Machines, and other deep learning algorithms.[15] Textual data can be organized in three different ways: structured, unstructured, and semi-structured. Structured data is organized typically in a tabular format, which helps storage, processing and readability. On the other hand, unstructured data corresponds to text that do not have a specific and predefined format produced by various sources - i.e. documents, social media posts, reviews, books - or rich media formats like, audio files and videos. Semi-structured data is a combination between structured and unstructured data formats: it has some organization, but is not sufficiently structured to meet the requirements of a relational database.

In our case, the data we will use for analysis is extracted from Twitter, thus is

organized in an unstructured format: thus, the preprocessing and cleaning phase represent a crucial step for a successful analysis.[14][15] Text Mining processes comprise various activities that enable to retrieve information from unstructured text data.

But, before applying any text mining technique, it is needed to preprocess text. Text preprocessing is the practice of cleaning and transforming text data into a usable data format, and represents a core aspect of natural language processing: typically, it involves the use of one or more of the following techniques: language identification, tokenization, stemming, lemmatization, part-of-speech tagging, chunking, and syntax parsing. These techniques are fundamental before implementing any statistical processing and for obtaining a proper outcome. When text preprocessing is complete, we can apply text mining algorithms to derive knowledge from the data.

Text Mining consists of a wide range of tasks and methods that can be combined together into a single workflow. We can identify and divide the areas in:

- Information Retrieval;
- Natural Language Processing (NLP);
- Knowledge Extraction;
- Data Mining.

2.1 Information Retrieval

Information Retrieval is a process that returns relevant information - usually documents - that satisfies an information need, typically based on a pre-defined set of queries or phrases. For this reason, Information Retrieval is commonly used for cataloguing books systems and search engines. Some common IR sub-tasks include:

- Tokenization: This is the process of splitting text from its original form - for example a document - into sentences and words called “tokens”. For

example, if we consider a dataset composed of a set of reviews, the collection of reviews is called *corpus*, which is the object containing various and distinguishable single textual elements, the reviews, which we call *documents*. We can further split a document into smaller items, which can correspond to sentences or words; these smaller elements are called *tokens*. By dividing into smaller and smaller pieces textual data, we are able to "simplify" and address machine's work. Then, the resulting tokens are used in the models for text clustering and document matching tasks.

- Stop Words removal: Stop words are very common words in a language like "the", "a", "for", "at", "on", "is", "are". Since, these words do not provide any meaning and are usually removed from texts, they are typically removed before proceeding with the analysis.

- Stemming: This technique refers to a process of word-normalization based on removing prefixes and suffixes from words, to derive the root word form and get the meaning. In this way, we map a group of words to the same stem, even if the stem is not a valid word in the given language. One of the main goal of this task is to improve information retrieval by reducing the size of indexing files, thus partially solve sparsity.

- Lemmatization: Lemmatization is another way to normalize textual data. This task reduces the inflected words to a root words which actually belong to the language, unlike stemming. The root word is called lemma, and it corresponds to the canonical form or dictionary form of a set of words.

- Bag of Words (BoW) building: After text tokenization and cleaning, by using this model a text can be represented as a set of its words, considering the multiplicity of terms by themselves, ignoring grammar and words order. This technique is typically used for computing terms occurrence.

2.1.1 Information Retrieval - Python Implementation

Within the tasks above, by using Python programming language in this project we implemented text tokenization, stop words removal, bag of words definition, and lemmatization.
We employed tokenization during data cleaning phase. As a matter of fact, we defined a function able to correct specific contractions, remove numbers and punctuation - by using Regular Expressions operations - and the lemmatize the corpus. In order to prepare data for that function implementation, we had to first extract single words: to do so, we split strings by white spaces, so each step provided in the function was performed on each term present in any string.
As regard stop words removal and lemmatization, we used some specific *spaCy* API tools.
SpaCy is a free, open-source library for advanced Natural Language Processing in Python. First we downloaded the trained pipeline called "en_core_web_sm", which is a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax and entities.[21]
The tool we employed is the one working of single tokens, named *token*, and we used the function for POS tagging (see section 2.2 "Part-of-Speech"), the function for selecting punctuation and stop words.

2.2 Natural Language Processing

Natural Language Processing (NLP) is based on computational linguistics techniques. It employs methods from various field, i.e. computer science, artificial intelligence, linguistics, and data science. Thereby, we enable machines to understand human language, somehow "translating" it into a language that the computer is able to "read" and understand. Some common NLP sub-tasks are:

- Summarization: This technique enables to summarize, shortening large texts to shorter once for quicker consumption but still pass the intended

message.

- Part-of-Speech (PoS) tagging: It performs a semantic analysis assigning a tag to every token present in a document based on its part of speech — i.e. denoting nouns, verbs, adjectives, adverbs etc.

- Text Categorization (also Text Classification): This task classifies documents based on predefined topics or categories, and may be also used for categorizing synonyms and abbreviations.

2.3 Knowledge Extraction

Knowledge Extraction (also know as Information Extraction) enables to detect the relevant pieces of data when working with various documents. We can summarize some Information Extraction sub-tasks as:

- Feature Selection (or Attribute Selection): It is typically used in predictive analytic models for selecting important features (dimensions) which would contribute the most to output.

- Feature Extraction: In case of classification models, this process enables to select a subset of features to improve the accuracy of a task and is also used for dimensionality reduction. It works by applying some weight to words in a document, giving a numerical representation of the terminology: in this way, we are able to use that information for understanding of the context of what we are dealing with.

- Named-Entity Recognition: It aims to detect and categorize given entities present in text: for example it can be used to find names - proper nouns - or locations or other elements.[15]

2.4 Data Mining

After building a structured database containing information extracted from the annotated documents provided by NLP algorithms, data is ready to be analyzed, i.e mined. By mining structured data, we are able to draw useful information and to build up new knowledge.[14]

2.5 Text Vectorization

In order to process natural language text and extract useful knowledge from it, we need to covert text into a set of numerical features.
Word vectorization - also known as word embedding - consists of mapping textual elements - which can be terms, keywords, n-grams or even entire sentences - into vectors of real numbers in a given vector space. In this way, documents can be represented as vectors of terms, where each dimension corresponds to a separate term. While, if a term does not occur in any given document, its vector is a zero-vector.
Since a corpus is a collection of documents, when we represent a corpus in the vectorial space, we can depict the dataset as a matrix which rows correspond to every document present in the collection and columns represent terms. Thereby, each single element $C_{dt} = w_{dt}$ of the term-document matrix can be seen as the weight of t-th term in the d-th documents.
The set of all unique terms in the corpus correspond to the vocabulary: $\mathbf{V} = \mathbf{|C|}$, so the dimension of the space n is the total number of the elements in the vocabulary.[16]
There are several ways to convert text into numerical representation, namely:

- Binary Term Frequency (or One-Hot Encoding): We capture in the matrix the absence or the presence of a word in the vocabulary of a given document, inserting in the cell respectively 0 or 1. The main disadvantage of this representation of documents lies in vectors sizes: as a matter of fact, it

is very likely for the resulting matrix to be very sparse, i.e containing a high number of zeros and few ones. In this way, the information provided is very small even the heavy size of the file containing data. Besides, this type of term-document representation lacks in highlighting the "weight" of one term or more terms in a given document, since it only captures the co-occurrence - or not co-occurrence - of words.

- Bag of Words Term Frequency: Starting from the concept of BoW - i.e. collecting a set of terms present in a document and their occurrences, independently from order - this method represents term-document matrix in which the individual cells denote the frequency of each word, computed as the sum of number times a given word appears in a document. As in Binary Term Frequency case, BoW Term Frequency matrix is easy to compute but also very sparse, thus inefficient for computation. In fact, if we consider a given corpus **C** containing a number **D** of documents, having themselves a number **N** of unique tokens in their vocabulary. This implies that the resulting matrix has dimension **D**x**N**. But we are likely to have quite different size of the vocabulary within documents: thus, the amount of cells providing no information - zero cells - would we quite high. Moreover, every time we modify or add new documents to the dataset, we may need to change/increase dimensionality of the matrix, since we would have to add new columns representing new terms in the vocabulary.

- Weighted Term Frequency: We may employ several types of weighted term frequency, based on different mathematical formulas. Some examples are Logarithmic Term Frequency

$$1 + log(tf_{\mathrm{t,d}})$$

computing the logarithm of the frequency and Augmented Term Frequency

$$0,5 + \frac{0,5tf_{\mathrm{t,d}}}{max_{\mathrm{t}}(tf_{\mathrm{t,d}})}$$

which weights the term-frequency. They start from the the conditions of BoW and then normalize the term frequency for all terms occurring in the document. In this way do not only consider simple occurrence, but we also normalize the result, taking into consideration other aspects, like the maximum term-frequency in that specific document.

- TF-IDF: TF-IDF score is based on the ratio between the simple term-frequency and the inverse document frequency.

$$TF - IDF_{\mathrm{t,d}} = tf_{\mathrm{t,d}} \times log(\frac{N}{df_{\mathrm{t}}})$$

  The frequency is computed multiplying simple term-frequency of a term t in document d with logarithm of ratio between the number of documents N and the total number of document containing the term t. In this way, very common terms which would have a high score with a simple term-frequency, in this case happen to have a lower value: in fact the outcome is weighted on the general occurrence of the word in other documents. The more a term is generally common, the more the final score will be affected and lowered. Thus, with TF-IDF we tend to give a higher weight to a given term in a document when it happens to be frequent in that document but not common in general in the corpus.

- Word2Vec: Word2Vec transforms words into vector by providing an embedded representation the words. Given a starting representation of terms, Word2Vec model trains a Neural Network on the corpus. In this way we define a multidimensional space determined by the set of words in the vocabulary. This algorithm is designed to assign close vectors - in dimensional space - to similar words; thus, we can state that the neighbourhood of a term can be considered as its context.

  For example, if we consider couples of terms$(w_{\mathrm{i}}, w_{\mathrm{j}})$, the objective is to maximise the likelihood to find a valid pair of words D, instead of "bad word pairs" $\bar{D}$.

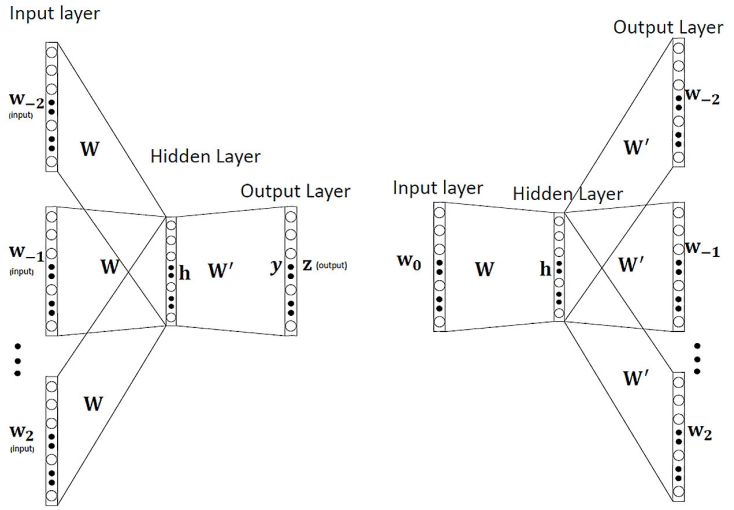

Figure 2.1: CBOW and Skip-gram representations[19]

There are two main algorithms to obtain a Word2Vec implementation: Continuous Bag of Words (CBOW) and Skip-Gram. The CBOW - which is also the default one - is based on the process that, starting from the context words, predicts the vector representation of the target term. The input layer of the neural network consists of the context words in one-hot encoding form with size $1 \times V$ [1]. For every context word, we get the hidden layer resulting from the weight matrix $W_{\mathrm{V} \times E}$. Then, we average them into a single hidden layer, which is passed onto the output layer. Once the training is completed, the weight matrix $W_{\mathrm{V} \times E}$ is used to generate the word embeddings from the one-hot encodings. On the other hand, Skip-Gram is based on the opposite process: starting form the "center", the target word, is able to predict the vector representation of the context words (see Figure 2.1).

[1] Where V is the size of the given vocabulary.

Vectorization	**Function**	**Pros**	**Cons**
Frequency	Counts term frequencies	Bayesian models	Most frequent words not always most informative
One-Hot Encoding	Binarizes term occurrence (0, 1)	Neural networks	All words equidistant, so normalization extra important
TF–IDF	Normalizes term frequencies across documents	General purpose	Moderately frequent terms may not be representative of document topics
Distributed Representations	Context-based, continuous term similarity encoding	Modeling more complex relationships	Performance intensive, difficult to scale without additional tools

Table 2.1: Overview of Text Vectorization Methods[30]

2.5.1 Text Vectorization - Python Implementation

In Python programming language we can implement text vectorization by using *Scikit-Learn* feature instruction functions.*Scikit-Learn* is a free software machine learning library in Python that provides various unsupervised and supervised learning algorithms. In particular, the *Scikit-Learn* functions that return text vectorization are *CountVectorized()* ot *TfidfVectorizer()*.
CountVectorized() converts a collection of text documents to a matrix of token counts, producing a sparse representation of the counts.[20] In our project we employed this function in order to create the term matrix and define the bag of words assigning to each term in the vocabulary its occurrence in the corpus provided.
On the other hand, *TfidfVectorizer()* is used in order to convert a collection of raw documents into a matrix of TF-IDF scores.

2.6 Sentiment Analysis

Sentiment Analysis, also called Opinion Mining, is the field studying people's opinions, sentiments and emotions towards entities such as objects, people or specific organizations.[22]
The origin of this NLP technique can be traced to 1950s, but it became a very active research area only after year 2000.
By implementing Sentiment Analysis we can detect positive or negative sentiment from internal or external data sources. This task can allowing to track, for example, changes in customer attitudes over time, automatically detect opinions in reviews, and so on. It is also commonly used to provide information about perceptions of brands, products and services, and determine user experiences.[15]
It is a proliferating research area in Natural Language Processing, which is also widely studied in data mining, Web mining, and text mining.
The analysis can be applied on three different levels: document, sentence and entity level. The wider level want to detect the opinion of the entire document,

assuming that each sub-element in it expresses an opinion on the same entity and/or a single entity. The second level works on single sentences, and determines whether a sentence expresses a positive, negative or neural opinion. The last one directly detects opinion, which consists of assigning a sentiment -which can be positive or negative - to a given target.[17]

We can use different methods to detect sentiment. One of the most straightforward ways of implementing this analysis is a sort of rule-based method: it consists of defining a lexicon of sentiment terms or opinion words - or phrases - that can be used as indicators for a given sentiment. But this method does not take into account the context, sarcasm and ambiguity issue: some terms may have completely different meaning when used in a manner with respect to another, and this may affect the opinion expressed. On the other hand, it may happen to have some typical sentiment words not implying any sentiment.[17]

Together with the two previous methods, we have also two main types of approach to sentiment analysis, namely: subjectivity/objectivity identification and feature/aspect-based sentiment analysis; the first classifies sentences into objective - if presents some factual general information - or subjective - if expressing some personal feelings or views. In this case the main challenge is that the meaning of a term - or a phrase - is often related to its context, does not only depend on its own meaning.

On the other hand, feature/aspect identification detects opinions and sentiments (features) in relation to different aspects of an entity. Unlike the previous type of identification, feature/aspect based identification allows to have a wider overview of opinions and feelings.[23]

2.6.1 Sentiment Analysis - Python Implementation

Sentiment Analysis in NLP field corresponds to a classification model, assigning positive, negative or neutral based on some given characteristics of a string Thus, the output consists of textual data (document) we want to classify and a

given set of possible classes we can assign to the input text. While the output is the predicted class.

Most of the times Sentiment Analysis is implemented textual data like reviews, by using a given sortable feature present in the dataset - for example, the number of "stars" given by a reviewer - to determine the opinion of the writer. The numerical score is used to classify ex ante the sentiment expressed towards a given entity.

In our case, the implementation of Sentiment Analysis by using Python programming language cannot refer to a label present in the dataset, but has to be derived directly from textual data - the tweets. To achieve it, we used the Sentiment Analysis package of the Python open source for NLP processing called Natural Language Toolkit - also known as *nltk*.

Valence Aware Dictionary for sEntiment Reasoning - i.e. *Vader* - is the *nltk* function for detecting opinions. Its algorithm combines sentiment lexicon approach with grammatical rules and syntactical conventions for defining sentiment polarity and intensity.[24] The sentiment lexicon approach builds a dictionary containing a list of sentiment features, which can be words, phrases, emoticons and sentiment-laden acronyms. These lexical features are manually rated for the polarity and intensity using a numerical scale starting from "Extremely Negative" to "Extremely Positive": the resulting score is the average score given to that feature. *Vader*'s lexicon dictionary contains around 7500 sentiment features in total. The words that are not listed in the dictionary, are automatically be scored as "0: Neutral".[25]

Besides arranging a labelled vocabulary, other structures can influence polarity of a sentence: some elements that may be inherently neutral, but can change the polarity of sentiment - for example "not" and "but" - or can modify the intensity of the entire sentence - for example "very" and "extremely". In *Vader* algorithm several heuristic rules have been incorporated, such as the cases of punctuation, capitalization and contrasting conjunctions.[25]

Vader developers, after collecting and labelling lexicon and grammar rules knowledge, defined the computation of the compound score, the value that classifies polarity of a give string, as the normalized sum (-1, 1) of the scores of features found within the text, using the function:

$$\frac{x}{\sqrt{x^2 + \alpha}}$$

where α is set equal to 15, since 15 is approximately the maximum expected value of x. The more compound score is closer to +1, the higher the positivity of the text, and vice versa.

Furthermore, *Vader* also returns other scores, such as the percentage of positive, negative and neutral sentiment features. Compound score represent the score summarizing the other three percentages.

In particular, *Vader* has a pre-define list on negations - for example "ain't"-, "special case" idioms - for example, "hand to mouth" - and "booster" terms - for example "absolutely". It first looks for negation words, then check if some preceding words increase, decrease or negate/nullify the valence of a word structure. Then it computes scores.[24]

In summary, *Vader* algorithm is based on an embedded lexicon and grammar rules, which can have some cons but also pros: on one hand the number of features are restricted and the lexicon and rules are fixed, even if it is difficult to include and consider all the possible meaning, contexts and exceptions a lexicon construction can have. On the other hand, it is easy to modify and extend the sentimental vocabulary and tailor it to a specific context. Furthermore, *Vader* is computationally economical with respect to other machine learning algorithms that requires massive operation for word embedding and long and heavy training processes.[25]

2.7 Topic Modeling

Probabilistic Topic Modeling is a Natural Language Processing that discovers and explains the enormous collection of documents by reducing them in a topical subspace. It is a unsupervised probabilistic algorithm which isolates the top K topics in a dataset starting from the most relevant N keywords. It can allow to discover hidden themes in a collection, classify documents, and organize or summarize or search the documents.

The idea behind Latent Semantics Indexing (LSI) - also called Latent Semantic Analysis (LSA) - is to find topics or latent concepts that explain data even when using a different terminology for expressing the same concept, by exploiting matrix factorization techniques. In fact, by only using vector space models, we can run into some issues, namely: high dimensionality, synonymy, and ambiguity. These problems can be solved by implementing LSI.

There exist different topic modeling applications, but typically the algorithm used is the Latent Dirichlet Allocation one. The intuition of LDA is to find topics a document belongs to, starting from the information provided: the words present in the document.

Namely, the documents are represented as sets of latent topics, where each topic is characterized by a Dirichlet distribution over a fixed vocabulary and "latent" implies that topics are inferred, rather than simply observed.[28]

LDA is a three-level hierarchical Bayesian algorithm, for which every item of a collection (document) is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities, and it is represented as a probability distribution over vocabulary words.[29] Thus, the number of times a word appear in a given document, depends on the probability that word has in the given topic.

The Latent Dirichlet Allocation model is based on some assumptions:

- A document is essentially a collection of words (bag of words), so the order

of terms in it is not taken into account:

- Every document can be seen as a mixture of topics, each of which has a given proportion in the document;
- Each topic is the summary of a set of words.

The generative process for LDA is given by:

$$\theta_d \sim Dir(\alpha), \phi^{(k)} \sim Dir(\beta), z_i \sim Discrete(\theta_d), w_i \sim Discrete(\phi^{(z_i)})$$

where θ_d is the document d distribution over the topics, $\phi^{(k)}$ is the discrete probability distribution over the vocabulary for the k-th topic, z_i is the topic index for word w_i, and α and β are the hyperparameters for generating Dirichlet distribution.

In the learning phase, after setting arbitrarily the number K of topics to discover

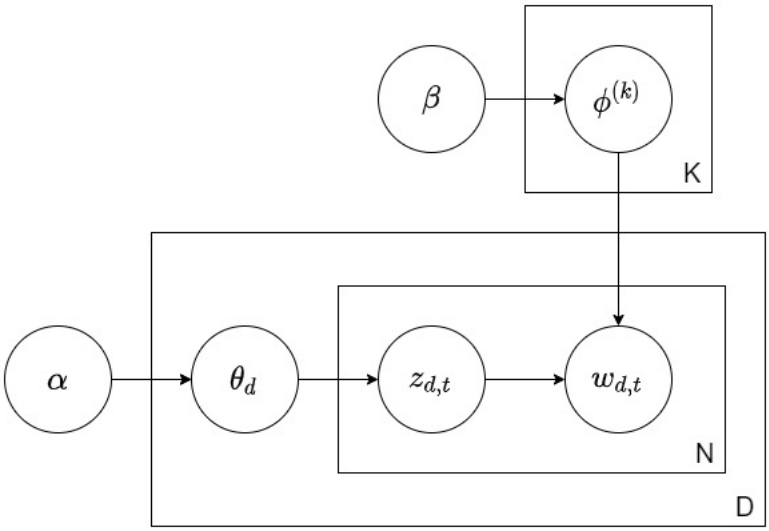

Figure 2.2: LDA process

- the algorithm randomly assigns every word to one of the K topics, creating an initial distribution of words between topics. Then, the algorithm computes the

probability of a given topic t to be assigned to a document d - computed as the proportion of terms belonging to document d assigned to topic t at the first stage - and the probability of a word w to belong to topic t - corresponding to the proportion of assignment to topic t over all documents containing word w. Then, words are reassigned to a new topic, based on the product of the previous probabilities: $P(t|d) \times P(w|t)$. This process is iterated until convergence is reached.[28][27][26]

2.7.1 Topic Modeling - Python Implementation

There are several software packages available in different programming languages for generating topic models; in case of Python programming language, one of the most popular topic modeling tool is provided by *Gensim*, a library based on *numpy* and *scipy* packages. Some of its advantages are scalability - is designed for processing huge amount of data - and a intuitive user interface.
Through *Gensim* workflow, documents are transformed from one vector representation into another, via a lazy fashion disk reading, i.e. one document at a time, without pass the whole corpus into main memory at once. This process allows to reach two main goals: bring out hidden structures lying in the corpus - which can be used to discover relationships between words and describe documents from a semantic point of view - and build a more dense document representation, improving efficiency and efficacy.[31][32]

Chapter 3

Experiment

The goal of the project was to study perception people have on data sharing in research and academical field. In order to achieve it, we collected statements and discussions published on Twitter by users.
For data collection we employed *Tweepy v2*, an open source Python package which provides access to Twitter API using Python programming language.
Tweepy offers a set of functions that can be implemented for various necessities, such as:

- Extracting tweets and data associated with them - e.g. date of publication, user ID, geographical location, conversation ID;
- Retrieve number of tweets published in a given period of time, even specifying a given query detailing features tweets need to have;
- Directly publish, delete, re-post tweets
- Add likes and write comments to other posts.

In particular, for data collection we employed the function *search_all_tweets()*, which is full-archive search endpoint returning the complete history of tweets matching a given search query. The first date tweets retrieved from historical database can have is March 26, 2006.[37]

Another endpoint we used is *get_all_tweets_count()*, which returns the count of tweets that match a query from the complete history of public tweets; also in this case the starting date for retrieval is March 26, 2006.
For extracting opinions on research data sharing published by Twitter users, we looked for tweets written in English language published between March 26, 2006 - starting date - and December 31, 2021.
Data collection was based on specific keywords and hashtags presence. Then, we filtered and cleaned tweets before doing the exploratory analysis; we also performed a sentiment analysis and topic modelling on data, in order to detect relevant terminology and topics addressed in texts.
Then, we did a in-depth study of potential differences emerged between the period right before and after Covid19 emergence: in particular, we repeated the same steps we performed for the entire dataset.

3.1 Data Collection

By using *get_all_tweets_count()Tweepy* endpoint, we collected tweets we used for the experiment. Our purpose was to explore Twitter users' perception about research data sharing; to do so, we first looked for tweets containing specific keywords and hashtags that summarized the topic, namely:

- #Opendata
- Opendata
- Open data
- #Openscience
- Openscience
- Open science
- #Openresearch
- Openresearch
- Open research
- #Datasharing
- Datasharing
- Data sharing

We selected these terms, since we found out that - both in common saying and also in literature - they are often used as synonyms. In fact, it is hard to find a common definition, endorsed by the entire scientific community. While for

European Commission Open data and Open research data are different entities - where the first one consists of the possibility to share data belonging to various contexts and industries, e.g. Government data - other parties depict Open data as data sharing exclusively of the research field.
While gathering information, we found out that Open data, Open science, Open research and Data sharing are frequently used as synonyms. Thus, we collected tweets containing at least one of the keywords above, intending to refine the data set afterwards with finer-mesh filters.
Thus, we run one separate query for each of the specified keywords and hashtags and we merged all the single sets together, dropping duplicates: the resulting initial dataset was a 485268-row data frame.

3.2 Data Cleaning

Starting with merged 485268-row data frame, we cleaned the initial set by removing tweets containing specific terms and phrases: in fact, by manual scanning the initial data frame, we found out that among the collected tweets we gathered also some bots' ones, recurring texts and tweets referring to different topics with respect to project theme field. In particular, *#Opendata* is widely misused on Twitter, sometimes also inappropriately added at the end of a tweet about a whatever topic, far apart from open data actual meaning.
Some examples of tweets pertaining to different contexts are the one promoting specific events - e.g. webinars, meeting - about Open data, or publication of books and articles about this theme. Other tweets refer to Governments data, or data behind journal articles or TV news reports. Others publish job advertisements, providing information about previous experience in data sharing, indeed.
First, we removed the collected tweets having a higher number of tags with respect to the actual number of other terms: in fact, we noticed that tweets having with this characteristic tended to be very short answers to other tweets, bringing just little - or even lack of - information. From collected texts, we derived

removed from tweets hashtags, tags and websites, and we stored them into separated and specific columns.
In text cleaning phase, we turned all words into lowercase, we removed numbers, punctuation, contractions and we lemmatized the tweets. Lemmatization is a word normalization technique which turns inflected words into their root form - namely, the lemma. After this, we saved the resulting cleaned string into a new column in the dataframe. In this step, we also expanded some contractions and dropped stopwords by using *spaCy* library.
Furthermore, by retrieving tweets written by the top 100 users with the highest number of published tweets inside the collected set of tweets - the total amount of unique users are 90934 - we were also able to detect potential spamming of tweets or bots, and, thus, remove them from data.
In order to be able to run better more specific analysis afterwards, we created a further column in which, starting from cleaned text, we also removed the terms we used in queries, during data collection. The query terms correspond to:

- Data
- Open
- Openresearch
- Research
- Science
- Datasharing
- Sharing
- Opendata
- Openscience.

The dataset resulting after running these passages was a 224784-row data frame: we discarded 260484 tweets, 46% of total number of rows of the raw initial set.

3.2.1 First Data Filter

Since the cleaned set of tweets we obtained by simply collecting and cleaning data appeared to be quite heterogeneous, and sometimes not fitting the reach goal field, we selected a specific terminology to look for in tweets. Thus, to obtain a set of tweets pertaining to academic/research area, we selected some terms that could summarize the field of interest, and we used them to filtered

data. Among tweets in the 224784-row data frame we built, we selected tweets containing at least one of the following stems:

- Universit
- Academ
- Research
- Publi
- Prof
- Availab
- Phd (or P.h.d)
- Discover
- Reproducib
- Scientific communit
- Scien

After merging the results and dropping duplicates, the resulting data frame was reduced from 224784 to 97645 rows: 20% of raw initial data and 43% of the cleaned dataset.

3.2.2 Second Data Filter

Apart from academic and research context, it is relevant for our project to have an in-depth view of open research data field. To have tweets only about data sharing - among the ones present in "first filter" data - we looked for tweets having at least one term suggesting the concept of datum, repositories, knowledge collected during the research.
We kept only tweets containing at least one of the following stems:

- Data
- Datum
- Openness
- Info
- Find
- Discover
- Stat
- Knowledg

The resulting data frame was reduced from 97645 to 11959 rows: it represents 2.5% of raw initial data, 5,3% of dataset after cleaning and 13.7% of first filter data. Furthermore, we found out that none of the tweets collected belonged to 2006, nor 2007: the oldest tweets extracted after this filter were published in 2008.

3.3 Exploratory Analysis

The first aspect we explored was the timeline of usage of each keywords and hashtags. We saw that the first one that has been used in a tweet is "open data" in 16/02/2007, while the last was "#openresearch" in 02/06/2009. In particular, we noticed that for all the keywords, the first format to be used was the non-hashtag form, with separated words - e.g. "data sharing" - then the merged form of the word - e.g. "datasharing" - and the last one in time was the hashtag format - e.g. "#datasharing".

As regards amount of tweets published using the keywords and hashtags, we can notice an increasing trend towards years, with a small flexion around 2017: the trend we found by counting amounts of tweets we extracted in each year is also supported by a counter-check we made by looking for timeline of amount of scientific publications registered on *PubMed.gov* website[1] Distinguishing between

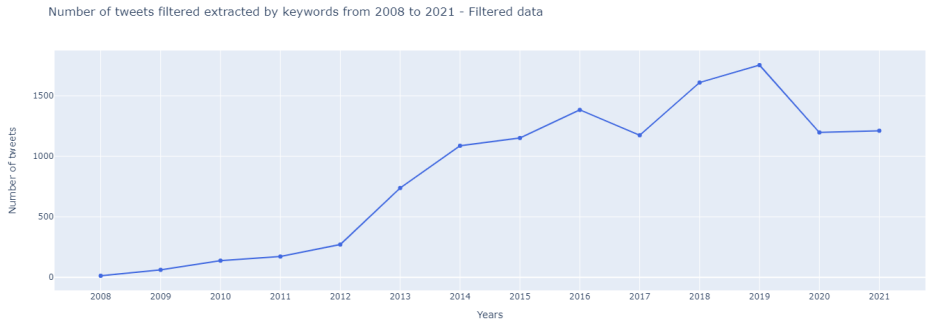

Figure 3.1: Number of tweets extracted by keywords from 2008 to 2021

[1] `https://pubmed.ncbi.nlm.nih.gov/`.

hashtags and keywords, we notice that "#opendata" is the most used in absolute terms - likely because of the reasons we stated previously, i.e. the overlapping of meaning and use of this terms as summary of all the other terms. Although, it is evident that the usage of "#opendata" is decreasing in the last years, in particular, starting from 2016.

As regards fluctuation of usage of the terms and hashtags within months, we can notice a seasonality: publication of tweets about open research data follow academic and scholar year - opening and holidays - and we have peaks in periods in which typically most of the events, conferences and summits are scheduled. This can imply that this kind of events are attended and values, and thus, there exists a spread interest towards the topic.

To show how much the cleaning phase was able to reduce the number of tokens

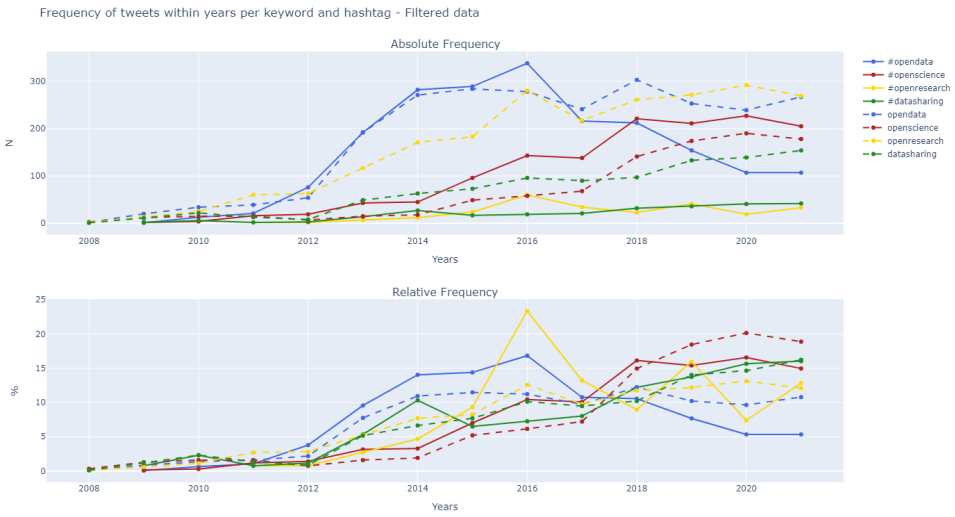

Figure 3.2: Frequency of tweets within years per keyword and hashtag

present in tweets, we plotted two histograms describing the amount of tokens - atomic elements in textual data - before and after cleaning tweets. In particular, while raw data had 27 token per tweets on average, after cleaning we have around 12 tokens, nearly a half of the starting texts. This is mainly driven by tags,

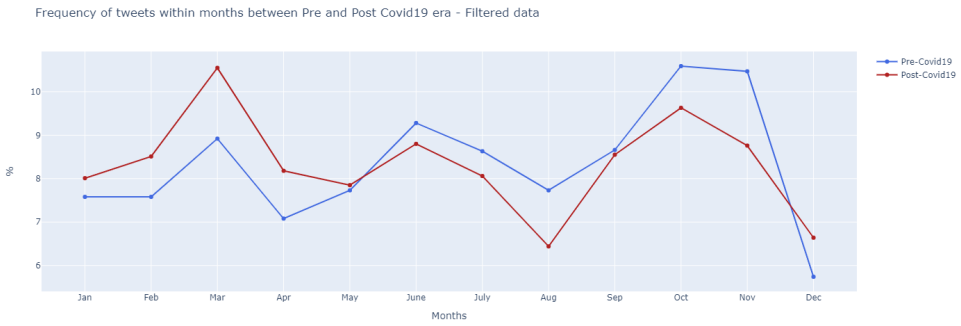

Figure 3.3: Frequency of tweets within months per keyword and hashtag

hashtags, links and stopwords removal.

Then, we focused on occurring hashtags: by looking for most popular hashtags

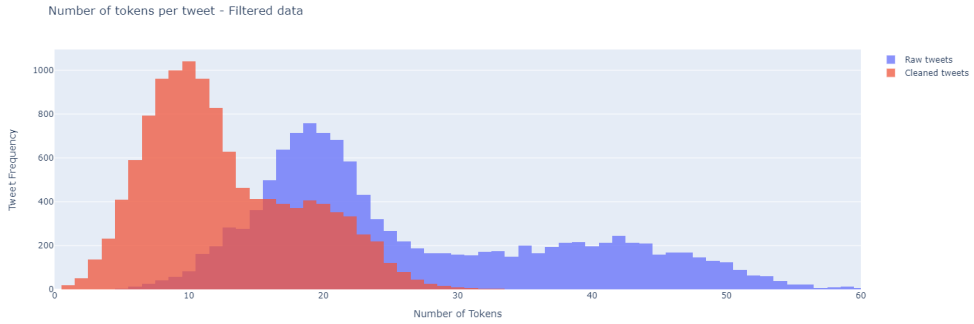

Figure 3.4: Number of tokens per tweet before and after cleaning data

in the set, we also study possible correlations between them. By looking at Figure 3.5 we notice a slight negative correlation between two of the query hashtags: opendata and openscience. This suggests that these two terms may be used as synonyms in tweets.

3.4 Relevant Terminology

Looking for differences between years in terms of lexicon, topic and perception changes along time, we looked for relevant terminology, both collecting most used

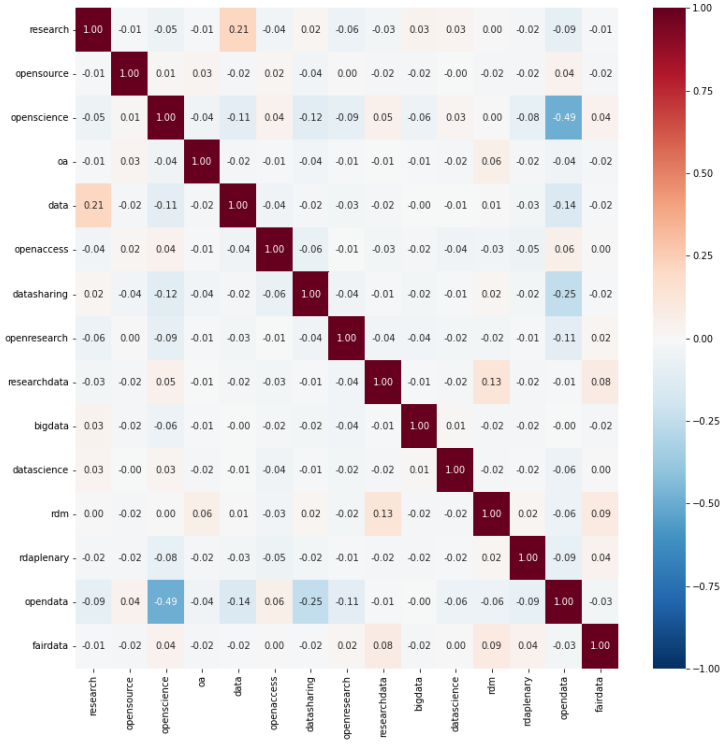

Figure 3.5: Correlations between top 15 most frequent hashtags

single words, bi-grams and tri-grams and also computing Tf-Idf score by classes - namely, dividend into year of publication.

Moreover, in this section we will refer to some tweet examples for valuing the outcomes presented.

3.4.1 Most Common Single Terms

Firstly, in this section we used the cleaned text in which we removed query terms from tweets (see Section 3.2): in this way we avoided finding in all classes the same set of terms, which are for sure present, since they represent the *contitio sine qua non* for being selected and extracted through *Tweepy* API.

After deriving the 50 most frequently used words in each year, we built the

Figure 3.6: Most frequent terms of the set after second filter

intersection between years, in order to visualize which top frequent terms are in common between all years[2]. From Figure 3.7 we know that words that happen to be frequent among selected years are: “publish”, “datashare”, “need”, “public”, “researcher”, “share”, “access”, “good”. They refer to openness and accessibility of data, and express a positive opinion towards this topic.

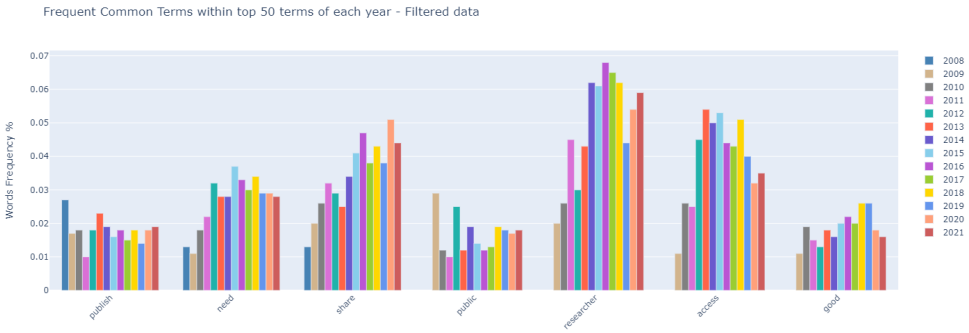

Figure 3.7: Frequent Common Terms within top 50 terms of each year

[2]Please note: we did not consider 2008 computing the intersection between top 50 most frequent terms, because of its lower numerousness with respect to the other years.

3.4.2 Most Common Bi-Grams and Tri-Grams

Moving on to bi-grams and tri-grams analysis, we looked for most occurring sets of subsequent terms. Basically, bi-grams correspond to two consecutive terms in textual data, while tri-grams are composed of three terms. These constructions are able to suggest and depict topics addressed in the tweets present in the set.

The most common bi-gram is "publicly fund": by reading tweets containing

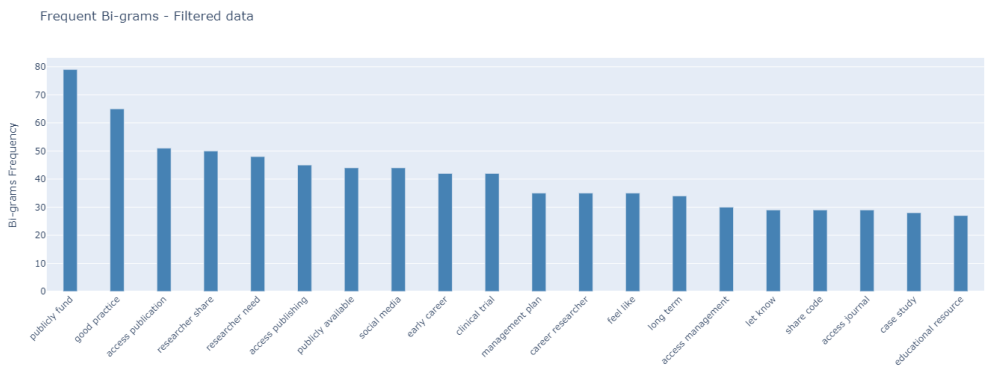

Figure 3.8: Frequent bi-grams

this expression, we notice that Twitter user stated that, apart form make public research in general, but especially researches that are publicly fund, should be publicly available - also "publicly available" is between most common bi-grams. For example:

> "(...) Publicly funded data sets, and research publications, should be open access.".

Another interesting result is the reference to clinical trial: as a matter of fact, a lot of the tweets collected refer to health and medicine area, stating that sharing data of researches may fasten and enhance the entire sector, leading to better quality of life and defeat diseases.

"#Datasharing has incredible potential to strengthen (..) research, the practice of medicine, & the integrity of the clinical trial system.", "Perspective: Data sharing can strengthen research, medicine & the integrity of clinical trials".

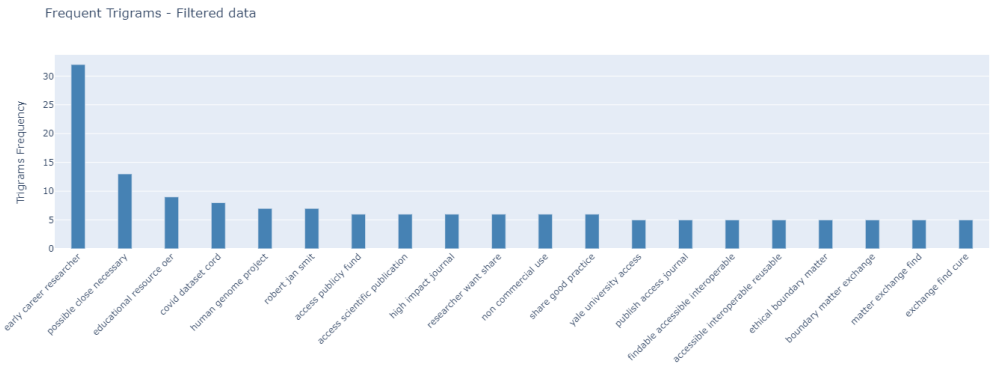

Figure 3.9: Frequent tri-grams

The most common tri-gram is "early career researcher" - where "early career" was also present between most frequent bi-grams - and points to tweets stating that open research data can be a useful tool also - and mainly - for beginner researchers. Also, some other tweets claim that younger researchers should be the ones to push and encourage this practice.

"#OpenResearch is the means to reward early career researchers for much of the work they do generating data and writing code by documenting it and making it available and citable.",

"Possible close necessary" is the tokenized form of the quote "as open as possible, as closed as necessary", which refers to the definition of FAIR data[38]. While, "human genome project" refers to HGP, the collaborative international research program, which goal was the complete mapping and understanding of all the genes of human beings[39]. This project has been taken as example to prove that by sharing data, research and knowledge science can reach significant goals.

> "Sharing data can save lives. The 'Bermuda Principles' for public data disclosure are a fundamental legacy of producing the first human reference DNA sequence during the Human Genome Project (HGP). #OpenScience #OpenAccess"

By looking at most occurring tri-grams, we also discovered that a fair share of tweets basically consist of quotes and references to something someone said: for example "Robert-Jan Smits" is a former Director-General of DG Research and Innovation (RTD) at the European Commission from 2010 to 2018. Some of his views and statements towards the future of research sharing in Europe.

> "From the Nature Briefing: Robert-Jan Smits: ' will be the ice-breaker for open access publications." The departing European Union research chief, will lead the EU's charge to make publicly funded scientific papers and data freely available by 2020(...)'",
>
> "Only 40% @ScienceEurope funders require a DMP still. Robert-Jan Smits calls for action to drive towards open data by 2020 target. #researchdata #rdm"

3.4.3 Tf-Idf by Classes

In order to fint relevant terminology, we also employed Tf-Idf score, computing significant lexicon for each year. We based the computation following the code of *cTFIDF* GitHub repository by *MaartnGr*[40].
In particular, class-based Tf-Idf computes typical Tf-Idf score after merging documents by class - in our case classes were years form 2008 to 2021. We set N[3] equal to the length of the dataframe - i.e. 11959 - instead of setting it equal to the number of pseudo-documents - i.e. 14, namely the number of years in the set of tweets. In this way, we weighted Tf-Idf scores on the total amount of tweets present in the corpus.

As shown in Tables 3.1, 3.2, 3.3, 3.4, we were able to gather very little pieces

[3]Total number of documents

2008	score	2009	score	2010	score	2011	score
community	0.024	public	0.012	university	0.008	researcher	0.008
astronomical	0.023	available	0.009	report	0.008	publication	0.008
responsibility	0.023	online	0.007	available	0.007	fund	0.008
network	0.022	pryor	0.007	project	0.007	share	0.007
online	0.022	pancake	0.007	good	0.006	academic	0.006
grp	0.021	publish	0.007	share	0.006	incentive	0.006
string	0.021	discovery	0.007	access	0.006	need	0.006
user	0.021	gain	0.007	publish	0.006	university	0.005
aids	0.020	set	0.006	freely	0.006	access	0.005
scl	0.020	result	0.006	researcher	0.005	good	0.005

Table 3.1: Tf-Idf scores from 2008 to 2011.

2012	score	2013	score	2014	score	2015	score
access	0.008	access	0.008	researcher	0.007	access	0.006
fund	0.008	researcher	0.005	access	0.006	researcher	0.006
need	0.007	academic	0.005	share	0.004	need	0.005
public	0.007	publish	0.005	need	0.004	share	0.005
publicly	0.006	need	0.005	talk	0.004	academic	0.004
publication	0.006	talk	0.004	publish	0.004	university	0.004
share	0.006	publication	0.004	public	0.004	use	0.003
scientific	0.006	share	0.004	big	0.003	good	0.003
publish	0.005	big	0.004	fund	0.003	work	0.003
researcher	0.005	new	0.004	academic	0.003	new	0.003

Table 3.2: Tf-Idf scores from 2012 to 2015.

2016	score	2017	score	2018	score	2019	score
researcher	0.007	researcher	0.006	researcher	0.004	work	0.003
share	0.005	access	0.005	access	0.004	access	0.003
access	0.005	share	0.005	share	0.004	share	0.003
need	0.004	need	0.004	need	0.003	like	0.003
good	0.003	talk	0.004	work	0.003	researcher	0.003
academic	0.003	new	0.003	good	0.003	good	0.003
university	0.003	work	0.003	think	0.003	people	0.003
use	0.003	good	0.003	like	0.003	use	0.003
publish	0.003	great	0.003	use	0.003	need	0.003
talk	0.003	source	0.003	way	0.003	think	0.003

Table 3.3: Tf-Idf scores from 2016 to 2019.

2020	score	2021	score
share	0.005	researcher	0.004
researcher	0.004	share	0.004
work	0.004	work	0.004
find	0.003	like	0.003
need	0.003	access	0.003
like	0.003	knowledge	0.003
time	0.003	community	0.003
paper	0.003	need	0.003
access	0.003	people	0.003
covid	0.003	university	0.003

Table 3.4: Tf-Idf scores from 2020 to 2021.

of information from TF-Idf scores computed: mostly, relevant terms provided happened to be shared by also the other classes. For example, university and academic references are present in 2010, 2011, 2013, 2014, 2015, 2016 and 2021. Also, "access" references were very common between 2012 and 2019, being one of the top three relevant terms for all those years.
On the other hand, we found also some peculiarities: tweets published in 2012 had a lot of references to making public (see 3.2 "public", "publicly", "publication", "publish"); while in other cases we found references to specific circumstances and events, for example "aids" in 2008 and "covid" in 2020.

3.5 Topic Modeling

By employing *gensim* Latent Dirichlet Allocation function, we generated five topics, defined by 10 different terms. It was trained on the 11959-rows set - second filter data - using the textual data from which we removed the query terms (see Section 3.2), which set chuck-size[4] equal to 200. From Figure 3.10 we

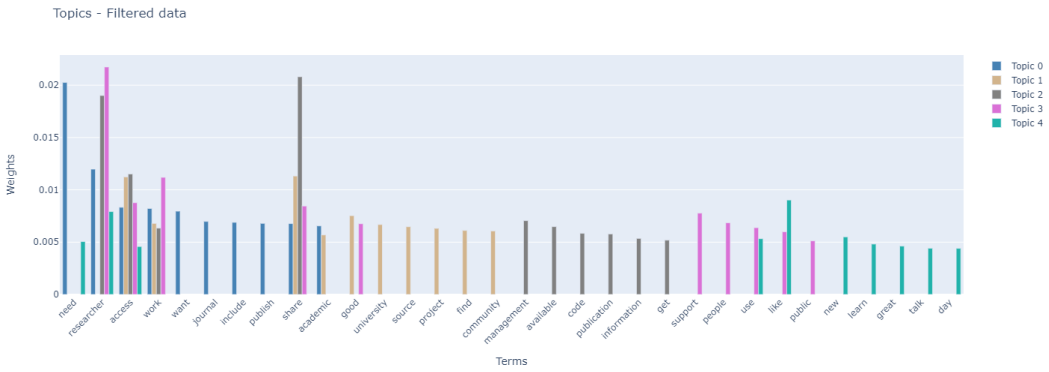

Figure 3.10: Topic derived by employing *gensim* LDA model

identify a group of tweets referring to group working and collaboration - Topic 4 - another around journals and publications - Topic 3 - while Topic 2 may point

[4]Chuck-size value corresponds to the number of documents to consider at once. It affects the memory consumption.

out difficulties and issues of this practice. Topic 0 representative terms, on the other hand, are mostly shared with the other topics, but we can notice that there exists also a group of tweets in which users talk about data management, research and health.

3.6 Sentiment Analysis

To discover sentiment expressed in collected tweets, we employed *Vader* sentiment analysis rule-based algorithm. Firstly, we studied sentiment changes in time: we focused on years 2018, 2019, 2020 and 2021. As Figure 3.11 shows, general manifested sentiment became more and more positive over the years. Although, the largest part of the tweets have been classified as neutral - value 0 of sentiment score: this happened because *Vader* package has a restricted number of labelled terms in the pre-defined vocabulary. Thus, if we provide textual data containing words that are not present within the labelled lexicon, the algorithm returns automatically a neutral score, i.e. zero. Then, we focused on

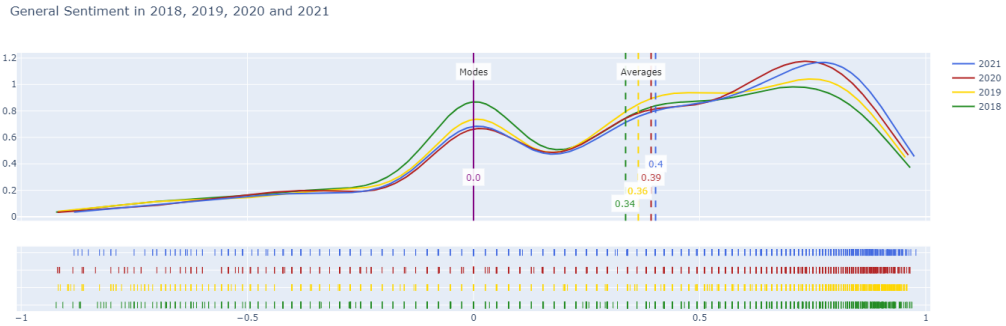

Figure 3.11: Sentiment score changes in time

differences between query hashtags: #opendata, #openscience, #openresearch, #datasharing. #Opendata had the highest concentration of tweets in neutral zone, and, accordingly, had a smaller amount of tweets classified as positive or negative. On the other hand, we can say much about the other hashtags: we can

point out that #openresearch and #openscience seemed to be the ones having higher average sentiment score(see Figure 3.12). However, their score - 0.33 - is not high enough to highlight positiveness, since typically in literature value equal and above 0.5 are considered positive.

Thirdly, we plotted sentiment variations over years for each hashtag. While #opendata, #openscience and #openresearch increase on average sentiment score over years, tweets having hashtag #datasharing dis not follow the same trend. Furthermore, the only class that really reached 0.5 is the average sentiment score of #openresearch in 2021.

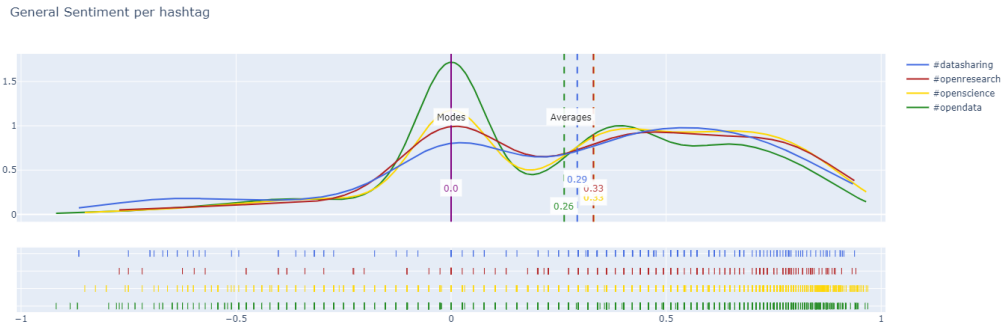

Figure 3.12: Sentiment score differences between hashtags

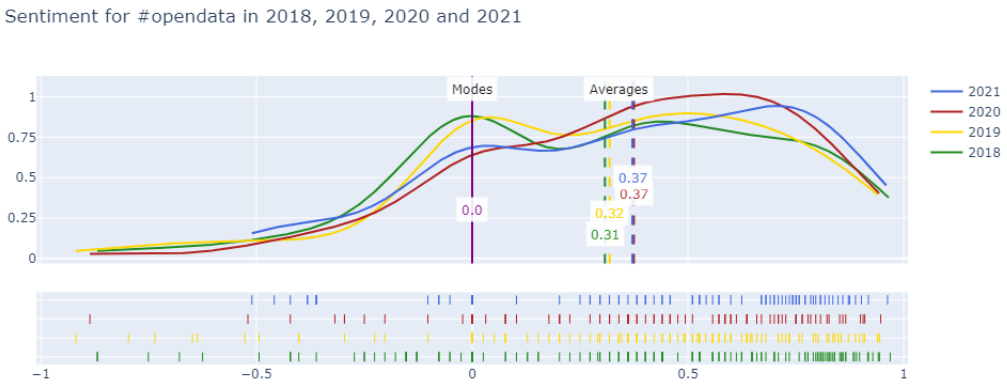

Figure 3.13: Sentiment score changes in time - #opendata

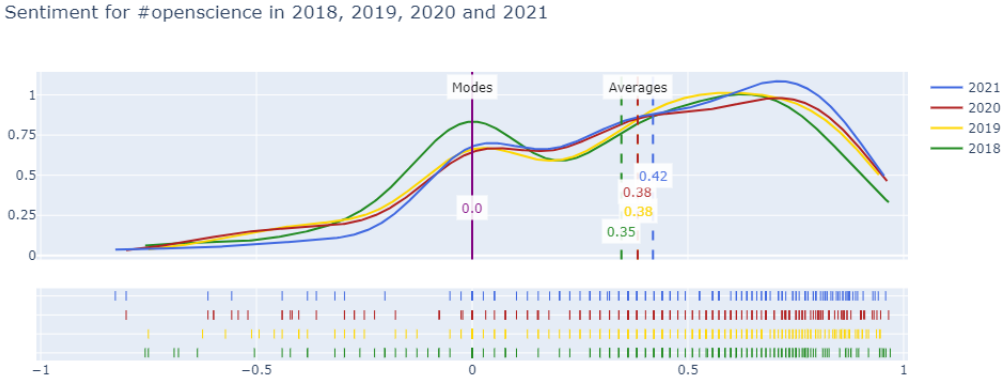

Figure 3.14: Sentiment score changes in time - #openscience

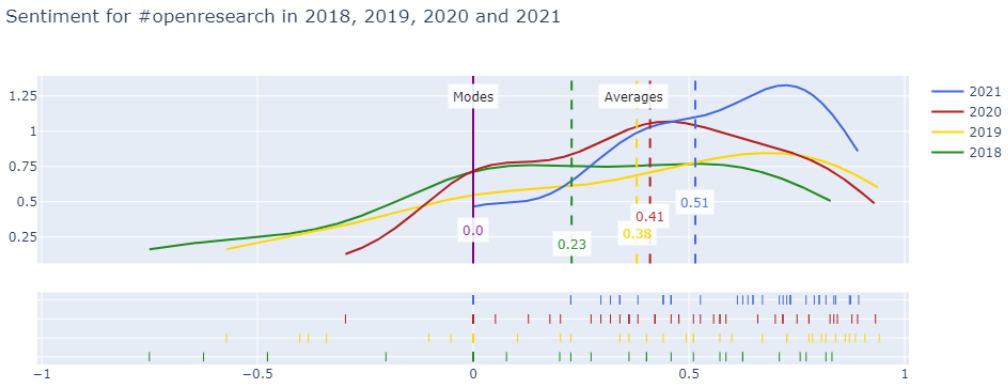

Figure 3.15: Sentiment score changes in time - #openresearch

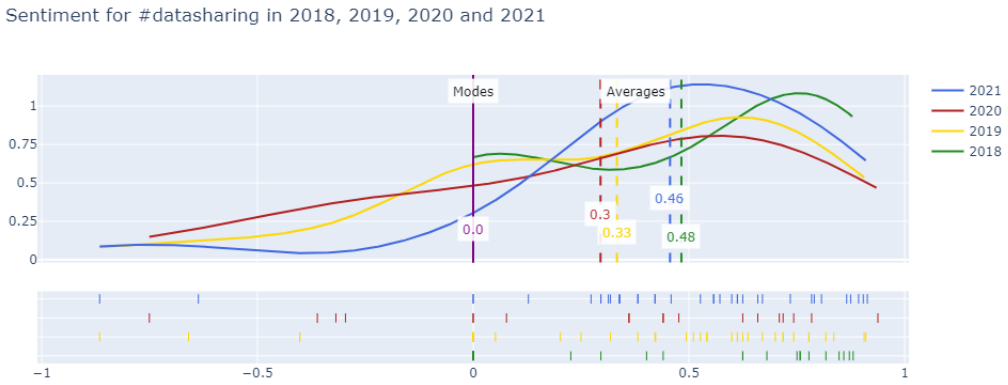

Figure 3.16: Sentiment score changes in time - #datasharing

3.7 Exploratory Analysis - Pre-Post Covid19 Era

Moving on, we implemented a in-depth study of potential differences between the period right before and after Covid19 emergence: to do so, we selected only the last four years of the set and divided into Pre-Covid19 era - years 2018 and 2019 - and Post-Covid19 era - years 2020 and 2021. In particular, we repeated the same steps we implemented on the set produced after the second data filter

(see Section 3.2.2).

By looking at the timeline on publication over months, we noticed the same seasonality found in the largest dataset: also in this case, the relative frequency follows academic year and conferences calendars.

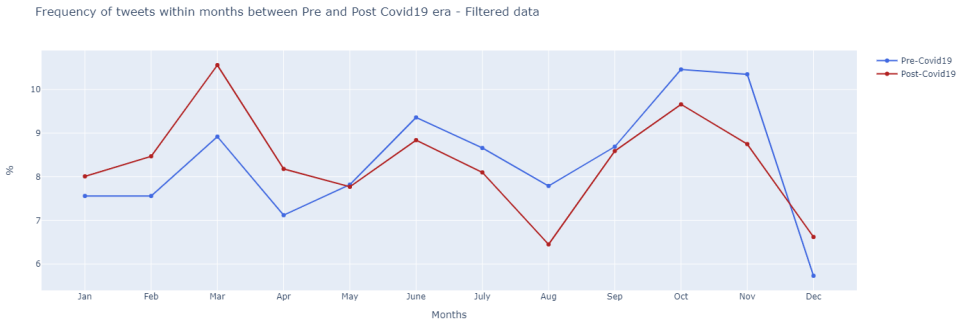

Figure 3.17: Frequency of tweets within months - Pre-Post Covid19 era

3.8 Relevant Terminology - Pre-Post Covid19 Era

As we did in Section 3.4, we collected most frequent occurring words, bi-grams and tri-grams and we computed Tf-Idf score by classes.

3.8.1 Most Common Single Terms- Pre-Post Covid19 Era

We retrieved the 50 most frequently used words in each class - Pre-Covid19 era and Post-Covid19 era- and we built the intersection between the two. In this way we obtained frequent lexicon common to the two periods. From Figure 3.18 we know that the two periods have most of the top 50 terms in common: in particular, 42 terms are part of the intersection. Only 8 words for each class - 16 terms in total - were not part of the intersection: In particular, we can notice that, while in Pre-Covid19 era the terms may refer to data analysis and data management, after Covid19 emergence tweets tended to highlight the knowledge, information, resource and value data sharing can provide when implemented (see Figure 3.19).

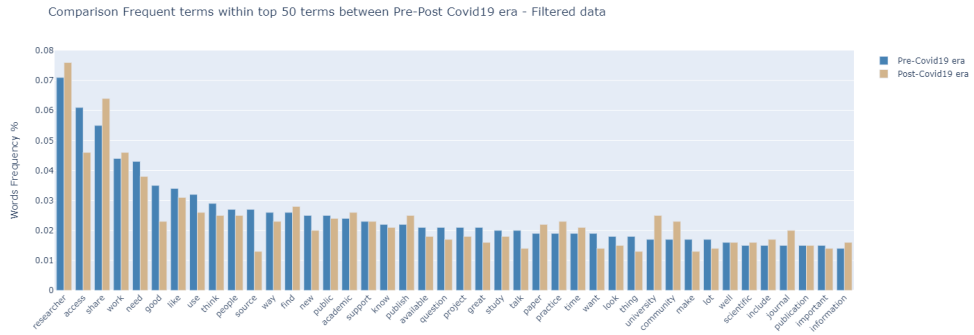

Figure 3.18: Comparison of frequent terms within top 50 terms between Pre-Post Covid19 era

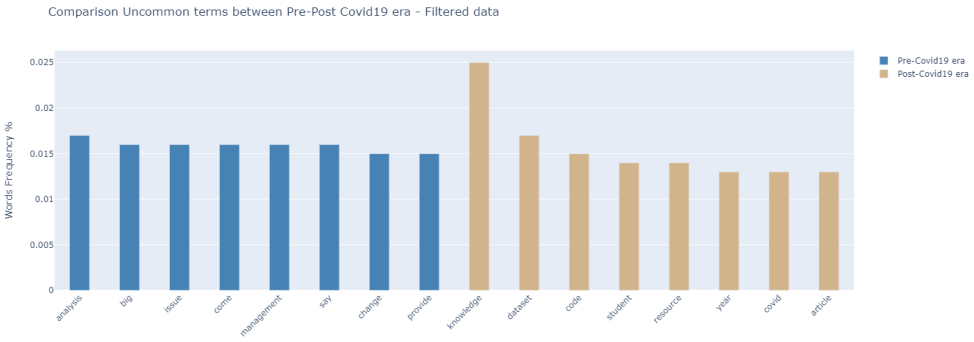

Figure 3.19: Comparison of uncommon terms within top 50 terms between Pre-Post Covid19 era

3.8.2 Most Common Bi-Grams and Tri-Grams - Pre-Post Covid19 Era

By comparing period before and after Covid19 emergence in occurring bi-grams, we cannot notice much differences nor relevant outcomes: we can point out is that in the Post-Covid19 class we had references to Coronavirus within the most common bi-grams - i.e- “covid dataset” and “dataset cord” (see Figure 3.20 and 3.21).

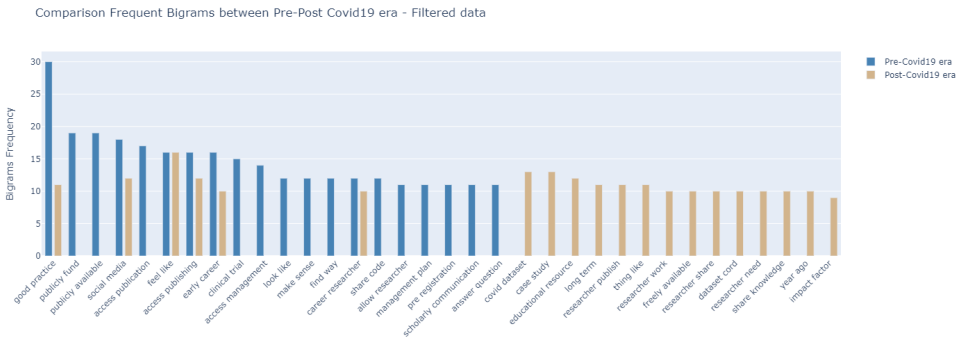

Figure 3.20: Comparison of frequent bi-grams terms between Pre-Post Covid19 era

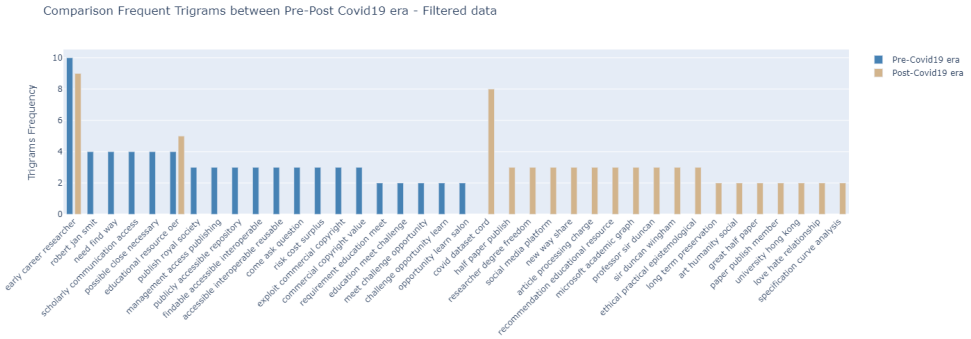

Figure 3.21: Comparison of frequent tri-grams terms between Pre-Post Covid19 era

3.9 Topic Modeling - Pre-Post Covid19 Era

By employing *gensim* Latent Dirichlet Allocation function, we generated five topics, defined by 10 different terms for each of the two classes.
Firstly from Figure **??** we noticed that topic-words found for the period before Covid19 world emergency are a little more shared between topics, while in the bottom part of the image sets of terms are more separated.
As we found in Section 3.5 when looking for frequent lexicon, Pre-Covid19 topics highlighted themes as management, practice and risk behind sharing data. We

also found references to the economic aspect: for example “grant” and “commercial”.

On the other hand, tweets published in 2020 and 2021 happened to draw attention to other topics: apart from academic and publication area, one of the main focus is on medical aspect - “death”, “healthcare” and “incidence”. Another piece of information we extracted, was derived from Topic 0 which points out the enlarged point of view in geographical terms: as a matter of fact, between Topic 0 terms we found “community”, “international”, and “world”.

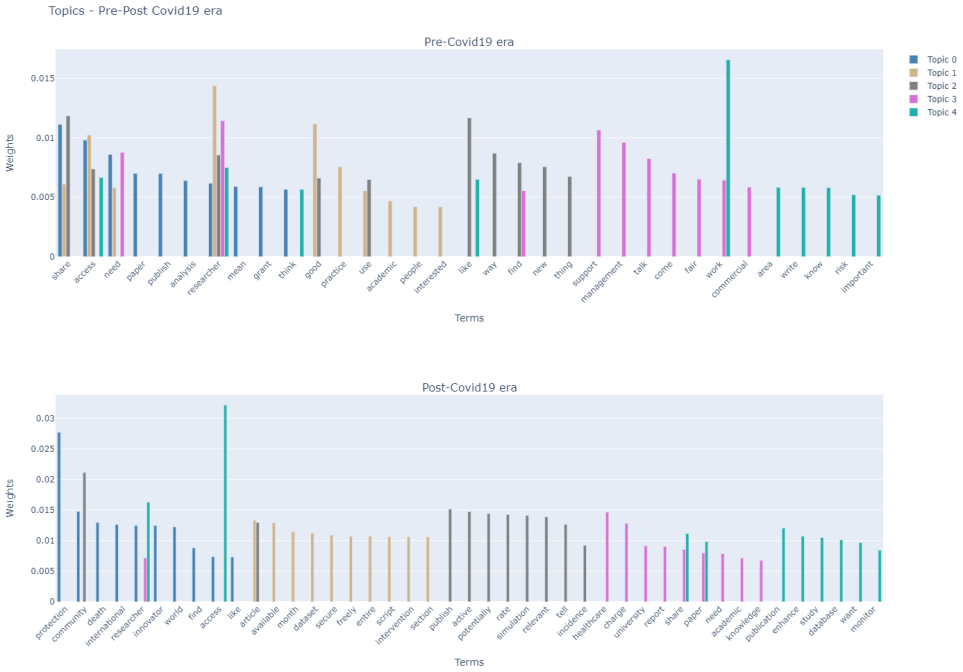

Figure 3.22: Topics - Pre-Post Covid19 era

Concluding Remarks

Through this project we collected, studied and analyzed opinions about research data sharing, by applying several text mining techniques. Considering tweets published between 2006 and 2021, we extracted opinions on data sharing theme in academical field published by Twitter users.

After data collection - based on specific keywords and hashtags - we cleaned and filtered the set of data: by cleaning sets of tweets, we found out that within all possible fields lying in the multifaceted open data - and open research, open science - topic, the subject of interest of this project - i.e. open research data - represents only a small part of it: in Section 3.2.2 we filtered for tweets explicitly referring to data, and they correspond to 2.5% of raw initial data - after simply collection from Twitter API using query on keywords -, the 5,3% of the cleaned dataset and 13.7% of first filter data, which represents the specifically academic/university field.

In the project we also studied the topic from a chronological point of view: we discovered that interest in data sharing and open science is increasing over time. Moreover tweets' publication tend to follow academic years and also dedicated events and conferences calendar, mainly scheduled in the period around spring - March and April - and autumn - from September to November, reflecting interest and attendance in dedicated summits.

Some of the main specific engaging points highlighted in tweets are: public availability of data produced by publicly funded researched; interest in increasing

usage of data sharing in medical and clinical industry; adduction of Human Genome Project as proof of well-functioning of data sharing system; attention towards early career researchers, who should also be data sharing promoters; claiming keeping data as open possible, as closed as necessary.

By studying specifically periods right before and after Covid19 pandemic world emergence, we found out some differences and peculiarity of those periods. In particular, while Pre-Covid19 era is characterized by discussions about data management, data analysis and economic aspects, in tweets belonging to Post-Covid19 emergence period users highlighted knowledge, information, resource and value data sharing could provide when fully employed, and geographical references emerged.

Bibliography

[1] Foster website, *Open Science Definition*. Available at: https://www.fosteropenscience.eu/content/open-science-scientific-research.

[2] Di Giorgio Sara. *Open Data, Open Science, Open Access*, 2017.

[3] Sonja Bezjak, April ClyburneSherin, Philipp Conzett, Pedro Fernandes, Edit Görögh, Kerstin Helbig, Bianca Kramer, Ignasi Labastida, Kyle Niemeyer, Fotis Psomopoulos, Tony RossHellauer, René Schneider, Jon Tennant, Ellen Verbakel, Helene Brinken, & Lambert Heller, *Open Science Training Handbook (1.0)*, 2018. Available at: https://doi.org/10.5281/zenodo.1212496 and https://open-science-training-handbook.github.io/Open-Science-Training-Handbook_EN/.

[4] Benedikt Fecher, Sascha Friesike, *Open Science: One Term, Five Schools of Thought*, 2013. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2272036.

[5] Gewin Virginia. *Data Sharing: An Open Mind on Open Data.*, Nature (London), 2016.

[6] European Commission, *Digital science in Horizon 2020: Concept paper*, 2020. Available at: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en.

[7] David Paul A., *The Historical Origins of 'Open Science': An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution*, 2013, Capitalism and Society, Vol. 3, Issue 2, Article 5, 2008. Available at: https://ssrn.com/abstract=2209188.

[8] Molloy JC *The Open Knowledge Foundation: Open data means better science*, 2011. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3232214/.

[9] Vogt Thomas, *Reinventing Discovery: The New Era of Networked Science*, Princeton U. Press, 2012.

[10] Green, Samantha. *An Illustrated History of Open Science*, 2017. Available at: https://www.wiley.com/network/societyleaders/open-science/an-illustrated-history-of-open-science.

[11] Paul A. David, *Understanding the emergence of 'Open science' institutions: functionalist economics in historical context*, Industrial and Corporate Change, 2004. Available at: https://doi-org.pros1.lib.unimi.it/10.1093/icc/dth023.

[12] *Budapest Open Access Initiative Declaration*, Budapest, 2002. Available at: https://www.budapestopenaccessinitiative.org/read/.

[13] Univeristy of Exeter website. *Open research definition*. Available at: http://www.exeter.ac.uk/research/openresearch/about/explained/

[14] Nadir Zanini, Vikas Dhawan, *Text Mining: An introduction to theory and some applications*, Research Matters: A Cambridge Assessment publication, 2015. Available at: https://www.cambridgeassessment.org.uk/Images/466185-text-mining-an-introduction-to-theory-and-some-applications-.pdf.

[15] IBM Cloud Education, Text Mining, 2020. Available at: https://www.ibm.com/cloud/learn/text-mining.

[16] Aggarwal C. C., Zhai C., *Mining text data*, Springer Science & Business Media, 2012.

[17] Liu B., *Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1)*, pages 1-167, 2012.

[18] McCormick Chris, *Word2Vec Tutorial - The Skip-Gram Model*, 2016. Available at: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/.

[19] Dhananjay Kimothi, Pravesh Biyani, James M. Hogan, Akshay Soni, Wayne Kelly, *Word2Vec architecture: The figure shows two variants of word2vec architecture—CBOW and Skip gram*, 2020. Available at https://figshare.com/articles/figure/_i_Word2Vec_i_architecture_The_figure_shows_two_variants_of_word2vec_architecture_CBOW_and_Skip_gram_26_for_a_sample_/11982951/1.

[20] Sklearn API guide website, *sklearn.feature_extraction.text.CountVectorizer* Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html.

[21] SpaCy API guide website. Available at: https://spacy.io/.

[22] Bing Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, 2012.

[23] Scott Sims, *Sentiment Analysis 101*. Available at: https://www.kdnuggets.com/2015/12/sentiment-analysis-101.html#:~:text=Sentiment.

[24] Nltk.sentiment.vader API guide website, *Documentation: nltk.sentiment.vader module*. Available at: https://www.nltk.org/api/nltk.sentiment.vader.html

[25] Ying Ma, *NLP: How does NLTK.Vader Calculate Sentiment?*, medium.com, 2020. Available at: https://medium.com/ro-codes/nlp-how-does-nltk-vader-calculate-sentiment-6c32d0f5046b

[26] Uttam Chauhan, Apurva Shah, *Topic Modeling Using Latent Dirichlet allocation: A Survey*, Article 145, 2021. Available at: https://doi-org.pros2.lib.unimi.it/10.1145/3462478.

[27] Shrivarsheni, *Text Summarization Approaches for NLP – Practical Guide with Generative Examples*, machinelearningplus.com, 2020. Available at: https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/.

[28] Rosaria Silipo, *LDA for Text Summarization and Topic Detection*, dzone.com, 2019. Available at: https://dzone.com/articles/lda-for-text-summarization-and-topic-detection.

[29] Blei, David M., Andrew Y. Ng, Michael I. Jordan, *Latent dirichlet allocation.*, Journal of machine Learning research 3, 993-1022, 2003.

[30] Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda, *Applied Text Analysis with Python*, Chapter 4: Text Vectorization and Transformation Pipelines, 2018. Available at: https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/ch04.html.

[31] Micah D. Saxton, *A Gentle Introduction to Topic Modeling Using Python*, 2018. Available at: https://serials.atla.com/theolib/article/view/2609/3271.

[32] Islam Akef, Juan S Munoz Arango, *Mallet vs GenSim: Topic modeling for 20 news groups report*, University of Arkansas at Little Rock, Information Science, Principles and Theory, 2016. Available at: https://www.researchgate.net/profile/

Islam-Ebeid/publication/331972126_Mallet_vs_GenSim_Topic_Modeling_Evaluation_Report/links/5c96e229a6fdccd46036707e/Mallet-vs-GenSim-Topic-Modeling-Evaluation-Report.pdf.

[33] European Commission, *European Data Portal, data.eu*. Available at: https://data.europa.eu/elearning/en/module1/#/id/co-01.

[34] European Commission, *Research and innovation strategy 2020-2024*. Available at: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024_en.

[35] European Commission, *Facts and Figures for open research data*. Available at: https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/facts-and-figures-open-research-data_en

[36] Ross-Hellauer, Tony. *What is open peer review? A systematic review*, 2017. Available at: https://f1000research.com/articles/6-588/v2#ref-59.

[37] Tweepy Documentation. Available at: https://docs.tweepy.org/en/stable/index.html.

[38] FAIR data definition, University of Cape Town website. Available at: http://www.researchsupport.uct.ac.za/fair-data#:~:text=Accessible%3A%20once%20someone%20has%20found,(ethics%20always%20trump%20openness).

[39] Nationa Human Genome Research Institute, *What is the Human Genome Project?*. Available at: https://www.genome.gov/human-genome-project/What.

[40] Maarten Grootendorst, *Creating class-based TF-IDF matrices*. Available at: https://github.com/MaartenGr/cTFIDF.