Umanistica Digitale – ISSN: 2532-8816 – n. 13, 2022
G. Minello, D. Paci – *A Benchmark Corpus for Topic Modeling on the Origins of Modern Antisemitism*
DOI: http://doi.org/10.6092/issn.2532-8816/14767

A Benchmark Corpus for Topic Modeling on the Origins of Modern Antisemitism

[1]Giorgia Minello, [2]Deborah Paci

[1]Ca' Foscari University of Venice, Venice, Italy
[2]University of Bologna, Bologna, Italy
[1]giorgia.minello@unive.it
[2]deborah.paci4@unibo.it

Abstract

The pace of digitized collective knowledge accumulation has become increasingly rapid in the last few years. That means we have tremendous amounts of information content to be organized, searched, and understood that can be arranged only by employing automatic methods. In the case of textual data analysis, topic modeling, a machine learning method, is definitely the most famous framework to uncover latent topics from text documents. Adopting topic modeling approaches for studying textual sources is a well-established practice in many scientific and humanities studies fields, including the historical research scope. In this paper, we present a benchmark corpus for topic models, a dataset containing an annotated real-world collection of texts focused on the antisemitism theme in 19th century France. The benchmark corpus has been developed to address a specific machine learning task but it can also support the enhancement of other natural language processing-based studies, in particular, those concerning the historical sphere.

Negli ultimi anni il ritmo di accumulazione della conoscenza collettiva digitalizzata è divenuto sempre più rapido. Ciò significa che abbiamo enormi quantità di contenuto informativo da organizzare, ricercare e analizzare: una serie di compiti che possono essere svolti soltanto impiegando metodi automatici. Nel caso dell'analisi dei dati testuali, il topic modeling, un metodo di apprendimento automatico, è sicuramente la via più nota per cogliere gli argomenti latenti all'interno dei testi. L'adozione di approcci di topic modeling per lo studio delle fonti testuali è una

pratica consolidata in molti campi di studi scientifici e umanistici, incluso quello della ricerca storica. In questo articolo presentiamo un benchmark per il topic modeling, un dataset contenente una collezione di testi annotati incentrati sul tema dell'antisemitismo nella Francia del XIX secolo. Il benchmark è stato sviluppato per affrontare un compito specifico di apprendimento automatico, ma può anche consentire il miglioramento di altri studi basati sull'elaborazione del linguaggio naturale, in particolare, quelli riguardanti l'ambito storico.

Introduction

Digital collections have increasingly become the standard way of storing data in various fields. Specifically, in humanistic studies areas, such as history, these collections are enormous amounts of text files obtained through Optical Character Recognition (OCR), a tool for converting images of handwritten, typed, or printed text into machine-encoded text. On the one hand, this growing volume of textual data represents a tremendous gain for the researcher, being a source enrichment for studies; on the other hand, the quantity and the unstructured format of these files make them difficult to handle and explore, thus paradoxically becoming an issue. For this reason, text mining approaches have gained more and more popularity in very several fields over the last decades. These computational tools help us automatically organize, search, and understand these vast amounts of information. For instance, in the context of history study, such processing may involve identifying the names of prominent and not persons, the relationships between currents of thought, as well ass latent themes. In this regard, topic modeling is undeniably the most famous approach for finding subjects in a set of documents automatically. In a nutshell, topic modeling approaches, by means of machine learning techniques, attempt to find word and phrase patterns within a series of documents, and automatically cluster word groupings and related expressions that best represent the set.

Benchmarks are important tools for computer science researchers as they represent the way to objectively measure how well they are doing on a particular problem. Notably, in machine learning (ML) field, benchmarking means to evaluate and compare (ML) methods with respect to their capability to recognize patterns in datasets that have been applied as 'standards'.

In this work, we present a novel benchmark dataset to be employed to test topic modeling methods based on an actual collection of documents concerning the origins of modern antisemitism in 19th century France. This corpus, completely written in French, has been conceived to be both challenging and convenient to evaluate the performance of novel devised tools for textual data analysis.

The paper is structured as follows : in section 2 we provide references of both methodological and historical aspects of the corpus. In section 3 we present and discuss the corpus in details. In addition, we provide the needed background to understand the use of our dataset in an actual topic modeling application. Finally, in section 4 we discuss results and future work.

Related Work

This section provides a brief literature review about benchmarking, topic modeling in the humanistic studies scope, and an overview of the historical context of the corpus. For a thorough description of topic modeling characteristics we refer the reader to section.

Benchmarking

A benchmark dataset is the basis of fair comparison and validation of computational methods. In other words, a well-designed dataset allows researchers to test their approach and compare their algorithm performance with others'. This ground truth necessity is a crucial aspect for conducting fruitful research in many scientific areas, from computer vision [34] to fake news detection [48], not to mention the biology [11]. In the NLP scope, a benchmark dataset takes the form of a corpus, that is, a set of documents, a large collection of texts, complemented by annotations. These annotations clearly depend on the task the corpus has been devised for. For example, in the case of a dataset for the spelling correction [12] the annotations are the spelling errors while a corpus for sentiment analysis task present word-level sentiment annotations Parupalli et al. [33]. Annotation can be at a word level as well as a document level. For instance, in Wang et al. [49] the authors designed a corpus for related work generation where each document is labelled with a target paper, a ground truth related work, and the corresponding reference papers. Similarly, in [36], for the event classification task, the authors provide a large dataset based on a manually curated event repository consisting of short event descriptions.

Topic Models

NLP stands for Natural Language Processing, an interdisciplinary field of computer science and linguistics. It allows computers to understand and manipulate human language — speech or text. Topic modeling (TM) belongs to that scope. In particular, TM techniques permit finding hidden patterns in documents, called topics. The employment of these techniques is a well-established and recommended practice for exploring textual data in many scientific fields and humanistic studies, from biology to psychiatry [24]; [1]; [31]; [6]. TMs have the main advantage of "unveiling" the

leading themes in a collection of texts in an automatic and unsupervised way, that is, by letting the machine find the hidden structures itself and without any supervision by the reader. The objective of detecting the salient points in a collection of texts is not exclusive to topic model methods. In the last decades, other alternative approaches were proposed, such as by establishing frequently occurring phraseologies, defined as *n*-grams Fletcher (2007) or *conc*grams [7], as well as by characterizing the semantic fields by means of semantic annotation [37] or by comparing word frequency with that in a more general corpus, e.g. adopting the Keywords function in [42]. In practice, all these methods analyze what is most frequent and/or what is most distinctive about the specialized corpus. Even if these approaches contextualize the information the researcher deals with, they did not gain the same popularity as topic models, likely because considered reductive [31].

Recently, the body of literature discussing how to extract information from written text has increasingly grown. In this respect, TM has become more and more popular in history, social sciences and, in a broader sense, in digital humanities projects [51]; [41]. For instance, in [13], authors applied topic modeling to the ACL Anthology to analyze historical trends in the field of Computational Linguistics from 1978 to 2006.

Examples of how historical researchers have exploited TM techniques during the last decades are not lacking. Newman [32] can be considered the pioneer work about the employment of topic modeling in historical studies. Here authors explore and test the efficacy of TMs application to structure the content of the issues of the American colonial newspaper *Pennsylvania Gazette*, between 1728-1800. Later, further works confirmed the effectiveness of this procedure. In [53], authors experimented the use of topical models as proper tool to assist historical research by identifying potential issues of interest for historians. Few years later, another work to mention was presented in [16]. This was the first case of employment of topic modeling for a historical research monograph. Interestingly, in [38] we can find a clear illustration on how TMs method can be employed as an alternative way of reading a corpus in lieu of other traditional approaches that historians have used to explore very large collections of texts. Even if novel NLP-based techniques for text analysis have emerged in the last years, we can still find TM applications in recent works. For example in [25], where authors attempt to capture discourse dynamics on a large set of historical newspapers with the purpose of investigating methodological issues in diachronic data analysis for historical research. However, on this matter, it is worth noting that other NLP based approaches have been developed to study diachronic conceptual changes, such as those based on word embeddings [14]; [46]; [45].

Although, topic modeling is defined as "a type of statistical model for discovering the abstract 'topics' that occur in a collection of documents",[1] its methodological application embodies the 'distant reading' framework proposed by literary historian Franco Moretti, as in Brauer and Fridlund [5] mentioned. Specifically, in 2000, Moretti in Moretti [30] presented a new way of reading and analyzing texts, the so-called 'distance reading'. Broadly speaking, the 'distance reading' approach adopts a distant position of the reader concerning the collection of texts to analyze. This reading manner would allow detecting principal structures and informative parts underlying the set of documents more fruitfully than the 'close reading' mode, which, in contrast, employs a magnifying lens-based practice.

Historical Context

The debate around anti-Semitism identifies several factors that were at the origin of anti-Semitism in France in the 1880s-1890s ([17]). The defeat of Sedan in the franco-prussian war of 1870 with the consequent ceding of the provinces of Alsace and Lorraine stoked the fires of revanchist patriotism while weakening the republican nationalism of the Jacobite matrix. There prevailed a faith in the armed forces which was seen as the bedrock upon which the nation would redeem itself. In this context one can situate the success of the Boulangist movement, guided by General Boulanger, who could count on the support of the monarch and the church ([8]). He started to impose a nationalism based on the principles of order, authority, social hierarchy and tradition, which represented all the anti-parliamentarian tendencies and veered towards authoritarianism. This nationalism aimed to call into question the postulates on which the republic was founded, starting with consideration of the decadence of modern times and the fall of the Third Republic. Zeev Sternhell [43] has traced the origins of fascism to the anti-enlightenment tradition which rose to prominence at the end of the 19th century. It was this period that gave rise to a synthesis and fusion between two political currents up until then antinomic : one of the left, the other of the right. This meeting of a right-wing and populist nationalism, and a left-wing nationalism, which followed a revision of Marxism and rejected the enlightenment tradition, established the conditions for the birth of fascism. Anti-Semitism, as with racism, anti-liberalism, Boulangism, and social Darwinism, were political currents, which according to Sternell, created the conditions in which fascism could emerge. As Stephen Wilson [52] observed, the reasons for resentment towards Jews can be attributed, to the fears of the period ([20]). Attributing blame to Jews of being at the root of all evil furnished a justification for all the changes that assaulted modern society. Further according to Jean Paul Sartre : The experience did not bring about the notion of the Jew, on the contrary, it is

1 https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

this that clarifies the experience; if the Jew did not exist, anti-Semitism would have invented him ([47]). The Jew became symbol of the ills of modern society, phenomena such as capitalism and secularization in particular ([9]). The political implications of Jews' entrée into society stimulated modern anti-Semitism. Anti-Semitism transformed alongside modernization and its accompanying social change. In the nineteenth century, Frenchmen used "the Jew" to negotiate and deliberate the meanings of citizenship and the role of the citizen in the nation. Modern anti-Semitism has been linked with the political successes of the Jews of the Republic, specifically those Jews who integrated into the administration ([39]). The recognition of the rights of citizenship of Jews in 1791 was a result of the French revolution. The step forward in terms of rights had the effect of favouring greater integration of Jews into the social and political fabric; however, their inclusion in the French nation was considered a disuniting factor to the point at which it started to be spoken of as a "Jewish invasion". In virtue of the pronouncements of the National Assembly and of the advantages Jews gained as a result of the recognition of their citizenship, the idea prevailed that they were directly responsible for the French revolution. Every problem of a political, economical, or social nature was attributed to the Jews: from the collapse of the Catholic bank Union Générale (1882) through the Panama scandal (1892) and the economic crisis of the mid-1880s to the general decline of standards. It is possible to identify five types of anti-Semitism that can be combined one with the other in equal measure. 1) social and economic anti-Semitism : 2) conspiratorial anti-Semitism 3) racial anti-Semitism; 4) religious anti-Semitism; 5) moral anti-Semitism. Social and Economic anti-Semitism was manifest in the stigmatization of the concentration of capital. In the face of profound changes in the economic structure of the country, seen in the process of industrialization and modernization, anti-Semitism shifted progressively from the religious to the social terrain. Jews were considered agents of capitalism, bankers, and speculators, responsible for financial crashes such as that of the Catholic bank Union Générale. Small businessmen and shop owners, under pressure from the heightened competition by large distribution chains, saw the diffusion of department stores through slitted eyes, while the agricultural depression of the 1880s 1890s induced worry among landholders, not to mention the growing poverty of the urban proletariat and the anxiety about the military fragility of the nation. Expression of social unease, anti-Semitism arose as a reaction to the socio-economic reactions in French society, offering an ideology able to exorcize this change in the social structure. Conspiratorial anti-Semitism placed the accent on the presumed propensity of Jews for conspiracy. The presence of Jews in the spheres of economics and finance, as seen by the influence exerted by the Rothschilds on political power, was denounced as foreign and anti-national interference. A myth spread about the existence of a grand plan to surreptitiously appropriate the assets of France. The cosmopolitan and international nature of the Jews and what was viewed as their innate tendency towards betrayal were characteristics that rendered to them by antonomasia foreign agents. The Jews were considered conspirators and money

manipulators, as well as capable of leading or supporting political and social insurrection with the purpose of dominating the nation. According to the anti-Semites, the Jews were able to organize their conspiracies - known as Jewish masonic conspiracies - effectively because their actions were guided by a collective subject, belonging to a race naturally wicked and deceitful. Racial anti-Semitism was founded on the presupposition that the subdivision of races was a natural phenomenon, consequently, the characteristics of the enemy were considered immutable over time and hence an accident of history. The conviction was that the Jewish race had not changed its nature and would continue, without intervention, its conspiracy to take over the world. For anti-Semites, the Jewish race was propagating through mixed marriage pathological defects polluting the French race, the "true French" and the "good French". According to them Jews had specific somatic, biological, and psychical traits of a race that was ontologically inferior and dangerous. Religious anti-Semitism was found in the context of the policy of the secularization of the state. Jews were accused of being responsible for anti-clerical laws. They were thus considered a deicidal people, traitors, dedicated to ritual sacrifices and instruments of demons. Moral anti-Semitism refers to the prejudice about perverse sexual practices and their responsibility for the decline of standards. These forms of anti-Semitism inter-linked ([50]).

Methods and Results

Background

This section offers an intuitive explanation of the bottom mechanism of topic modeling, focusing mainly on the most famous topic model approach, namely the Latent Dirichlet Allocation (LDA).

In the next sections, we will use the following terminology and notation as adopted in [3].

- A *corpus* is a collection of D documents.
- A *vocabulary* is the list of all unique words in a corpus ("stop words" excluded).
- A *word* (or unigram) is a distinct item from a dictionary (or vocabulary), derived from the training *corpus*, indexed by $1,\ldots,W$. A specific instance or occurrence of a word within a document is called *token*.
- A *document* is a sequence of N words. In natural language processing, a document is usually represented by a bag-of-words (BoW) that is a word-document matrix.

- A *topic* is a set of words that occur frequently together. In [2], a topic is formally defined as a distribution over a fixed vocabulary.

Topic Modeling: Text mining means extracting meaningful insights from written resources through advanced analytical techniques. TMs fall into this very category. In fact, by TMs, we refer to that suite of algorithms used to discover themes that run through words of texts. TMs algorithms are of enormous importance in every field where data is represented by a huge amount of raw text. Indeed, these statistical methods enable us to outline and catalog electronic archives at a scale not feasible by human annotation. In addition, they do not depend upon any prior annotations of the documents to be analyzed. Technically speaking, topic modeling is a machine learning technique. It can be seen as the textual counterpart of the clustering method : while by clustering approach, we attempt to find (unknown) natural groups in items represented by numeric and/or categorical values, with topic model techniques, we try to identify groups of words related to certain unknown subjects. It is precisely for this reason, i.e. the unknown prior knowledge of topic content and presence, that we refer to topic modeling as an unsupervised classification method, like the clustering method is. The model learns that there are repeating patterns of co-occurring terms in a corpus but the learning is not preceded by any training stage (since there are no labeled data to learn). For example, a topic model may derive that the words "knives", "forks", "spoons" all fall under the umbrella of "Cutlery" but no label is associated with "Cutlery" as a known category.

We underline that, from the grouping text view point, topic modeling is not aiming to find similarities in documents, as text classification does, since topic modeling works at word level. In layman's terms, a topic model algorithm is able to detect topics in a text by counting words and grouping similar word patterns. The assumption behind these algorithms' functioning is that each document is composed of a mixture of topics. This, in turn, implies documents consist of a fixed number of topics. Then the goal is trying to find out the proportion of presence each topic has in a given document, where a topic is a mixture of words itself. In most of the TM algorithms, that translates into getting a term-topic matrix, which decomposes topics in terms of their word components, and the document-topic matrix, which describes documents in terms of their topics. The parallelism with the human process for text generation is quite evident. When we want to write a piece of text, first, we decide on the subjects and the importance (i.e. the weight) of each of them for our reasoning; then, we articulate each argument based on a pool of words related to that topic.

There exist several technical methodologies to carry out topic modeling. These can be divided into two categories: TMs performed by vector space models and TMs based on probability models. Some of these models include PachinkoAllocation Model (PAM) [23], Non-negative Matrix Factorization (NMF), TextRank [29], Parallel Latent Dirichlet Allocation (Plda) [49], to name but

a few. However, as TM is not the crux of this work, in this section, we describe a representative algorithm for each category, eventually focusing on Latent Dirichlet Allocation, which is the most common method. The basic vector space model is the Latent Semantic Analysis (LSA) [21]: here the assumption is that words and expressions that occur in similar pieces of text will have similar meanings or, rephrased, words with similar meanings appear frequently together (distributional hypothesis). The core of this approach is decomposing via singular value decomposition (SVD) the matrix representing information about documents and terms, in order to obtain a document-topic matrix and a term-topic matrix. In fact, LSA maps terms and documents into a latent semantic space via SVD. On the other hand, probabilistic topic models are those statistical algorithms aiming at discovering the latent semantic structures of the corpus by relying on a document generation process. The idea behind the document generation process, as mentioned above, comes from the human written articles. At heart, probability topic models do simulate the behavior of the human article generating process. For instance, Probabilistic Latent Semantic Analysis [15] (pLSA), to face the same matter of LSA, uses such strategy, that is, a probabilistic approach. Here the goal is thus to find a probabilistic model (instead of the SVD) able to generate data contained in the document-term matrix. Without going into mathematical details, we may affirm that pLSA was certainly a turning point for the probabilistic text modeling sector but did not gain the same success as LDA because it was incomplete, as lacking a probabilistic structure at the level of documents.

LDA: Unlike pLSA, Latent Dirichlet Allocation (LDA) ([3]) is the first fully probabilistic model for text clustering, the Bayesian version of pLSA. LDA is a probabilistic generative model, i.e. a machine learning technique that generates an output by considering the prior distribution of some objects. Specifically, LDA draws on Dirichlet priors for the document-topic and word-topic distributions. In plain language, a Dirichlet process is a distribution over distributions. In this section, we will try to understand the model's mechanism without diving into the math. LDA is built atop the premise that each document can be described by a probabilistic distribution of topics. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups. Moreover, LDA also assigns a distribution of words to each topic. In other words, the presumption is that a document exhibits multiple topics, usually not many, and each document is generated by a process. Similarly, a topic is generated by a process over a fixed vocabulary. First, each topic is generated, then the documents are produced. This is exactly the generative model for a collection of documents that LDA tries to backtrack from texts, by specifying only the number of topics in advance. LDA attempts to estimate both the mixture of words associated with each topic and the mixture of topics describing each document. It should be noted that LDA, as well as pLSA and LSA, is based on the "bag-of-words" assumption, i.e. within a document the order of words has no importance, meaning words are exchangeable. This exchangeability assumption, which also

holds for documents, underpins the whole LDA mechanism. Indeed, following De Finetti's theorem that states that any set of exchangeable random variables has a representation as a mixture distribution, if exchangeability for documents and words within documents is assumed, a mixture model for both is needed, which is exactly the LDA foundation. Practically speaking, LDA takes as input a document-word matrix (by counting the word presence in each document) and gives as output two matrices: 1) the document-topic matrix containing for each document the topic distribution, and 2) the topic-word matrix providing for each topic the word distribution, as shown in Figure 1. There are a number of existing implementations of this algorithm. From the computational view point, in this work, we used the LDA variant based on Gibbs sampling (a form of Markov chain Monte Carlo) introduced by Tom Griffiths to prevent correlations between samples during the iteration [44].

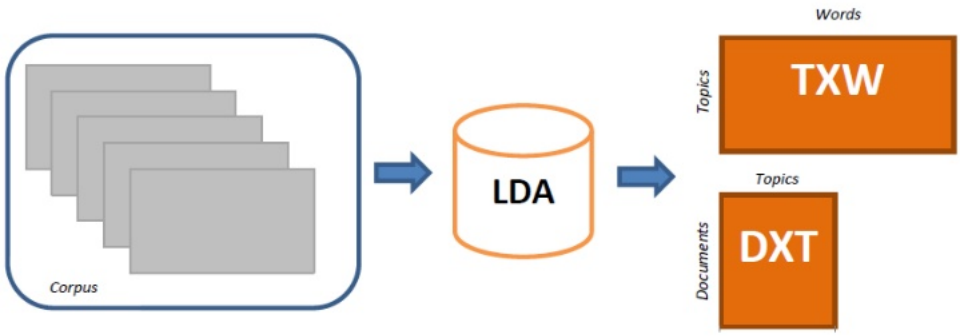

Figure 1 LDA summary schema.

Benchmark Corpus Criteria

In many sciences, benchmark datasets are essential to rigorously compare the performance of different methods, determine the strengths of each technique, and provide recommendations regarding suitable choices of methods for analysis. Yet, there are no official guidelines nor protocols shared by the scientific community concerning the creation of a benchmark dataset, to the best of our knowledge. Nevertheless, unofficially certain essential characteristics exist and are widely acknowledged by the scholarly community. For this reason, to devise a proper and fair corpus, we gathered a set of requirements we know by experience to be critical for a benchmark dataset and translated them into the following criteria:

— *Practicality* : it must deal with the machine learning task taken into consideration

— *Difficulty* : it must have a sufficient level of difficulty in order not to result trivial

— *Reality* : it must be an actual representative of real-world data science problems to be solved with the given tool

— *Diversity* : it must have enough features to be interesting

— *Ground truth* : it must have clear labels to use as ground-truth

— *Accessibility* : it must be open and well documented

— *Quality* : it has to be clean but not too clean, otherwise, it would result unrealistic.

— *Size* : it should not be too large.

The Corpus

This section provides a detailed description of the corpus from a historical and content viewpoint. Essentially, the following summary is the result of a close reading of the entire corpus carried out by a human being, next used to derive the dataset annotations.

We have put together a corpus of twenty-three works of authors linked in various respects to Édouard Drumont [18] freely accessible in the *Gallica* the online library of the *Bibliothèque Nationale de France* and *Internet Archive: Digital Library of Free & Borrowable Books.*[2] Journalist of Catholic orientation, he became intellectual point of reference to French anti-Semites in the 1880s-1890s. Born in 1844 in Paris to a working class family, his father was employed at the Hôtel de Ville, his mother was related to Alexandre Buchon, historian of the crusades and the medieval period, both fervent republicans. Drumont started a journalistic career in 1861 under the patronage of Alfred d'Aunay, putting his name to articles in *Moniteur du Batiment* and *Presse théatrale et musicale*. In 1864 he did work for *Contemporain*. Shortly before the war of 1870 he contributed to the journal *Liberté*, a collaboration which lasted until 1885. Despite the support of Louis Veuillot, manager of the Catholic newspaper *L'univers*, he remained a little-known journalist. As a writer he met with little success, as evidenced by two works from 1878 and 1879 respectively : *Mon Vieux Paris e Les Fetes nationales à Paris.* Our corpus includes a selection of works on the question of Jews in France during the second half of the 1800s : half of these texts, marked as "friends of Drumont", were written by authors of antisemitic orientation close to Drumont, the other half, labelled as "enemies of Drumont", includes volumes of opponents of anti-Semitism. To this work a reference text to frame Drumontian thinking has been added.

2 https://gallica.bnf.fr/ and https://archive.org/ respectively

In 1886 a hefty work in two volumes entitled *La France juive. Essai d'histoire contemporaine* (148), a true and proper manifesto of anti-Semitism was released to the press. Although the subtitle of the work suggests an essay of contemporary history, on reading it is as if one is before an enormous cauldron of commonplace assumptions on Jews which includes Catholic, social, racial, economic, and conspiratorial anti-Semitism. The author of the work, which sold 65000 copies in the year of its publication, was Édouard Drumont. The success of his work - between 1886 and 1912 it was reprinted 200 times - depended on the waves it made in the intellectual milieu of the era and its impact on the popular masses attracted by the synthesis of anti-Semitism of the right, of a church worried about laicization, and anti-Semitism of the left, anti-capitalist and laical. In 1889 Drumont founded along with Jacques de Biez, Albert Millot and the Marquis de Morès la Ligue Antisémitique de France to which they gave the motto "France to the French". In 1897 at the initiative of Jules Guérin he set up la Ligue Antisémitique Française with the intent of reviving the moribund Ligue Antisémitique de France [22]. The members of Ligue came from the popular classes and distinguished themselves for their violent methods. Drumont was honorary president. Within our corpus we have included the collective volume *Rothschild, Ravachol & Cie* (1892) by the organization "Morès et sesamis" founded in 1891 by the Marquis de Morès which describes the influence exercised by Jews on French society through an enquiry into the strategy put in place to take control of the managing class and the social and economic structures of the country. The authors underline the role played by freemasonry, by the war conducted against the catholic world and unions, to the alliance with the English up to the condemnation of the Franco-Russian alliance. As a result "Morès et ses amis" examines the role of the Bank of France revealing how it was entrusted to Rothschild, considered to have preferential access to state funds. The agents "knowing and unknowing" of Rothschild are "the freemasons, the press, the markets, the red spectrum" (p. 6). Therefore it was necessary to intervene politically so that the Jews, classed as foreigners, lost a privilege that instead belonged to the French people and which was used to sustain agriculture, commerce, and industry. From an analysis of the network surrounding the collaborators we can deduce that Drumont assiduously cultivated relations with people linked to the, such as the literary salon run by Alphonse Daudet, and press, in particular his immediate colleagues at the journal *La Libre Parole*, founded by him in 1892, that was first to report the arrest of Dreyfus on 29th October 1894. One of the texts examined is by Léon Daudet, son of Alphonse : *Souvenirs des milieux littéraires, politiques et médicau*x (1920). In these memoirs Daudet outlines how it was initially the literati and journalists who made known the writings of Drumont. People who would then go on to espouse the anti-Dreyfusard cause, were already actively engaged in the diffusion of anti-Semitic ideas well before the Dreyfus case exploded. Somewhat different are the volumes of the corpus written in the journal "La Libre Parole". We can mention Caroline Rémy de Guebhard known as Séverine ([4]), socialist journalist and feminist who wrote in *La Libre Parole* in

the years 1894-1896, author of *Vers la lumière...impressions vécues: affaire Dreyfus* (1900), Jean de Ligneau, pseudonym of François Bournand, special secretary to Drumont and coordinator of the news section of "La Libre Parole", author of *Juifs et antisémites en Europe* (1891), the anti-Semitic adventurer and editor of "La Libre Parole", the Marquis de Morès ([19]) or again Isaac Blümchen, psuedonym of Urbain Gohier, a rather complex, and in many ways contradictory figure ([28]). Dreyfusard, friend of Émile Zola, anti-Semite, socialist e anti-militarist Gohier collaborated with "La Libre Parole" in 1908. Victor Célestin Méric, pseudonym of Henri Coudon, defended him from those who attacked him for his anti-Semitic position in an article published in "Les Hommes du Jour", 6 March 1909: "Without doubt, whether at Libre Parole or at Intransigeant, on the right or the left, Gohier continued tirelessly the same work and found himself, forever poor, forever indifferent to material gain, on the same side of the barricade".[3] His collaboration with Drumont made Gohier the subject of hostility of many colleagues, among them the anarchist Jean Grave, who accused him of being an armchair revolutionary and who in his novel *Les Malfaiteurs* (1903) found it amusing to portray him as Roguier to reveal his bad faith.[4] A good part of the text of the corpus is about l'Affaire Dreyfus. Of particular note in the section of the corpus on "enemies of Drumont" are two works by Émile Zola, credited with the celebrated editorial *J'accuse* published on13 January 1898 in the journal *L'Aurore* in which, addressing the then President of the Republic Félix Faure, the writer denounces the irregularities that occurred in the trial to the detriment of Dreyfus. In the corpus we have *Lettre à la jeunesse: l'affaire Dreyfus* (1897) and *Lettre à la France: l'affaire Dreyfus* (1898). The first is an open letter to French youth invited by Zola to rebel against what he viewed as a social injustice : the trial and the conviction of Dreyfus, was more correctly a "judicial error" (p.6). Zola, turning to his interlocutors exhorts them not to squander what has been achieved to date by the preceding generation and instead carry forward the social and political gains achieved. In *Lettre à la France : l'affaire Dreyfus* (1898) Zola calls upon the entire nation, France, to pay attention to the danger contained in anti-Semitism. He emphasized the role of newspapers, in particular *L'Écho de Paris* and *Le Petit Journal*, in having influenced public opinion by publishing false news and generating fear and intolerance. Zola denounced the silence and more in general the lack of interest shown by parliament with respect to the intrigue. The text is constructed around the crucial importance of truth, which he stressed several times : "I will dare to say everything, because I have always had one passion in my life, the truth" (p.4). L'affaire Dreyfus is furthermore object of discussion by authors close to Drumont, such as Séverine, who in *Vers la lumière... impressions vécues: affaire Dreyfus* proposed as a court reporter, to report the facts, to throw light on the protagonists and on their conduct seeking to depict the climate created in that period characterized

3 https://gallica.bnf.fr/ark:/12148/bpt6k442295x/f5.item

4 http://militants-anarchistes.info/spip.php?article2303

by disorder both in and outside the halls of the tribunal. Séverine accused certain sections of the press of having diffused false and distorted information. In the course of the narration it is possible to discern an evolution of the position of the author on the guilt of Dreyfus: initially firmly convinced of his guilt, she gradually evinces a certain perplexity about the trial, the irregularities and errors committed at the judicial level and concluded with an affirmation of the innocence of Dreyfus. Even so Séverine did not nurture any special empathy for Dreyfus and expressed anti-Semitic sentiments, in particular when she referred to the wife of Dreyfus ("She is not of my religion, and not of my race [...] And if I don't like the colour of their skin anymore than I don't like, in general, the Jewish soul, this was not a reason to approve of their torture!" (p. 68). In the end she fell into the camp of those who fought for the liberation of Dreyfus in the name of the struggle for "Truth") (p. 461).

In the corpus we have included a series of works that from two points of view diametrically opposed interpret the role of the Jew in society. Among the authors that dominate the corpus is Bernard Lazare, poet-anarchist, a convinced Dreyfusard, who in the volume *L'antisémitisme: son histoire et ses causes* (1894) set out to reconstruct the origins and the deeper causes and long duration of anti-Semitism. Lazare proposed a historical-sociological study that was impartial and that made a lie of all the voices that depicted him as anti-Semitic one moment, philo-Semitic the next. He analyzed the social, economic, and political context within which the transformations and changes in anti-Semitism had matured, and at the same time enquired into the role played by Jews in the intellectual, economic, social, and political fields. According to Lazare, to feed the anti-Semitic sentiment of many were elements innate to the Israelites, that is, congenital traits, such as asociality, resistance to integration because of the unity of morality and politics in Jewish religion. Beside these endogenous causes, the author identifies exogenous causes strongly linked to the historical epoch and particular context. According to Lazare, modern anti-Semitism originates from hate for the stranger and as a consequence, for Jews, who were viewed as a foreign minority. However, in his opinion, anti-Semitism was destined to disappear thanks to the progressive assimilation of Jews in contemporary society and to the development of the ideals of socialist internationalism (p. 408-409). On the other hand we have included in the corpus authors who are strongly anti-Semitic, such as forementioned Gohier, who in *À nous la France* (1913) claimed, writing in the first person singular and plural, and making extensive use of irony, all the weaknesses revealed how much France and the French were under the yoke of Jews, whose ultimate objective was the conquest and control of the world. Therefore he eulogizes ironically the conduct of several French ministries and politicians that served the Jews, who in virtue of French naturalization, had succeeded in climbing the social ladder and obtaining prominent positions. Of the same tenor is the volume *Droit de la race supérieure* (1914) in which Gohier describes the right of the superior race, the Jew, to dominate

that which is inferior, the French, as a law of nature. To support this thesis he mentions the dominance of Jews in education, politics, the judiciary, the press, and freemasonry. Gohier dwells on the situation of the colonies, revealing how the Arabs, despite having served France, had never obtained, in contrast to Jews, citizenship. Finally a conspicuous part of the corpus deals with the question of anti-Semitism in Algeria. Anti-Semitism in Algeria was also a relevant aspect of Drumont's personal network. In 1892 la Ligue anti-juive d'Algers was founded at the initiative of Émile Morinaud and Max Régis. Morinaud was socialist republican deputy and member of the Parti Radicale Antijuif, a leading light in the Algerian anti-Semitic movement. Régis was a naturalized Frenchman and distinguished himself in his university days for his fervour as a political agitator. He refounded in Algeria a "Comité Central Républicaind'UnionAntijuive" in 1898 drawing inspiration from Ligue Antijuive established in 1892 by Fernand Grégoire ([35]). In January 1898, a crowd led by Régis torched central Algiers for five days, killing two people and destroying the synagogue and several homes ([27]). Max Régis had a central role in the candidacy and election of Drumont in Algeria. It was at the suggestion of Régis that Drumont put himself forward for election in Algeria in May 1898 and was elected deputy as part of the anti-Semite group ([39]).

One of the texts of the corpus is L*es mémoires du prisonnier Max Régis* (1899) consisting of the memoire written in prison by Max Régis collected by his friend Louis Gardais and published in "Mustapha imprimerie spéciale de *L'Antijuif*". In this account Régis talks of his incarceration for four months and rails against the French government, the local administration, and in particular the prefect of Algiers Charles Lutaud. The government in Paris was accused of showing no consideration for the people of Algeria and at the same time of being manipulated by Jews. Régis did not neglect to outline how popular he was in Algeria reproducing extracts of the letters that he received from supporters, friends and political figures such as Drumont but also from his mother who urged him to terminate his "struggle" as "the Jews will always have the power to impose the governor they wish, and it won't be our revolt, however beautiful or legitimate it might be, that will impede this"(p. 6). From another point of view, opposed to that of Drumont, is the volume *L'antisémitisme algérien* (1899) that reproduces a speech from the chamber of deputies on the 19 and 24 May 1899 by Gustave Rouanet ([40]), socialist deputy for la Seine, addressed to Algerian anti-Semitic deputies. The intention of Rouanet was that of demonstrating the inconsistency of the anti-Semitic discourse, confuting the arguments, in particular those that outlined the financial and economic wealth of Jews. In his view social criminality did not originate from the supposed wealth of Jews, but from the harm generated by capitalism. Rouanet outlined how before the Crémieux decree of 1870 bestowing French citizenship on Algerian Jews, they had done their military duty fighting beside the French. The naturalization of the Jews in Algeria brought about a hardening of

anti-Semitism, which until then was not widespread, and helped aggravate an already tense situation due to the presence of Italians and Maltese who cultivated an "ancient hatred of Jews" (p.33) and that with their traditions, habits, and mentality had undermined the French spirit distancing Algeria from France.

We remind the reader that the corpus can be easily built by downloading from Gallica and Internet Archive: Digital Library of Free & Borrowable Books the document raw files, in txt format. Alternative, all files are also available at the following Link.

Table 3 contains some information about the size of each documents. In particular, we can find the number of characters and words in the original texts, the number of words obtained after tokenization and stop word removal, the number of unique tokens in each literary work and the percentage of each documents based on the total number of characters in the original versions. This last column of the table deserves some remarks. *Souvenirs des milieux littéraires, politiques artistiques et médicaux de 1880 à 1905* occupies 21.23% of our corpus because it deals with a central theme, namely the relationships matured within the intellectual salons in the period under discussion. These memoirs, by Léon Daudet, son of Alphonse (main entertainer of the Parisian literary circles of the time as well as a Drumont's friend) allow us to reconstruct the overall picture of the network of acquaintances within the anti-Semitic milieu as well as the dynamics that involved the opposing sides, that of the dreyfusards and the antidreyfusards. Still substantial (15.05%) is *La France juive*, Drumont two-volume work that represents the sum of anti-Semitic arguments, in which the main topic of our study dominates: anti-Semitism. Although they do not have a similarly large percentage (respectively 1.65% and 0.8%), the two texts by UrbainGohier - *A nous la France* and *Le droit de la race supérieure* - focus on the theme of the influence of Jews in French society according to a pattern of interpretation of conspiracy theorists. The topic relating to the Affaire Dreyfus is present in some works, some of which are relatively short since they are letters (*Lettre à la jeunesse* and *Lettre à la France* by Zola) in which, nevertheless, the subject is examined in-depth and where we can identify homogeneous sub-themes, such as the concept of law, justice, truth. Other documents, written by non-anti-Semitic authors, with a percentage between 0.42% (*Antisémitisme et révolution by Lazare*) and 9.23% (L'antisémitisme: son histoire et ses causes by Lazare), are uniform in terms of topics: the Dreyfus Affaire and the origins of anti-Semitism. Those texts addressing the issue of anti-Semitism in Algeria and the status of Jews in that country are also homogeneous (they represent 26% of our corpus). Four of them are argumentative books, with a percentage between 4.78% and 7.46%, one book is a biographical account of Max Régis (0.43%), a leading exponent of anti-Semitism in Algeria, the other one is a condemnation speech of the anti-Semitic movement in Algeria pronounced by Gustave Rouanet.

Based on human interpretation (close reading), it was possible to identify the topics, divided into macro-topics and sub-topics, present in the collection of documents. It should be stressed that these are the themes detected by a human expert blindly, that is, without thinking in terms of ground truth for the machine. The addition of the related keywords was performed as a subsequent stage. It was derived from a set of words, a dictionary. This set of words was retrieved by selecting the most relevant terms by topic obtained through the implementation of the LDA method over our corpus, as described in 3.4 but fixing the number of topics parameter to $k = 15$, to have a larger basin of terms. Specifically, in Table 1 the macro topics and related subtopics are reported along with their keywords while Table 2 is an outline of the documents in the corpus, annotated with topic information : title (original and translated), author(s) name, document typology, position of the author(s) with respect to Drumont where F stands for Drumont's friend while E indicates an opponent, number of topics present in the work and content of these topics. Since the corpus is completely written in French we provided keywords both in English and French, to help the user in understanding the content.

It is of interest to compare words we expect in documents based on Table 1 and Table 2, and most common words contained in them, by frequency. To show word frequency, we used a well-known graphical representation, i.e. word clouds, as depicted in Figure 2. Here, we can immediately notice as the word *juif* is the most frequent, along with *france*, *homme* and *peuple*. Even if they are explanatory for an initial investigation, in terms of topic modeling viewpoint they are too general. Also from the content viewpoint they are vague. For instance, in *Souvenirs des milieux litteraires, politiques, artistiques et medicaux de 1880 à 1908* and *Les grandes juives de l'histoire* the wordclouds highlight words such as "man", "great", "father", "Jewish woman" which are not significant at all for the actual topic, i.e. the intellectual circles in France at the time. Based on the close reading we expected to find words such as "Daudet", "Hugo", "literature", and "Paris". Another example of fuzzy terms concerns the two wordclouds of Urbain Gohier's works, *A nous la France* and *Droit de la race supérieure*. In this case there should be words related to the theme of Jewish conspiracy and Jewish influence in French society, such as "Rothshild", "money", and "bank", instead we find "army", "France", and "Jew", which unarguably common terms. On the other hand, *Les mémoires du prisonnier Max Régis* wordcloud shows adequate results: in particular, it emphasizes the names of Charles Lutaud, prefect of Algiers in 1898, and Julien Édouard Laferrière, governor general of Algeria from 1898 to 1900 and responsible for revoking Régis' mandate as mayor of Algiers.

Results

In this section we will discuss how the created corpus meets the benchmark criteria we set and we will see a basic utilization of the corpus as benchmark dataset for topic modeling. It should be

pointed out that the corpus may be employed in several other NLP research cases since each document is annotated and the most prominent words are contextually clustered.

How does the Corpus Meet Benchmarking Criteria? We attempted to fulfill all set requirements. In the following, we report how we tackled criteria.

1. *Practicality* → The dataset deals with topic modeling task (it is a collection of documents of which we are interested in finding out topics).
2. *Difficulty* → The number of topics and subtopics makes the task quite challenging.
3. *Reality* → The corpus is an actual example of a collection of documents to be studied by historians.
4. *Diversity* → There are different types of documents (books, letters, etc.) and themes. Novels have not been considered for the construction of the corpus intentionally, as they are not suitable for the topic modeling tasks.
5. *Ground truth* → There are labels both at document and word level.
6. *Accessibility* → The dataset is freely accessible.
7. *Quality* → All documents are from official online repositories with policies about OCR level.
8. *Size*→ The number of documents is sizeable but at the same time allows to perform LDA analysis quickly (a few minutes via Google Colab Intel(R) Xeon(R) CPU 2.20GHz, RAM 12 GN Disk 108 GB).

LDA Experiment We carried out the LDA experiment in Python, by adopting standard NLP libraries, such as gensim and nltk. To preprocess the text we tokenized the text documents (by splitting the entire corpus into smaller units, i.e. individual terms) and removed stop words. Stop words are the set of commonly used words in any language. The list of French stop words is provided at the following Link. We did not perform any stemming process. Tokens have not been filtered out by their frequency (we kept tokens which are contained in i) at least 1 document ii) in no more than 1 expressed as fraction of total corpus size, that is, the whole corpus). We performed the LDA Mallet Model, by Andrew McCallum, McCallum ([26]) in Python through the gensim wrapper gensim.models.wrappers.LdaMallet. We fixed the number of topics to be 5 (as many as the macro topics) and the number of training iterations equal to 200 (default value is 50).

In Table 4 results of LDA application with $k = 5$ are reported. Each column represents a topic and contains the words the machine detected. Words are in descending order, that is, the most important words are at the beginning of the list. Then, to exploit our annotations, we employed the

Jaccard similarity (Js) to assess the quality of the predicted topics. Js is a measure commonly adopted to compute the similarity between two objects, such as two text documents. Jaccard similarity can be used to find the similarity two sets and it is defined as

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}$$

where A and B are the two sets, ∩ and ∪ denote intersection and union, respectively. The Jaccard similarity index (sometimes called the Jaccard similarity coefficient), ranges from 0 to 1. The higher the percentage, the more similar the two populations. In other words, we count the number of members which are shared between both sets and divide by the total number of members in both sets. In in the bottom row of Table 4 we reported the Js between each topic found via LDA and the actual topics (in bold the highest value). Without studying the word lists but only these values, we can affirm that by using $k = 5$ as parameter we did not obtain a perfect clustering. Indeed, if LDA had identified all the themes correctly, we would have just one sharp Js match for each predicted topic. Nevertheless, we can keep appraising LDA results by studying these predicted labels. For instance, by comparing Table 5 and Table 2, we can say that 16 out of 23 topic assignments are correct.

Notably, the topic about the anti-Semitism (topic #1) appears to be the most important in both works written by anti-Semitic authors, including *La France juive* (0.3338), and by strong opponents of anti-Semitism [*Juifs et antisémites en Europe* (0. 4116), *Le nationalisme juif* (0.5443), *L'antisémitisme son histoire et ses causes* (0.707), *L'antisémitisme algérien* (0.3374), *Antisémitisme et révolution* (0.4412)]. The close reading confirms the results of the distant reading as concerns works focusing on anti-Semitism and its history. The close reading also confirms the assignment of topic #1 and #2 (about anti-Semitism and Algeria respectively) to the volume *L'antisémitisme algérien* by Gustave Rouanet.

Topic #2 regards the French conquest of Algeria. Indeed, we can find it in both *L'Algérie de 1830 a 1840* (0.7878) and *La conquête d'Alger* (0.7821). Here again, we can observe a correspondence between the close reading and the distant reading.

Topic #3, related to the discussion about the Dreyfus Affair, is correctly prevalent in those works dealing with the French army officer story. Both distant and close reading give us this result without equivocation: *Vers la lumière* (0.6027), *Lettre à la jeunesse* (0.7289), *Lettre à la France* (0.8321), *L'Affaire Dreyfus* (0.7409), *La verité sur l'affaire Dreyfus* (0.7432), *Comment on condamne un innocent* (0.7024). It is surprising, but only apparently, that *Les memoires du prisonnier Max Régis* is among works where topic #3 prevails, followed at a short distance by topic #4 concerning

Algeria (0.2238). The close reading suggests that the text focuses on Algeria. However, although the Dreyfus affair is not mentioned in these memoirs, we can observe a vocabulary in which, coherently with the themes dealt with by topic #3, criticism of the government and of the alleged injustices suffered by the prisoner Régis emerges. In these memoirs, for example, there is a letter signed by Régis' mother in which she underlines her son's combative attitudes against the injustices he had to undergo by the French government. It means that Régis's condition can be compared, from the lexical terms viewpoint, to the Dreyfus one.

Topic #4, the one without a clear label, unexpectedly emerges in those documents concerning the presence of Jews in Algeria, such as *Les juifs en Algérie* (0.6093), *L'Algérie juive* (0.753). It is also the main topic for the work about the symbol par excellence of Jewish conspiracy, namely the Rothschild family, *Rothschild and Ravachol et Cie* (0.5148) and for *Droit de la race supérieure* (0.3569). While for those texts definitively about Algeria the assignment is clearly misplaced, actually for Gohier work it is not, from a historical expert perspective, as the volume treats all the anti-Semites arguments. Instead, *A nous la France* does not have a prominent topic as expected since it does discuss a larger set of issues. Finally, we find the correspondence between distant reading and close reading by looking at topic #5 assignments. This topic focuses on the relations between intellectuals and Jews. It is the most prevalent in *Souvenirs des milieux littéraires, politiques artistiques et médicaux de 1880 à 1905* and in *Les grandes juives*. The former depicts the intellectual milieu of the second half of the 19th and the beginning of the 20th century; the latter provides the new generations of French Jewish women some fundamental lessons drawn from the stories of the great Jewish women, highlighting their virtues and intellectual gifts. In both cases we can affirm that there was a correct assignment.

Conclusions

In this work, we presented a benchmark corpus to evaluate and compare NLP methods. The corpus has been devised to satisfy criteria widely yet unofficially recognized as necessary for a benchmark dataset. The corpus has a historical slant, being the origins of modern antisemitism as the primary subject matter, which makes it original in its kind. We provided explicit annotations and labels at the word and document levels and tested the convenience over a basic LDA application. Outcomes are promising and encourage future work in this direction. For instance, as concerns the applicability of labels for testing topic modeling approaches, we plan 1) to quantify numerically the importance of words within topics and topics within documents, so that to make labels more machine-friendly 2) to expand the basin of words to use for the topic description. Finally, the

employment of this corpus for a fair comparison between close reading (human beings) and distant reading (machine) is already on the schedule.

Funding

This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement No 732942 and was written in the context of the ODYCCEUS H2020 Research Project (https://www.odycceus.eu/).

Author Contributions

This article is the result of a collaboration between the two authors. They both thought and discussed together all the sections. Specifically, Deborah Paci wrote "Historical context" and "The Corpus". They both contributed in producing the section "number annotations" and the "Results". The remaining parts were written by Giorgia Minello.

Figure 2 Wordclouds.

#	*Macro Topic*	*MT keywords*	*Sub-Topics*	*ST keywords*
1	Anti-Semitism	juif, nation, peuple, drumont, aryen, léon,gambetta, république, terre (jew, nation, people, drumont, aryan, léon, gambetta, republic, land)	Characteristics of Jews	rabbin, race, sang, nez (rabbin, race, sang, nose)
			Religion	rabbin, antijudaïsme, judaïsme, église, talmud,catholiques, chrétien, saint, dieu, foi, israélites, jésus (rabbi, anti-Judaism, judaism,

				church, talmud,catholics, christian, holy, god, faith, israelites, jesus)
			Jewish influence in society	argent, banque, capital, commerce, rothschild, pereire (money, bank, capital, trade, rothschild, pereire)
2	Dreyfus Affair	dreyfus, droit, juge, justice, vérité, bordereau, accusation, procès, capitaine, enquête, homme, libre, innocent, douleur, loi, ministre, peuple, raison, silence (dreyfus, law, judge, justice, truth, accusation, trial, captain, investigation, man, free, innocent, pain, law, minister, people, reason, silence)	Protagonists of the Dreyfus Affair	esterhazy, georges, picquart, emile, zola, joseph reinach, léon, gambetta
			Role of Press	journal, livre, dossier, article, documents, lettre (newspaper, book, file, article, documents, letter)
3	History of France	allemagne, autriche, baron, civilisation,empire, espagne, état, europe, france, léon,gambetta, gauche, gouvernement, guerre, histoire, ministre,monde, napoléon, nation, paris, politique, pouvoir,président, puissance, restauration, russie, république, révolution, siècle, société, époque germany, austria, baron, civilization, empire, spain, state, europe, france, léon, gambetta, left, government, war, history, minister, world, napoleon, nation, Paris, politics, power, president, power, restoration, russia, republic, revolution, century, society, era		
4	Intellectual Circles	daudet, hugo, littérature, ami, amour, années, auteur, caractère, cœur, esprit, fantômes, léon, gambetta, homme, jeunes, lemaître, lettre, lockroy, madame, loynes, maitre, milieu, musique, mots, paris, salons, souvenir, théâtre, ville, vivants, œuvre, milieu		

		(Alphonse Daudet, Victor Hugo, literature, friend, love, years, author, character, heart, mind, ghosts, Léon Gambetta, man, young people, Georges Lemaître, letter, Édouard Lockroy, Madame de Loynes, master, environment, music, words, Paris, salons, memory, theatre, city, living, work, environment)		
5	Algeria	naturalisation, décret, crémieux, rouanet, afrique, morinaud, oran, lutaud, laferrière (naturalisation, Crémieux decree, Gustave Rouanet, Africa, Émile Morinaud, Oran)	Colonization	colonie, indigènes, empire, arabes, civilisation (colony, natives, empire, Arabs, civilisation)
			French conquest of Algeria	alger, armée, bey, kabyles, capitaine, gouverneur, général, lieutenant, clauzel, colonel, commandant, abd, kader, kanoui, officier, soldats, tribu, troupes, artillerie, bataillon (Algiers, army, bey, kabyles, captain, governor, general, lieutenant, Bertrand Clauzel, colonel, commander, Abd el-Kader, Simon Kanoui, officer, soldiers, tribe, troops, artillery, battalion)

Table 1: Actual Topics and keywords

Year	Original Title	Title	Author(s)	Type	Pos	# topics	Content
1879	*La conquête d'Alger*	The conquest of Alger	Rousset	Book	F	2	French conquest of Algeria colonization history of France
1882	*Les grandes juives*	The great Jewish women	Weill	Book	E	2	intellectual circles characteristics of Jews
1886	*La France juive*	Jewish France	Drumont	Book	F	3	Jewish influence in society characteristics of Jews religion history of France
1887	*L'Algérie de 1830 à 1840*	Algeria from 1830 to 1840	Rousset	Book	F	2	French conquest of Algeria colonization History of France
1887	*L'Algérie Juive*	The Jewish Algeria	Meynié	Book	F	2	French conquest of Algeria naturalization
1888	*Les juifs en Algérie*	The Jews in Algeria	Meynié	Book	F	2	French conquest of Algeria naturalization
1891	*Juifs et antisémites en Europe*	Jews and anti-Semites in Europe	De Ligneau	Book	F	3	Jewish influence in society characteristics of Jews religion
1892	*Rothschild, Ravachol &Cie*	Rothschild, Ravachol&Cie	Morès et ses amis	Book	F	1	Bank of France Jewish influence in society
1894	*L'antisémitisme son histoire et ses causes*	Antisemitism: its history and causes	Lazare	Book	F	2	Jewish influence in society religion
1896	*Contre l'antisemitisme: histoire dune polémique*	Against anti-Semitism: a history of controversy	Lazare	Book	F	2	Jewish influence in society religion
1897	*Lettre a la jeunesse:*	Letter to Youth : The Dreyfus	Zola	Letter	E	1	Dreyfus affair

	l'affaire Dreyfus	Affair					
1897	*La vérite sur l'affaire Dreyfus*	The truth about the Dreyfus affair	Lazare	Book	E	2	protagonists of the Dreyfus affair role of press
1898	*Lettre a la France*	Letter to France	Zola	Letter	E	1	role of press
1898	*Le nationalisme juif*	Jewish nationalism	Lazare	Book	E	2	Jewish influence in society religion
1898	*Comment on condamne un innocent:* *l'acte d'accusation contre le capitaine Dreyfus*	How an innocent man is condemned: the indictment against Captain Dreyfus	Lazare	Book	E	2	protagonists of the Dreyfus affair role of press
1898	*Antisémitisme et révolution*	Antisemitism and revolution	Lazare	Book	E	1	Jewish influence in society
1899	*Les memoires du prisonnier Max Regis*	The memoirs of the prisoner Max Régis	Régis	Memoirs	F	1	Algeria
1889	*L'antisémitisme algérien*	Algerian anti-Semitism	Rouanet	Speech	E	1	Algeria Jewish influence in society
1900	*Vers la lumière...* *impressions vécues: affaire Dreyfus*	Towards the light... lived impressions: The Dreyfus affair	Séverine	Book	F	2	protagonists of the Dreyfus affair role of press
1901	*L'Affaire Dreyfus: la vérité en marche*	The Dreyfus Affair: the truth on the march	Zola	Book	E	2	protagonists of the Dreyfus affair role of press

1913	*À nous la France*	Ours is the France	Blümchen	Book	F	4	Jewish influence in society characteristics of Jews religion naturalization
1914	*Souvenirs des milieux littéraires, politiques artistiques et médicaux*	Memories of literary, political, artistic and medical circles	Daudet	Book	F	1	intellectual circles
1914	*Droit de la race supérieure*	Right of the superior race	Blümchen	Book	F	4	Jewish influence in society characteristics of Jews religion Algeria

Table 2: Corpus Details

Title	# characters (original)	# words (original)	# words (post)	# unique (post)	Perc.(%) (original)
Vers la lumiere	415282	68937	28003	10800	6.11
Rothschild, Ravachol et Cie	68996	11308	4781	2472	1.02
Les memoires du prisonnier Max Regis	29299	4712	2111	1552	0.43
Les juifs en Algerie	339443	55929	23519	8428	5.0
L'Algerie de 1830 a1840	506886	82405	34786	9131	7.46
La conquete d'Alger	324783	52869	22435	7209	4.78
Juifs et antisemites en Europe	568291	95554	38610	11772	8.37
Algerie juive	333358	53957	21958	6942	4.91
Lettre à la jeunesse	10456	1752	695	528	0.15
Lettre à la France	21086	3537	1343	885	0.31
Le nationalisme juif	40201	6505	2525	1604	0.59
L'antisemitisme: son histoire et ses causes	626991	98468	43175	13679	9.23
L'antisemitisme algerien	176571	29133	11680	4947	2.6
L'Affaire Dreyfus	352466	58934	23262	7613	5.19
La verité sur l'affaire Dreyfus	106384	17467	7308	3414	1.57
Contre l'antisemitisme: histoire d'une polemique	56354	9372	3540	2087	0.83
Comment on condamne un innocent	62051	9776	4877	3304	0.91
Antisémitisme et revolution	28558	4846	1851	1265	0.42
Les grandes juives	92948	16494	7188	3527	1.37
La France juive	1022238	168658	73703	23844	15.05

A nous la France	112393	18698	8053	4374	1.65
Souvenirs des milieux litteraires	1441578	238687	100001	24875	21.23
Droit de la race superieure	54601	8570	4197	2707	0.8

Table 3: Numbers

antisemitism (0.20) history (0.17) dreyfus (0.01) intellectual (0.03) algeria (0.00)	juifs, juif, peuple, chrétiens, antisémitisme, pays, race, israélites, israël, siècle, saint, paris, rothschild, temps, juive, religion, russie, monde, dieu, société, esprit, chrétien, peuples, histoire, allemagne, foi, antisémites, france, lois, nation	*Topic*#1
Antisemitism (0.00) history (0.07) dreyfus (0.01) intellectual (0.01) algeria (0.15)	général, alger, chef, hommes, france, conquête, armée, ville, ordre, français, temps, oran, heures, ennemi, troupes, jours, duc, arabes, guerre, jour, commencements, afrique, comte, cents, commandant, ministre, place, française, grande, gouvernement	*Topic*#2
Antisemitism (0.01) history (0.09) dreyfus (0.23) intellectual (0.05) algeria (0.05)	dreyfus, justice, homme, guerre, affaire, vérité, lettre, jour, france, conseil, père, capitaine, honneur, nom, colonel, doute, cour, procès, président, paris, ministère, zola, presse, ministre, peuple, raison, commandant, question, lettres, monsieur	*Topic*#3
Antisemitism (0.054) history (0.052) dreyfus (0.017) intellectual (0.016) algeria (0.051)	juifs, français, algérie, france, juive, juif, argent, arabes, francs, gouvernement, droit, grâce, jour, pouvoir, indigènes, grand, française, travail, banque, titre, moment, lieu, chambre, décret, arabe, fortune, pays, payer, années, terres	*Topic*#4
Antisemitism (0.00) history (0.03) dreyfus (0.01) intellectual (0.13) algeria (0.00)	grand, temps, homme, petit, yeux, voix, ans, père, ami, mort, petite, tête, monsieur, monde, vieux, hugo, vie, pauvre, esprit, rue, maître, main, daudet, bonne, grande, rire, jeune, paris, belle, cœur	*Topic*#5

Table 4: Topics by LDA (word distributions)

Title	*Topic*#1 AS	*Topic*#2 AL	*Topic*#3 DRE	*Topic*#4	*Topic*#5 INTEL
Vers la lumière	0.0381	0.09	0.6027	0.0772	0.1919
Rothschild and Ravachol et Cie	0.2328	0.0607	0.1526	0.5148	0.039
Les memoires du prisonnier Max Regis	0.0831	0.1924	0.3163	*0.2238*	0.1844
Les juifs en Algerie	0.056	0.1456	0.1458	0.6093	0.0433
L'Algerie de 1830 a 1840	0.013	0.7878	0.0849	0.0749	0.0394
La conquete d'Alger	0.0247	0.7821	0.0881	0.0656	0.0395
Juifs et antisemites en Europe	0.4116	0.0661	0.1687	0.2492	0.1043
Algerie juive	0.0413	0.1194	0.075	0.753	0.0112
Lettre à la jeunesse	0.0835	0.04	0.7289	0.059	0.0886
Lettre à la France	0.045	0.0357	0.8321	0.039	0.0482
Le nationalisme juif	0.5443	0.0377	0.1551	0.2153	0.0475
L'antisemitisme: son histoire et ses causes	0.707	0.0521	0.0606	0.1351	0.0453
L'antisemitisme algerien	0.3374	0.0631	0.2367	*0.3243*	0.0386
L'Affaire Dreyfus	0.0345	0.0882	0.7409	0.0779	0.0585
La verité sur l'affaire Dreyfus	0.0231	0.0909	0.7432	0.1064	0.0363
Contre l'antisemitisme: histoire d'une polemique	0.3595	0.0292	*0.313*9	0.2049	0.0925
Comment on condamne un innocent	0.0728	0.0904	0.7024	0.1006	0.0338
Antisemitisme et revolution	0.4412	0.05	0.1921	0.2018	0.1148
Les grandes juives	*0.255*	0.1037	0.1453	0.1644	0.3316
La France juive	0.3338	0.1026	0.1877	*0.2087*	0.1672
A nous la France	0.2055	0.1446	*0.2322*	0.2572	0.1606
Souvenirs des milieux litteraires	0.0389	0.0655	0.1366	0.0767	0.6823
Droit de la race superieure	*0.2663*	0.0705	0.1548	0.3569	0.1515

Table 5 Topic Distribution

References

[1] Ambrosino, A., Cedrini, M., Davis, J. B., Fiori, S., Guerzoni, M., and Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology*, 25(4) :329–348.

[2] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4) :77–84.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022.

[4] Braude, B. (1976). Séverine," écrivain de combat". *Nineteenth-Century French Studies*, 4(3) :404–412.

[5] Brauer, R. and Fridlund, M. (2013). Historicizing topic models, a distant reading of topic modeling texts within historical studies. In *International Conference on Cultural Research in the context of "Digital Humanities", St. Petersburg : Russian State Herzen University*.

[6] Carron-Arthur, B., Reynolds, J., Bennett, K., Bennett, A., and Griffiths, K. M. (2016). What's all the talk about? topic modelling in a mental health internet support group. *BMC psychiatry*, 16(1) :1–12.

[7] Cheng, W., Greaves, C., Sinclair, J. M., and Warren, M. (2009). Uncovering the extent of the phraseological tendency : Towards a systematic analysis of concgrams. *Applied Linguistics*, 30(2) :236–252.

[8] Combeau, Y. (1993). Le boulangisme dans tous ses mouvements (1886-1891). *Mappemonde*, 3 :93.

[9] Crubellier, M. (1990). *Ce que l'on dit des juifs en 1889. Antisémitisme et discours social*. Presses Universitaires de Vincennes.

[10] Drumont, É. (1886). *La France juive : essai d'histoire contemporaine*. Vol. II in *La France juive*. C. Marpon& E. Flammarion.

[11] Faessler, E., Modersohn, L., Lohr, C., and Hahn, U. (2020). Progene-a large-scale, high-quality protein-gene annotated benchmark corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4585–4596. Fletcher, W. H. (2007). kfngram (version 1.3. 1). *Annapolis, MD : United States Naval Academy*.

[12] Flor, M., Fried, M., and Rozovskaya, A. (2019). A benchmark corpus of English misspellings and a minimallysupervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 76–86, Florence, Italy. Association for Computational Linguistics.

[13] Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 363–371.

[14] Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv :1605.09096*.

[15] Hofmann, T. (2013). Probabilistic latent semantic analysis. *arXiv preprint arXiv :1301.6705*.

[16] Jockers, M. L. (2013). *Macroanalysis : Digital methods and literary history*. University of Illinois Press.

[17] Kalman, J. (2010). *Rethinking antisemitism in nineteenth-century France*. Cambridge Univ. Press.

[18] Kauffmann, G. (2008). *Édouard Drumont*. Librairie Académique Perrin.

[19] Kauffmann, G. (2014). L'affaire de la viande à soldats. Une campagne antisémite en 1892. *Archives Juives*, 47(1) :28–36.

[20] Kreis, E. (2011). *"Quis ut Deus?" Antijudéo-maçonnisme et occultisme en France sous la IIIe République*. PhD thesis, EPHE.

[21] Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3) :259–284.

[22] Larsen, F. (2005). *Quand Paris était antisémite : Jules Guérin, roi de Fort Chabrol*. Ars Magna Edition.

[23] Li, W. and McCallum, A. (2006). Pachinko allocation : Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, 577–584.

[24] Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1) :1–22.

[25] Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., and Tolonen, M. (2020). Topic modelling discourse dynamics in historical newspapers. *arXiv preprint arXiv :2011.10428*.

[26] McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

[27] McDougall, J. (2017). *A history of Algeria*. Cambridge University Press.

[28] Méric (1989). "Urbain Gohier," Dictionnaire de biographie française.

[29] Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404–411.

[30] Moretti, F. (2000). Conjectures on world literature. *New left review*, 1 :54.

[31] Murakami, A., Thompson, P., Hunston, S., and Vajn, D. (2017). 'What is this corpus about?' : using topic modelling to explore a specialised corpus. *Corpora*, 12(2) :243–277.

[32] Newman, D. J. and Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6) :753–767.

[33] Parupalli, S., Rao, V. A., and Mamidi, R. (2018). Bcsat : A benchmark corpus for sentiment analysis in telugu using word-level annotations. In *ACL (3)*, 99–104.

[34] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 724–732.

[35] Petrucci, F. (2011). *Gli ebrei in Algeria e in Tunisia*, 1940-1943. Giuntina.

[36] Piskorski, J., Haneczok, J., and Jacquet, G. (2020). New benchmark corpus and models for fine-grained event classification: To bert or not to bert? In *Proceedings of the 28th International Conference on Computational Linguistics*, 6663–6678.

[37] Rayson, P. (2008). From key words to key semantic domains. *International journal of corpus linguistics*, 13(4) :519–549.

[38] Riddell, A. B. (2014). How to read 22,198 journal articles : Studying the history of german studies with topic models. *Distant Readings: Topologies of German culture in the long nineteenth century*, 91–114.

[39] Roberts, S. B. (2017). *Citizenship and Antisemitism in French Colonial Algeria, 1870–1962*. Cambridge University Press.

[40] Rouannet, É. (2016). Gustave Rouanet, un publiciste et parlementaire socialiste face a l'émergence de l'antisémitisme français (1885-1895). *Cahiers Jaurès*, (3) :57–84.

[41] Schnober, C. and Gurevych, I. (2015). Combining topic models for corpus exploration : Applying lda for complex corpus research tasks in a digital humanities project. In *Proceedings of the 2015 Workshop on Topic Models : Post-Processing and Applications*, TM '15, page 11–20, New York, NY, USA. Association for Computing Machinery.

[42] Scott, M. (1998). *WordSmith tools manual*. University Press.

[43] Sternhell, Z. (1996). *Neither right nor left: fascist ideology in France*. Princeton University Press.

[44] Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7) :424–440.

[45] Sullam, S. L., Minello, G., Tripodi, R., and Warglien, M. (2021). Representation of jews and anti-jewish bias in 19th century french public discourse: Distant and close reading. *Frontiers in big Data*, 4.

[46] Tripodi, R., Warglien, M., LevisSullam, S., and Paci, D. (2019). Tracing antisemitic language through diachronic embedding projections : France 1789-1914. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 115–125, Florence, Italy. Association for Computational Linguistics.

[47] von Wroblewsky, V. (1997). The early Sartre and the Jewish question. *Sartre Studies International*, 3(2) :21–28.

[48] Wang, W. Y. (2017). Liar, liar pants on fire : A new benchmark dataset for fake news detection. *arXiv preprint arXiv :1705.00648*.

[49] Wang, Y., Bai, H., Stanton, M., Chen, W.-Y., and Chang, E. Y. (2009). Plda : Parallel latent dirichlet allocation for large-scale applications. In *International Conference on Algorithmic Applications in Management*, 301–314. Springer.

[50] Weinberg, H. H. (1983). The image of the jew in late nineteenth-century french literature. *Jewish Social Studies*, 45(3/4) :241–250.

[51] Weingart, S. (2012). Topic modeling for humanists : A guided tour. *The Scottbot Irregular*, 25.

[52] Wilson, S. and Wilson, S. (1982). *Ideology and experience : antisemitism in France at the time of the Dreyfus affair*. JSTOR.

[53] Yang, T.-I., Torget, A., and Mihalcea, R. (2011). Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 96–104.