Università degli Studi di Milano
Data Science and Economics

Text Mining

Ivo Bonfanti
985892

Abstract

This work has a twofold objective, on one hand it aims to detect valuable insights concerning the academic debate on nuclear plants safety, while on the other, it points to build a model able to classify related unseen documents according to the topics treated. The goals are met mainly adopting two unsupervised text mining techniques, the hierarchical clustering algorithm and the latent Dirichlet allocation. Also, the topics polarization is investigated through the application of the joint sentiment-topic model, a semi-supervised technique. The methods are applied to a dataset built ad hoc collecting journal article abstracts from the International Nuclear Information System.

ACADEMIC YEAR 2022-2023

1 Dataset

The International Nuclear Information System (INIS), managed by the International Atomic Energy Agency, provides a useful repository collecting full-text documents from the nuclear literature. For the purpose of the analysis, 1250 abstracts, complete with their pubblication year, author names and title, are retrieved from the database. Only those english journal articles reporting the key words *'nuclear safety'* within the abstract are retrieved. The resulting dataset is organized as follow,

Table 1: Nuclear plants safety dataset head.

Abstract	**Title**	**Author**	**Year**
Although adequate levels of nuclear safety have been attained, the societal and institutional ap..	Commentary on the cost of nuclear safety	Mariani, L.P.	1991
This paper discusses how corporate nuclear safety committees use the principles of self-assessme...	Nuclear utility self-assessment as viewed by the corporate nuclear safety committee	Corcoran, W.R.	1992
The political changes in Europe broadened the scope of international nuclear safety matters cons...	The use of probabilistic safety assessments for improving nuclear safety in Europe	Birkhofer, A.	1992
In 1988, responding to a recommendation by the Inspector at the Sizewell B Public Inquiry, the H...	The 'tolerability' of nuclear risk?	Barker, F.	1992

2 Data Preprocessing

Before proceeding with the analysis the abstracts must first undergo to some sort of transformation. Once the punctuation is removed, they are tokenized by converting the texts into set of words. Among them those considered stop words, which are those most commonly used, thus not carrying any specific information about the text, are filtered out. Avoiding this step would make the comparison between documents really hard given that they would share most of the features. Moreover, all the words are converted to lower case and lemmatized. This last technique consists in turning each word in its base form since the machine would otherwise recognize as different tokens like *'going'* and *'go'*, which would be inconvenient for the purpose of this work. For the sake of clarity a sample abstract and its transformed version are reported here below,

[*Site or multi-unit (MU) risk assessment has been a major issue in the field of nuclear safety study since the Fukushima accident in 2011. There have been few methods or experiences for MU risk assessment because the Fukushima accident was the first real MU accident and before the accident, there was little expectation of the possibility that an MU accident will occur. In addition to the lack of experience of MU risk assessment...*]

[*'site', 'multi', 'unit', 'mu', 'risk', 'assessment', 'major', 'issue', 'field', 'nuclear', 'safety', 'study', 'since', 'fukushima', 'accident', '2011', 'method', 'experience', 'mu', 'risk', 'assessment',*

'fukushima', 'accident', 'first', 'real', 'mu', 'accident', 'accident', 'little', 'expectation', 'possibility', 'mu', 'accident', 'occur', 'addition', 'lack', 'experience', 'mu', 'risk', 'assessment'...]

3 Corpus Overview

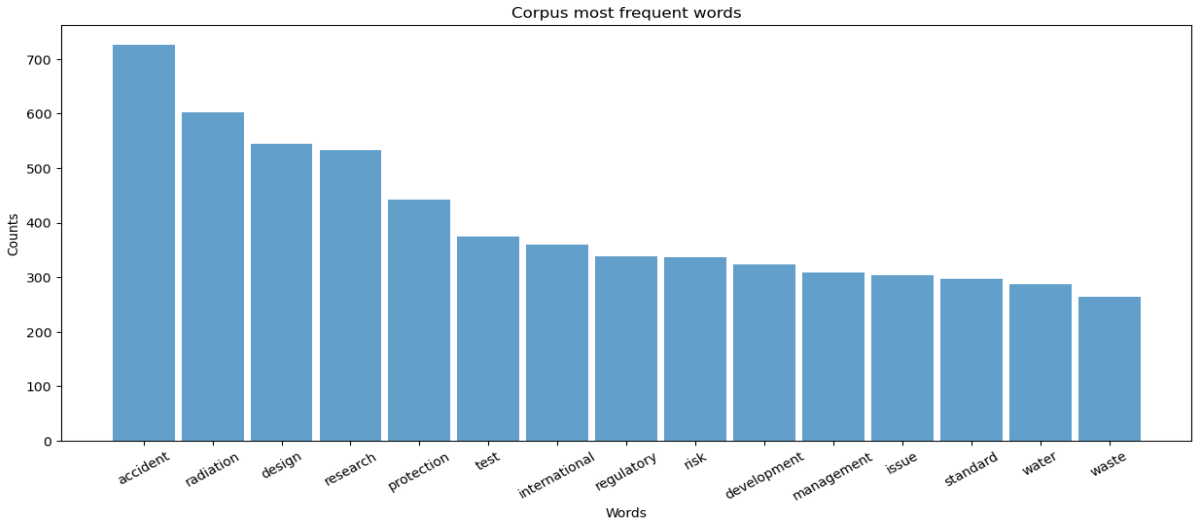
As a first step to determine whether the academic debate on the safety of nuclear plants is somewhat polarized, a lexicon-based approach is adopted. The lexical resource employed to extract a sentiment score for each token is SentiWordNet[1]. It assigns to each term three scores, determining how positive, negative and objective is the associated sentiment. For a more accurate evaluation, the score related to each token is the average of the scores of the terms belonging to its synset[1]. A very intuitive way to proceed is to compute for each document the average of the sentiment scores of its tokens. In this case the mean of the scores of the documents, representing the overall polarity of the corpus, is slightly positive: 0.015 out of a maximum of 1. Such a neutral result is not surprising given that in most cases academic statements are very specific, somewhat cautious and devoid of sensationalist tones. Out of 1250 documents only 214 report a negative score. Hereunder, those with the highest and the lowest scores are shown,

[*Highest score: 0.0799180751846193. The program of the improvement of nuclear safety in Eastern Europe offers both the Electricity Sector and the Spanish Nuclear Industry an opportunity to apply their wide technological capabilities and to contribute to strengthening of the nuclear option, which is perceived as being of fundamental importance to ensure the supply...*]

[*Lowest score: -0.056682266356339. The recent U.S. Department of Energy (DOE) nuclear safety policy establishes safety goals, which require that the risk to an average individual for prompt fatalities that might result from accidents should not exceed 0.1% of the sum of prompt fatalities resulting from other accidents and the risk...*]

The most frequent features, excluding those known to be numerous a priori as *'nuclear'* or *'safety'*, or those least informative as *'analysis'* or *'result'*, are shown in figure 1.

Figure 1: Most frequent tokens bar graph.



[1]Grouping of words expressing the same concept.

At first glance the security of nuclear power plants seems that can be challenged by *accidents*, unexpected events as natural disasters or unintended mistakes due to a negligent *management*. For this reason, to grant the highest possible security level, is very important to evaluate the *risks* and potential *issues* ex ante, during the *design* and *development* phase. The main danger is the dispersion into the environment of toxic radioactive *waste* due to the complexity of its disposal or due to the overheating of the coolant, *water* in most cases, which eventually can cause a nuclear meltdown. Somewhat more interesting and less obvious is the relation between nuclear safety and the probable lack of an effective *international regulatory* framework providing common *standards*.

4 Hierarchical Clustering

With the aim of inspecting in more detail the dataset, the hierarchical clustering algorithm is run. It is an agglomerative unsupervised classification method to group syntactically similar documents in disjoint clusters. It follows a bottom-up approach, by first assigning each document to a different cluster and subsequently merging those most similar until they all belong to the same one. The two clusters to be merged at each step are chosen following the Ward's method, which is based on the minimum variance criterion[2]. The documents similarities are computed using the cosine distance[3] and they are stored in the distance matrix. For this reason, the documents need to be first transformed into vectors, in this case using an embedding technique called Term Frequency-Inverse Document Frequency. Essentially, the TF-IDF technique creates a numerical representation of a document in relation to a corpus. Indeed, it turns it into a sparse vector of numbers, each representing the relevance, within the document, of a specific term of the corpus vocabulary. The score is computed as the product of the next two metrics,

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \cdot \log \frac{|D|}{|d \in D : t \in d|} \tag{1}$$

The first represents the relative frequency of a term t within document d, while the second expresses the rarity of t with respect the corpus D. It follows that a high score indicates that a word is frequent in a specific document but scarce in the overall corpus.

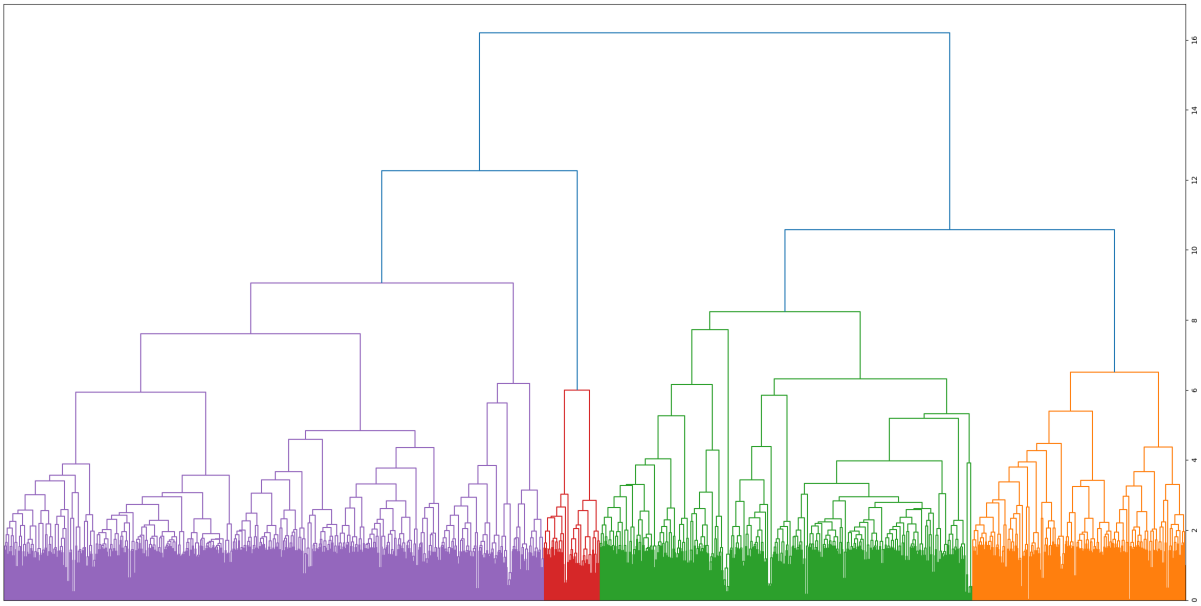
The HC dendrogram is cut in a way to obtain four clusters. The reason is that among all the partitions, the one resulting in four groups looked more distinguishing and in general more meaningful. The graph is shown in figure 2.

To get insights about the topics covered by each cluster, the most relevant tokens are retrieved from the TF-IDF matrix. Once again that is done excluding some uninformative terms. Also, the polarity score combined with the TF-IDF value, the average publication year, and the cluster size are reported for all. Next, a brief description of the contents of the clusters is given. The most relevant words are written in italic.

[2]It is a linkage method that merges clusters minimizing, at each iteration, the total within-cluster variance. Mathematically, at each step: $(x_k, x_l) = argmin_{(x_r, x_s) \in X^2} \|x_r - x_s\|$, where X is the dataset. Intuitively, and geometrically speaking, the merging is picked so that the resulting new cluster is the most compact among all the potential ones.

[3]Obtained by subtracting the cosine similarity to one. The cosine similarity is a metric measuring the angle between two vectors. Mathematically: $similarity(a, b) = \frac{a \cdot b}{\|a\| \|b\|} \in [0, 1]$, where a and b are two vectors.

Figure 2: Hierarchical clustering dendrogram.



The greatest cluster comprehends 571 documents and seems to deal with *neutron* monitoring solutions to ensure a safety management of nuclear reactors. Indeed, the *measurments* of the *neutron flow* released by the fission chain allows continuous control over the power released without the need to access to the reactor *core*.

The green cluster, counting 394 abstracts, regards two topics, not necessarily disjointed, one related to radioactive *waste management*, and one to *international conventions* and *regulations* aiming to establish a common nuclear *safety culture*.

The topic addressed by the orange cluster concerns safety by *design* and the importance of a *probability safety assessment* to minimize the *risk* of *severe accidents* like the one happened in *Fukushima* in 2011. The cluster counts 226 elements.

Last, the smallest in size, counting only 59 documents, deals with a very specific topic: *cladding*. Fuel cladding is the outer layer of fuel *rods*, which prevents *reactivity-initiated accidents* as the *pellet deformation* at high *temperatures*. That is important to avoid the contamination of the coolant with radioactive products.

While the mean years are similar between the clusters, ranging from 2004 to 2010, the polarity scores are quite different. They are all positive, but if we consider as a reference the scores of the orange and the purple groups, which are very similar, the green one is about one and a half times greater while the red one is definitely smaller, about one tenth.[4]

[4]Given the combination between polarity and terms relevance, the resulting scores make sense only if compared each others. Their actual value is of little importance since the range is no more $[-1, 1]$.

5 Latent Dirichlet Allocation

The outcome of the HC algorithm is useful to determine the number of topics treated in the corpus. This value can be exploited to set the a priori number of topics required by the latent Dirichlet allocation[2]. This last is a generative probabilistic model that can be employed, among all applications, also in an unsupervised text mining context.

LDA relies on Bayesian probability theory to estimate the distributions of the documents over the latent topics[5] and the distributions of the topics over the words[6], assuming both words and documents exchangeability[7]. The document-topics θ and the topic-words ϕ latent variables are assumed to be drawn from two Dirichlet distributions with parameters α and β, respectively: $\theta \sim Dir(\alpha)$ and $\phi \sim Dir(\beta)$. The LDA ultimate objective is the estimation of these two parameters, and the above mentioned distributions represent the researcher's prior beliefs on them before experiencing the data. The likelihood function instead captures the probability of generating the observed documents given the latent variables. It is derived from the two step generative process reported here below,

For each of the N words w_n within document i:

1. Draw a topic $z_n \sim Multinomial(\theta_i)$
2. Draw a word $w_n \sim Multinomial(\phi_n)$

Leading the likelihood of word j of document i to the form:

$$P(w_{ij}|\theta_i, \phi_j) = \sum_k (P(z_{ij} = k|\theta_i) \cdot P(w_{ij}|z_{ij} = k, \phi_j)) \tag{2}$$

Eventually, given the exchangeability assumption, the corpus likelihood is:

$$P(W|\theta, \phi) = \prod_{i=1}^{M} \prod_{j=1}^{N_i} P(w_{ij}|\theta_i, \phi_j) \tag{3}$$

The posterior distribution, accordingly with the Bayes theorem[8], can be approximated as the product between the likelihood function and the prior distributions. However, due to its intractability, the Gensim Python package providing the LDA model approximates it using a variational Bayesian method. Briefly, the original distribution is substituted by a different one chosen from a simpler family of distributions, found by minimizing the Kullback-Leibler divergence[9].

[5] Documents are intended as a mixture of topics.

[6] For each topic, words are assigned different probabilities. To give an example, the word 'cloud' is associated to a higher probability if related to topic 'nature' rather than topic 'technology', or if in relation to topic 'technology' rather than topic 'emotions'.

[7] Their ordering is ignored by the model.

[8] $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \simeq P(B|A) \cdot P(A)$

[9] $KL(P||Q) = \sum P(x) \cdot \log \frac{P(x)}{Q(x)}$, where P and Q are two probability distributions.

The model is run with the premise that the corpus focuses on four latent topics. As already mentioned, such an intuition comes from the HC outcome. However, the LDA result, i.e. the parameters posterior distribution, differently from HC, can be exploited to determine the topic mixture of new unseen documents. Moreover, while HC labels each document with its dominant topic, LDA assigns a mixture of topics, which is more reasonable. Indeed, because of topics interconnections, it is more likely for a document to treat different arguments instead of a single one. As can be noticed from table 2 the topics extracted match closely to those detected by the HC algorithm.

Table 2: Top 20 words per topic.

International law	**Dosimetry**	**Cooling system**	**Accident**
radiation, protection, waste, international, agency, state, management, iaea, research, regulatory, programme, convention, european, regulation, country, radiological, environmental, national, atomic, cooperation	neutron, radiation, dose, measurement, source, measure, sample, field, monitoring, standard, protection, dos, value, dosimetry, detector, concentration, obtain, test, parameter, developed	test, accident, water, condition, heat, temperature, flow, core, simulation, research, experiment, rod, cladding, experimental, containment, cool, transient, calculation, coolant, thermal	accident, design, risk, development, research, issue, regulatory, event, approach, public, security, npp, management, make, need, industry, standard, important, support, improve

To demonstrate its functioning, the pretrained LDA model is applied to the unseen document reported beneath. It classifies it coherently, as 35% dealing with cooling system and 65% dealing with accident.

[*Fuel cladding, the protective shield around nuclear fuel pellets in reactors, is essential for nuclear power plant safety. However, it can pose challenges when compromised, leading to accidents. Fuel cladding serves to contain radioactive materials, ensuring they do not escape into the environment. It also maintains fuel rod structural integrity under extreme conditions. Historical nuclear disasters, like Chernobyl and Fukushima, underscore the importance of fuel cladding. Failures in cladding contributed to the severity of these accidents. Stringent international safety standards, advanced materials, and proactive measures, such as inspections and research, help mitigate fuel cladding-related risks. As the world seeks sustainable energy, nuclear power remains an option. Prioritizing fuel cladding integrity, adhering to safety standards, and continuous research are essential to secure the future of nuclear power while ensuring public safety.*]

LDA is run also considering bigrams instead of single words, however the results are somewhat disappointing given that only two out of four topics are clear and well defined, one dealing with radioactive waste and the other with risk assessment.

6 Joint Sentiment-Topic Model

The joint sentiment-topic model [3] aims to capture the relation between topics and words sentiments. It is based on the intuition that the sentiment associated to a term is not absolute, but strictly dependent on the context in which the term is used. A trivial example might be the

word *thin* which carries a positive sentiment if used to describe the compactness of a computer, while it can turn negative if related to the insulation capacity of the glass of a window. The JST model architecture replicates that of LDA, except for one additional layer on top referring to sentiments. Indeed sentiment labels, positive and negative in this case, are associated with documents, topics are associated with sentiments and words with both, topics and sentiments. JST does not assign only one topic distribution per document, but as many as the sentiment labels, therefore introducing one additional latent variable.

To estimate the posterior distribution the Jointtsmodel Python package makes use of the Gibbs sampling inferential procedure, a Markov chain Monte Carlo method. It is a process which iteratively updates each variable by drawing it from its conditional distribution given the current value of the others. On one hand it is more accurate than variational Bayes, while on the other it is more time consuming.

Another distinguishing factor from LDA is that JST is a semi-supervised learning method. Indeed it takes into consideration within the initialization phase of the Gibbs sampling a partial list of words labelled with their sentiments.

Table 3: Top 20 words per topic and sentiment.

International law +	**International law -**	**Accident +**	**Accident -**
international, waste, iaea, country, management, european, research, protection, state, programme, regulatory, convention, regulation, national, agency, cooperation, atomic, development, act, area	design, regulatory, standard, iter, component, technical, inspection, program, control, operating, commission, maintenance, licensing, test, canadian, engineering, current, operational, procedure, must	research, year, technology, report, issue, japan, world, well, period, accident, unit, development, environmental, start, strategy, institute, cost, 2011, document, chernobyl, korea	accident, risk, public, event, security, human, factor, fukushima, culture, approach, improve, fire, propose, show, need, important, psa, npps, impact, china

Unfortunately the model does not fit very well the data. Only two out of four topics show a fairly clear distinction between positive and negative sentiments, although still it leaves room for interpretation. As reported by table 3, on one hand international *regulations* and *cooperation* between *countries* can enhance the *development* and the safety of the nuclear technology, while on the other these norms require more duties and higher costs for the involved countries since new *standards*, new *procedures*, new *licenses* and *inspections must* be respected. With regard to the second topic, *accidents* due to unexpected *events* pose a severe *risk* to the *public* security as what happened in *Fukushima*. However it is also true that *accidents* as those occurred in *Japan*, *Korea* or *Chernobyl* boosted the *research* towards the *development* of an effective *strategy* to improve the nuclear plants safety and protect the *environment*.

Following some intuitions from recent literature[4] the reversed joint sentiment-topic model is also applied to the data. Intuitively it reverses the JST assumptions, considering documents as a mixture of topics first and only then as a mixture of sentiments. Therefore JST works better at identifying the sentiment associated to a document, while rJST at assigning a sentiment to each topic. Some authors suggest following this reasoning that JST is then more appropriate to product review, where a dominant sentiment is likely to emerge, while rJST works better with

more complex documents, as political speeches or scientific debates. However, in this work the rJST outcome is almost identical to the one of JST and under some aspects even worst. This is in accordance to older literature[5] assigning to JST the primacy for joint sentiment and topic detection.

7 Conclusion

Despite being usually less accurate than supervised methods, unsupervised and semi-supervised techniques are increasingly widespread for natural language processing. Indeed their strength consists in requiring little or any additional information to train the models if not those provided by the dataset analyzed.

This project makes use of them to extract valuable information from a collection of academic abstracts concerning nuclear plants safety. After due data cleaning and transformation the hierarchical clustering algorithm allowed to group the most similar documents. Then the number of clusters detected is used to set the a priori number of topics required by the latent Dirichlet allocation, a Bayesian model able to determine the topic mixture of the abstracts. Once trained, the model can identify the latent topics of unseen documents and classify them accordingly, in a completely unsupervised manner. A concrete application of this model could be the automated management of large documents archives.

Furthermore other experiments have been conducted within the analysis. A lexicon approach based on SentiWordNet tried to capture the polarization of the debate, while the joint sentiment-topic model, along with its reversed version, attempted to model the relation between topics and sentiments. However the results of these last experiments are disappointing.

References

[1] Andrea Esuli Stefano Baccianella and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pages 2200–2204, 2010.

[2] Andrew Y. Ng. David M. Blei and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[3] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 375–384, 2009.

[4] Martijn Schoonvelde Max Boiten, Christian Pipal and Gijs Schumacher. Validating the joint sentiment topic model and the reversed joint sentiment topic model on political text. 2019.

[5] Yulan Lin, Chenghua; He and Richard Everson. A comparative study of bayesian models for unsupervised sentiment detection. *The 14th Conference on Computational Natural Language Learning.*, 2010.