Where to land after the pandemic? Semi-supervised thematic clustering on a post COVID-19 world survey

Salomé Do[1, 2], Jean-Philippe Cointet[2], Thierry Poibeau[1], and Donato Ricci[2]

[1]LATTICE, CNRS & École Normale Supérieure/PSL & Univ. Sorbonne nouvelle
[2]Sciences Po, médialab, Paris, France

Abstract

Where to land after the epidemic? This is the question asked by Bruno Latour in a digital questionnaire about the world-to-come after the COVID-19 crisis. This work dives into respondents' answers to two questions: Which activities would you see not coming back after the crisis? Which activities would you like to see developed? In order to discover the prominent themes raised by the respondents, we provide an original semi-supervised thematic clustering pipeline, and a dedicated evaluation tool. With this approach, we find that respondents would like to reduce hypermobility, over-consumption and polluting activities to the benefit of local, sustainable supply chains and transports. Besides these experimental results, our pipeline successfully deals with the issue of modelling topics on small corpora composed of short texts, notably by using a transfer-learning based sentence encoder yielding near state-or-the-art results on Semantic Textual Similarity (SST) benchmarks. Interpreting topics and assessing their quality is also easier than for traditional topic models. Yet, our work is simple to implement and runs on a relatively modest hardware configuration.

1 Introduction

At the peak of the COVID-19 pandemic in Europe, around mid-March 2020, the sociologist and philosopher Bruno Latour published a provocative article about the world-to-come after the crisis caused by the virus. (Latour, 2020). Latour wanted to take profit of the suspension of most productive activities and of the enforcement of social restrictions to make people think of a better future, through an exercise of personal and collective imagination. In the form of an open digital *questionnaire*, the exercise proposed to describe a) which activities, services and situations people wished to stop and, b) which ones they wanted to see arise and to support, once the pandemic is over.

This work aims at providing a reliable topic detection method to analyze the answers to the questionnaire. Leveraging open-ended answers in a digital survey is particularly challenging because of the small corpus volume (1083 users took part in the survey), but also because of the diversity of writing styles. In order to address these linguistic challenges, we developed an original pipeline (see Figure 1). First, the answers are broken into smaller parts, that we call segments. Second, the propositions are embedded in a shared vector space. We tested several popular sentence embedding algorithms for this. Because of the limited amount of text available, we make use of transfer learning techniques. The various embeddings are then concatenated and reduced in dimension, in order to eliminate redundancies. Third, propositions are clustered either in coherent topics or in a noise cluster. Noisy segments are then re-affiliated to coherent topics via a supervised classifier. We finally propose a quantitative evaluation of the algorithm, and compare our results against LDA, one of the main topic modelling algorithm used nowadays. Although we are illustrating our method on this specific questionnaire, we think our workflow could be applied in various settings.

The main contributions of this article are the following:

- We provide an original and reproducible thematic clustering pipeline, especially suited for small corpora of short texts.

- We provide a tool for the thematic analysis of a survey on the consequences of the COVID-19 pandemic. this makes it possible to model how respondents considered human activities such as mass tourism, mass consumption, transports, work, agriculture, etc.

- We evaluate our sentence embedding model on both SentEval and a human-annotated semantic similarity task applied to our corpus. We also provide a thematic clustering evaluation task called the *sentence intrusion* test and we report our results.

2 Related work

Topic modeling has become the archetypal technique for the automatic detection of topics from a corpus of texts. The most popular method is LDA (Latent Dirichlet Allocation), proposed in the early 2000s by (Blei et al., 2002). This model exploits a generic bayesian framework to statistically infer a generative model from a set of documents. It assumes that each document discusses a set of topics, and that each topic is composed of specific words. Topics are then defined as a probability distribution of words. Conversely documents are defined as a probability distribution of topics. As such, topic modeling adopts a traditional bag of words model of text, inferring structure in the corpus from the sole statistics of co-occurrences between words. In the case of LDA, the distribution is assumed to have a sparse Dirichlet prior. Since then, various alternative forms of topic models have been developed and popularized for text mining tasks (e.g. Structural Topic Models (STM) is very popular among political scientists (Roberts et al., 2013)), incorporating multiple hypotheses but still inferring the parameter of the generative model as a joint mixture of probabilities. However topic modeling of short documents is challenging (Cheng et al., 2014), calling for the use of more complex language model. (Qiang et al., 2016; Dieng et al., 2019) suggest that incorporating word embeddings in a topic model is helpful to produce a more accurate model. We adopt a similar strategy here, to encompass the limitations of simple bag of words models.

In (Demszky et al., 2019), authors analyze a large set of tweets and cluster them through topical categories. GloVe embeddings are trained on the corpus, tweet embedddings are then inferred using SIF (Smooth Inverse Frequency) (Arora et al., 2016), and are grouped thanks to a k-means clustering algorithm using a cosine distance between embeddings. Our method shares similarities with this workflow while addressing the supplementary challenge of data scarcity, notably through stronger segment embeddings and by adding a supervised step to the clustering part.

3 Questionnaire design

On March 30th 2020, a few weeks after the beginning of the lockdown measures in Europe, Bruno Latour (Latour, 2020) invited his readers to take advantage of the lockdown measures to question the pre-COVID situation. The aim of the experience was to describe, through a digital questionnaire *a) which activities were to encourage to start and b) which ones to stop definitely in the post-pandemic world.* In the same way as "protective measures" had been put in place against the virus, Latour spurred readers to imagine some measures against the return to a productivist and capitalist pre-COVID system.

The questionnaire has been conceived as a tool – hosted on a digital platform[1] – aiming at gathering and harvesting first-hand experiences and propositions rather than floating and general opinions. The answers to the questionnaire were subsequently used to propose thematic online workshops, where participants could further discuss their ideas in small groups. Our study work was first motivated by the desire to use automatic clustering to help organizers design thematic workshops.

In total, 1083 users participated in the online experiment, amongst which 696 answered in French and 387 in English. The answers are written in diverse styles , answers are of various length, sometimes being very personal and developed, and sometimes with a more impersonal enumerative style. Answers are also very composite, pointing out a large scope of issues, ranging from transportation to well-being, through topics like "gift economy" or consumption.

As a result, the main challenge when working on this corpus is to jointly deal with its high diversity and its small size, which harm the ability of traditional algorithms to work properly (because of the lack of repetitions and redundancies in our corpus). A classic LDA model applied on the English answers did not give satisfactory nor properly interpretable topics.

In the following section, we show how these specificities were handled by using transfer learning to extract themes from the answers.

[1] https://ouatterrir.medialab.sciences-po.fr/

I don't want the return of such extensive use of cars, ships and planes, along with global trade and travel. I don't want a return to the existing economic system which is so unbalanced. I don't want to lose the spirit of co-operation which emerged as a result of the crisis.
1)Commute every day to work at the same hours *2) Mindset which assumes that work from home is not possible.* *3) Poor meeting discipline: Since I run all my meetings on webex, I have noticed better meeting discipline, more effective and more constructive - leading to more fruitful discussions.* *4) Wear smart clothes for work everyday when a clean pair of trousers and 'smart casual' is just doing fine* *5) I would like to see cars not coming back (petrol engine)*
Rush

Table 1: Some examples of answers to the question "Which activities would you like to stop?". Writing styles may vary a lot among the samples, but generally share an enumerative structure.

4 Semi-supervised thematic clustering

Topic modelling aims at discovering topics in a set of documents. Topics are defined as a distribution on words, and documents as a distribution over topics. Thus, topic modelling algorithms such as LDA (Blei et al., 2002), or variational autoencoder (VAE)-inspired models (Miao et al., 2017; Srivastava and Sutton, 2017) are expected to learn these distributions from the corpus. However, these models only use co-occurrences to learn the distributions, which implies that the corpus is large enough and repetitive enough to contain this information. More recently, models like Embedding-based Topic Model (ETM) (Qiang et al., 2016), or Gaussian Bi-directionnal Adversarial Training (Gaussian-BAT) (Wang et al., 2020) incorporate word embeddings to include more global information about (contextual) word meanings.

These models use word-level transfer learning through word embeddings, and usually represent documents and topics with bag-of-words (ETM), TF-IDF (BAT), or multivariate Gaussian (Gaussian-BAT) approaches. Staying at the word level, and not assuming a compositional structure on the text is necessary to stick with the topic modelling assumption (topics are word distributions and documents are topics distributions). This assumption is generally more adapted to a large scope of corpora, where texts are composed of full-size paragraphs containing intertwined topics.

In our corpus, answers contain few co-occurrence information, are highly composite, but are generally structured as enumerations, due to the nature of the proposed task (*Can you list suspended activities that you would like to see (not) coming back?*). This prior on the structure of our answers, and the distribution of the topics in our answers (one topic per sentence, per bullet point, per line, ...) allows us to move from a word level to a "sentence level" topic analysis. In order to take advantage of this well-defined structure, we propose the following method (schematically represented in Figure 1):

4.1 Segmentation

We start by breaking down the answers in short, coherent segments. Segmentation is rule-based, corresponding to the observed structure in the dataset. At first, answers are separated on the line escape (`\n`) character. While this rule captures the prevailing "bullet point enumeration" pattern, it does not capture a less common, but still present, pattern of in-line enumeration. The in-line enumeration pattern corresponds to respondents concatenating many propositions in one sentence. To capture this, we run a sentence tokenizer algorithm, and then split long sentences on commas or semicolon characters. The criterion for splitting long sentences is a length threshold, empirically fixed at 100 characters. This rule-based segmentation proves to be generally robust, even though some specific sentence constructions (short sentences apposing keywords, for instance) could not be captured.

4.2 Segment embedding

Once the segmentation is done, the segments are embedded with three different sentence embedding models. Segment embedding is the core part of our thematic clustering model, as it enables us to take advantage of transfer learning methods to leverage the challenges brought by the size of our dataset and its high diversity.

To do so, we use SIF embeddings (Arora et al., 2016), Universal Sentence Encoder (USE) (Cer et al., 2018), and BERT [CLS] embeddings (Devlin et al., 2018). SIF (Smooth Inverse Frequency) embeddings aim at finding weights to aggregate word-level embeddings. In this setting, words with low frequency are given a higher importance, which is interesting in a thematic clustering context. Universal Sentence Encoder (USE) consists of a Transformer (Vaswani et al., 2017) architecture trained on both an unsupervised Skip-Thought (Kiros et al., 2015) objective and a supervised objective on the

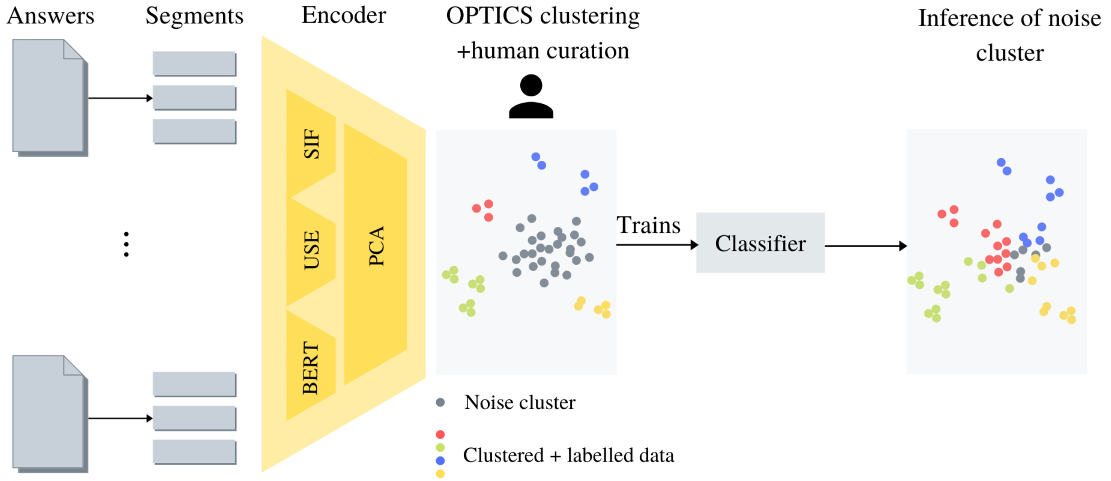


Figure 1: Schematic representation of our pipeline. Answers to the questionnaire are split into segments, which are in turn embedded in a vectorial space. Segments are then clustered using the OPTICS algorithm. Clusters are manually curated to merge or discard certain groups. Finally, a classifier is trained to categorize a larger set of segments that were still unassigned.

SNLI (Bowman et al., 2015) dataset, and is similar to InferSent (Conneau et al., 2017). BERT is a pre-trained language model, also using Transformers. It is pre-trained on a Cloze and a next sentence prediction task, and has proven to give robust results when fine-tuned on a large scope of downstream tasks. It has been showed that the algorithm gives deceiving results (Table 4) on unsupervised sentence embedding benchmarks such as SentEval (Conneau and Kiela, 2018), but fine-tuned versions as S-BERT (Reimers and Gurevych, 2019) or SBERT-WK (Wang and Kuo, 2020) have proven to be strong on SentEval (Table 4).

In order to avoid dimension problems, the embeddings resulting from SIF, USE, and BERT are scaled, concatenated, and reduced to dimension 100 with PCA. The scaling and dimension reduction is only possible on small datasets, as it requires to keep the three embeddings for the whole corpus in the computer's RAM. This model is thus not adequate for large corpora.

4.3 Clustering

In order to retrieve the main themes covered by respondents, we used OPTICS clustering (Ankerst et al., 1999), with cosine distance chosen as distance metric. We chose OPTICS rather than k-means clustering because we prefer to identify small, highly homogeneous clusters (Table 2) rather than assigning a cluster to every single segment. In OPTICS clustering, high-density zones are identified, and the rest is considered as noise. This clustering algorithm choice is guided by our thematic clustering objective: discovering reliable clusters is more important than covering the whole dataset.

4.4 Supervised re-affiliation of noisy segments

As reliability is chosen over exhaustiveness for the clustering phase, a large number of sentences are discarded as noise. In order to correct this, in a second phase, we use the well-clustered segments as a supervised text classification dataset, and then do inference on the noisy cluster.

Cluster curation. OPTICS clusters are thus first curated by a human annotator, who labels homogeneous clusters and discards non-coherent ones. Some homogeneous clusters can be annotated with the same label (*transports - air travel* for instance), and are then merged together. Some examples of homogeneous cluster are given in Table 3. In this table, we took the first six homogeneous clusters, and reported three segment samples for each of them.

	OPTICS Nb. of non-noisy clusters	Proportion of noisy segments	Nb. clusters after curation
Stop-En.	17	82.9 %	11
Stop-Fr.	29	88.8 %	15
Develop-En.	14	88.6 %	9
Develop-Fr.	21	93.7 %	11

Table 2: OPTICS results for English and French corpora on "Which activities do you want to stop?" and "Which activities do you want to develop?" answers. First columns shows the number of non-noisy clusters detected by OPTICS. Second columns shows the proportion of segments labeled as noisy (not clustered) by OPTICS. Third columns shows the number of final clusters after human curation. Human curation involves : a) labelling clusters as homogeneous or heterogeneous, b) merging homogeneous clusters sharing the same themes. Noisy segments, and segments in heterogeneous clusters will be re-affiliated to curated clusters via a supervised text classification algorithm.

k	Segment samples
1	*mass consumption of superfluous products* *compulsive consumption* *unconscious consumption*
2	*resource depletion* *desertification* *biodiversity loss*
3	*manufactured agricultural products promoting the culture ecologically* *deforesting for monoculture agriculture and livestock in extensive areas* *ban destruction of wild animal habitats so they stop moving towards human habitats*
4	*neoliberalism growth oriented economic thinking* *a capitalist western patriarchal mono* *capitalistic economy without limits*
5	*frequent shopping* *shopping in large malls* *frivolous shopping in shopping malls*
6	*short-haul flights* *cheap far distance flights* *frequent air travel*

Table 3: Clustering on "Stop" answers in English finds 11 homogeneous cluster, after human curation. This table presents three segment samples chosen at random for 6 out of these 11 clusters, in order to show their thematic coherence. These 11 clusters are then used as a supervised dataset for the supervised re-affiliation of noisy segments.

Re-affiliation. After this cleaning stage, pairs of well-clustered segments and their clusters' label are used as a training dataset for text classification. We use a simple logistic regression model, taking the segments embeddings as inputs. The model is used on inference on the noise cluster, and the probability relative to the predicted class is reported for each segment.

5 Evaluation

5.1 Segment Embedding

Our model is evaluated first on the Semantic Textual Similarity task of SentEval (Conneau and Kiela, 2018) (Table 4). Semantic Textual Similarity consists of comparing a human-rated similarity between couples of sentences to their embeddings' cosine distance. We compare our results with the results reported for the rest of the sentence embedding literature on (Wang and Kuo, 2020)[2]. Our model yields encouraging results given its straightforward nature. Interestingly, the concatenation of the three embeddings performs better than any of these embeddings alone, indicating that they may each capture different aspects of the segments.

We also evaluated the model on our dataset, by picking 60 couples of segments and having three annotators rating their similarity. Human-rated similarity is a score between 0 and 5, with 5 coding for total similarity, and 0 coding for totally non-similar segments. Segment couples were sampled so that approximately all scores between 0 and 5 could be equally represented in the annotation. We report a Krippendorf's alpha of 0.75 for the inter-annotator reliability, and a Pearson correlation coefficient of 0.68 between the median of the three annotation and the cosine similarity calculated with our embeddings. These results are considered satisfactory.

5.2 Sentence Intruder Task

Finally, we evaluate the complete pipeline with a *sentence intruder* task. For this task, a cluster k (defined as the original OPTICS cluster plus segments re-affiliated to this cluster with a probability greater than a threshold p) is chosen. Four segments are randomly drawn for cluster k, and one segment is drawn from another random cluster. A human annotator has then to find the intruder among the

[2] Their evaluation of Sentence-BERT gives lower Pearson coefficients than the Spearman coefficients reported in (Reimers and Gurevych, 2019). We chose to report (Wang and Kuo, 2020) results for compatibility reasons. We don't know whether the difference is due to coefficients differences or not, but remain critical about it.

Model	SST 12	SST 13	SST 14	SST 15	SST 16	Avg.
GloVe BOW	52.3	50.5	55.2	56.7	54.9	53.9
SIF	56.2	56.6	68.5	71.7	-	63.5
InferSent	59.2	58.9	69.6	71.3	71.5	66.1
USE	61	64	71	74	74	68.8
BERT [CLS]	27.5	22.5	25.6	32.1	42.7	30.1
Avg BERT	46.9	52.8	57.2	63.5	64.5	56.9
SBERT	64.6	67.5	73.2	74.3	70.1	69.54
SBERT-WK	**70.2**	68.1	**75.5**	**76.9**	74.5	**73.0**
Our encoder	63.9	**70.3**	72.7	76.6	**77.1**	72.1

Table 4: Results on Semantic Textual Similarity (SST) benchmark, reported as $100 \times \rho$, where ρ is Pearson correlation coefficient between human-rated similarity and cosine similarity. Evaluation of other models' performance are taken from (Wang and Kuo, 2020).

Pred. thres.	Stop-En.	Stop-Fr.	Dev-En.	Dev-Fr.
$p > 0.5$	72.7	71.1	59.2	45.4
$p > 0.75$	87.8	82.2	92.5	78.7

Table 5: Sentence intruder results for different probability thresholds ($p > 0.5$, $p > 0.75$). Probability thresholds are used in the reaffiliation process : when $p > 0.75$, we only keep segments where the theme has been predicted by the classifier with probability $p > 0.75$. In the sentence intruder task, pools of 5 segments are created, with four segments belonging to the cluster, and one segment drawn at random. A human annotator has to find the intruder among the 5 segments. Results are given as the accuracy (Acc.) averaged over all clusters. Cluster-by-cluster accuracy are given in appendix (Tables 7a, 7b, 7c, 7d). "Stop-En." stands for English "Which activities do you want to stop" answers, "Dev.-Fr" stands for French "Which activities do you want to develop" answers, etc.

five segments. For every corpus, we create three *sentence intruder* sets per cluster, for every cluster in the corpus. Results reported on Table 5 show the accuracy for all corpora, for two probability thresholds: 0.5 and 0.75. A "chose intruder at random" baseline would achieve 20% accuracy with this setting. For every corpus, our model largely exceeds the random baseline. Probability threshold $p > 0.5$ shows mitigated results, especially for the "Which activities do you want to develop?" question, but $p > 0.75$ achieves good results, indicating a more reliable model. In order to be able to distinguish problematic themes from well-predicted theme, Tables 7a, 7b, 7c, 7d (Appendix) provide theme-wise *sentence intrusion* accuracy for probability threshold $p > 0.75$. As only three different sets of *sentence intrusion* tests are evaluated per theme, theme-wise accuracy can only take values in $[0.333, 0.667, 1]$, not allowing a high-level analysis of theme-wise performances, but nevertheless indicating most problematic clusters.

Finally, qualitative samples of predictions are provided in Table 6. These six samples are randomly drawn from all available predictions.

Sample	Topic	Prob.
overdose of fuel oil powered automobiles	commuting/work	0.40
over-consumerism especially in terms of clothes and food	consumption	0.77
flying everywhere just for individual pleasure	air travel	0.78
excessive middle and long distance mobility of humans	commuting/work	0.47
require work-at-home for every conceivable job	work	0.96
these thoughts distract you sometimes from the things that really matter and to focus on the person you are	corporations/ business/ banks	0.61

Table 6: Some samples of final predictions, with associated probability for English answers to the question "Which activities do you want to stop?".

6 Results

Aggregated results for English answers to the question "Which activities do you want to stop?" are presented in Figure 2. Results to the answers to the other question "Which activities do you want to develop?", and French answers are presented in Appendix (3a, 3b, 3c). These figures show the proportion of segments predicted as belonging to a given theme (in percent), with a colour scheme indicating the probability they were predicted with, so that distributions can be put into perspective. A prediction probability of 1 corresponds to the segment being in the original OPTICS cluster.

Emerging themes come from answers to the question: "Which activities do you want to stop?",

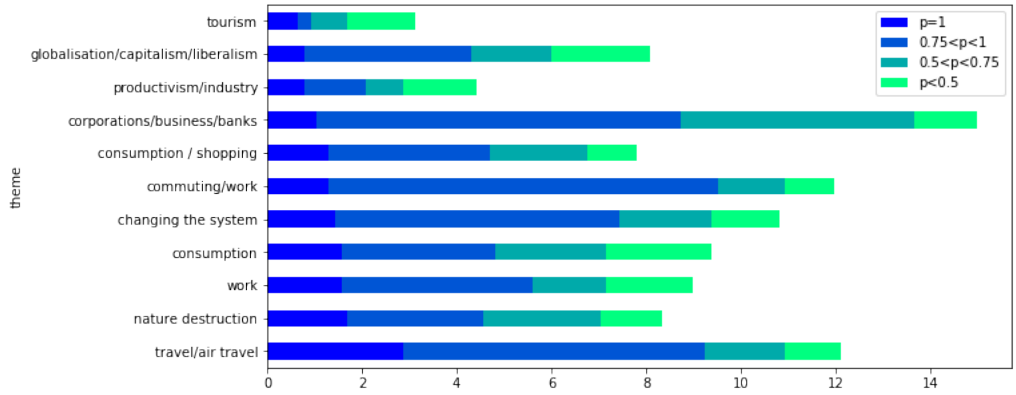


Figure 2: Theme distribution for Stop - English. The graph indicates the distribution of the discovered themes among the English "Which activities do you want to stop?" answers dataset (percentages). These distributions are showed for different predicted probability thresholds, for instance the segments predicted as *consumption*-related with $p > 0.75$ account for approximately 5% of the dataset. $p = 1$ means that the segment was present in the OPTICS cluster in the first place.

as well as to the question: "Which activities do you want to develop?". First, consumption, overconsumption, and more generally consumerism are major themes to be stopped, echoed by sustainable and local consumption in "Which activities do you want to develop?" themes. Hyper mobility (mass tourism, air travel, cruise ships) and day-to-day transports (cars, commute, traffic) also represent major themes to be stopped and replaced by transport reduction, local or alternative tourism, and the development of alternative transports, with bike lanes construction. Globalisation and imports/exports are problematic themes combining consumerism and hyper-mobility, that respondents want to stop in favor of local supply chains. Nature destruction, in the English dataset, and pollution together with intensive agriculture and over-construction in the French dataset, are important ecology-related themes that do not seem to have a direct, specific countermeasure in "Which activities do you want to develop?" answers, apart from aforementioned local/sustainable measures. Banking and trading, although the English corporation/business/banks cluster does not seem reliable enough to be consistently analyzed, are minor but existing themes, especially in French corpus. The daily commute to work, work schedules, and the workplace in general are less salient themes, but are associated to a work transformation cluster in the English dataset.

Though, respondents do not seem to have fully followed the instructions of not designating abstract entities as "capitalism", especially in the English corpus where two clusters are dedicated to liberalism and capitalism, together with "changing the system". However, these rather political clusters seem to be associated with local activism-related clusters in longer answers, where cooperation, gathering, communication, local policies and activism seem to be considered as a good way to be involved in a post-COVID transformation.

7 Discussion

Our task was to cluster short and diversified text segments under extreme data sparsity conditions. Instead of trying to learn semantic similarities and patterns from scratch in order to discover the most salient topics, we use already-acquired "knowledge" (i.e. already trained language models) and transfer it to our task. Transfer learning to model topics is not new, especially in short text topic modelling (Dieng et al., 2019). However, because we deal with short segments of text where units of meaning (our segments) are easy to define thanks to the questionnaire design, we can analyze topics at the segment level rather than word level, allowing for more comprehensive models. Additionally, it requires a reasonable hardware configuration (32 Go of RAM, 16 CPU cores but no GPU), builds on open-source software, and, apart from word embeddings required for SIF, can be used with multilingual USE and BERT model and is thus adaptable to a large variety of languages.

Moreover, we argue that our semi-supervised framing of the topic modelling problem gives our model interesting interpretability and evaluation capabilities. Far from "reading tea leaves" (Chang et al., 2009), topic interpretation by directly reading sentences sampled from a thematic cluster is an easy process. Predictions, with their associated level of probability, also enables to be more critical over the topic distributions estimated by the model. Evaluation, made difficult by the unsupervised nature of topic modelling, is realized through the *sentence intrusion* test. This test, albeit requiring human annotation, gives overall and theme-wise metrics of discovered themes. Nevertheless, *sentence intrusion* has two drawbacks. First, annotating enough *sentence intrusion* sets to provide robust theme-wise results can be time-consuming. Second, *sentence intrusion*, as other topic modelling evaluation metrics (such as topic coherence) does not guarantee exhaustiveness of the retrieved topics.

Finally, from a more general natural language processing perspective, it should be noted that the evaluation of our sentence embedding pipeline on the SentEval benchmark obtained surprisingly good results. This shows that concatenating existing models and reducing dimension is enough (at least, in our case) to get near state-of-the-art results and, more interestingly, better results than any of the three embedding techniques used alone. This might indicate that the three embedding techniques encode different, and complementary sentence information.

8 Conclusion

In this work, we have presented a semi supervised pipeline for thematic clustering. This pipeline is specifically adapted to small datasets, that traditional machine learning algorithms would fail analyzing due to data sparsity and data heterogeneity (few mentions of the different topics with low redundancy in the vocabulary). It also requires a reasonable hardware configuration, making it highly portable to different research environments, while still benefiting from recent advances in the field. More generally, our pipeline is potentially useful for collections of short texts like tweets, comments and posts on social media, press article titles, etc.

References

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, page 49–60, New York, NY, USA. Association for Computing Machinery.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but though Baseline for Sentence Embeddings. *Iclr*, 15:416–424.

David M. Blei, Andrew Y. Ng, and Michael T. Jordan. 2002. Latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 3:993–1022.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*, pages 288–296.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 670–680.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. Topic Modeling in Embedding Spaces.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. (786):1–9.

Bruno Latour. 2020. Imaginer les gestes-barrières contre le retour à la production d'avant-crise.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. *34th International Conference on Machine Learning, ICML 2017*, 5:3721–3731.

Jipeng Qiang, Ping Chen, Tong Wang, and Xindong Wu. 2016. Topic Modeling over Short Texts by Incorporating Word Embeddings.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

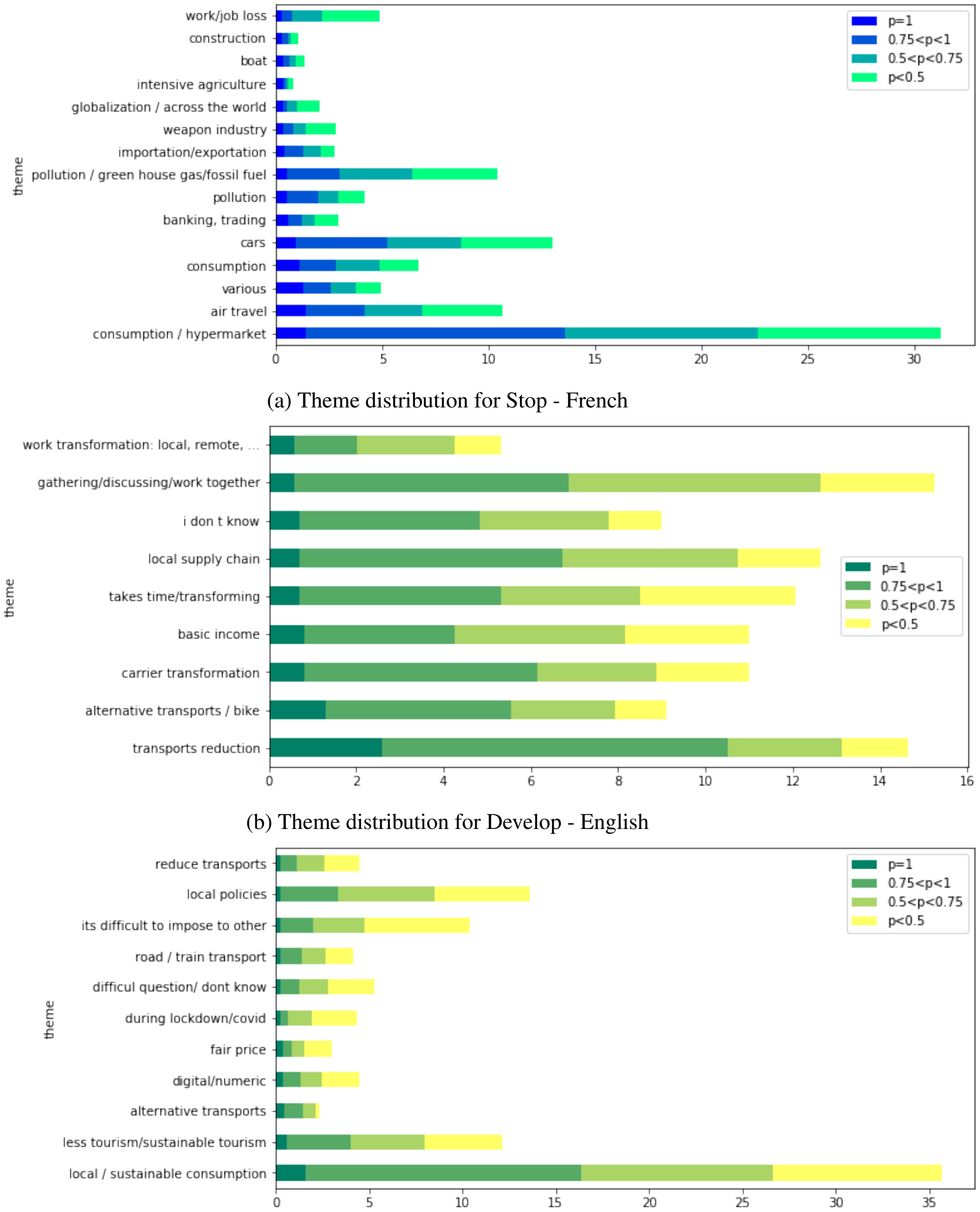
Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, volume 4. Harrahs and Harveys, Lake Tahoe.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. pages 1–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.

Bin Wang and C. C. Jay Kuo. 2020. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-based Word Models. 14(8):1–13.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural Topic Modeling with Bidirectional Adversarial Training.



(a) Theme distribution for Stop - French

(b) Theme distribution for Develop - English

(c) Theme distribution for Develop - French

Figure 3: Theme distribution for Stop and Develop answers in English and French. The graph indicates the distribution of the discovered themes among the English Stop answers dataset (percentages). These distributions are showed for different predicted probability thresholds. For instance, in 3a the segments predicted as *consumption/hypermarket*-related with $p > 0.75$ account for approximately 13% of the dataset. $p = 1$ means that the segment was present in the OPTICS cluster in the first place.

Theme	Accuracy
changing the system	1.0
commuting/work	1.0
consumption	1.0
consumption / shopping	1.0
corporations/business/banks	1.0
globalisation/capitalism/liberalism	0.3
nature destruction	1.0
productivism/industry	1.0
tourism	0.3
travel/air travel	1.0
work	1.0

(a) Thematic cluster reliability for Stop-En. $p > 0.75$

Theme	Accuracy
air travel	1.0
banking, trading	1.0
boat	0.67
cars	0.67
construction	1.0
consumption	0.67
consumption / hypermarket	0.3
globalization / across the world	1.0
importation/exportation	1.0
intensive agriculture	0.67
pollution	1.0
pollution / green house gas/fossil fuel	1.0
various	0.67
weapon industry	1.0
work/job loss	0.67

(b) Thematic cluster reliability for Stop-Fr. $p > 0.75$

Theme	Accuracy
alternative transports / bike	1.0
basic income	1.0
carrier transformation	0.67
gathering/discussing/work together	1.0
i don t know	1.0
local supply chain	0.67
takes time/transforming	1.0
transports reduction	1.0
work transformation: local, remote, ...	1.0

(c) Thematic cluster reliability for Develop-En. $p > 0.75$

Theme	Accuracy
alternative transports	0.67
difficul question/ dont know	1.0
digital/numeric	0.67
during lockdown/COVID	1.0
fair price	0.67
its difficult to impose to other	1.0
less tourism/sustainable tourism	0.67
local / sustainable consumption	0.3
local policies	0.67
reduce transports	1.0
road / train transport	1.0

(d) Thematic cluster reliability for Develop-Fr. $p > 0.75$

Table 7: Results of the sentence intruder evaluation task, cluster-by-cluster, for all corpora. In the sentence intruder task, pools of 5 segments are created, with four segments belonging to the cluster, and one segment drawn at random. A human annotator has to find the intruder among the 5 segments. Here, we annotated 3 pools of 5 segments per thematic cluster. An accuracy of 0.67 means for instance that 2 out of 3 intruders were correctly distinguished by the annotator. Probability thresholds are used in the reaffiliation process : when $p > 0.75$, we only keep segments where the theme has been predicted by the classifier with probability $p > 0.75$.