

Extracting Business Insights through Dynamic Topic Modeling and NER

Muhammad Arslan^a and Christophe Cruz^b

Laboratoire d'Informatique de Bourgogne (LIB), 9 Avenue Alain Savary, Dijon 21000, France

Keywords: Artificial Intelligence (AI), Business Intelligence (BI), Dynamic Topic Modeling, Natural Language Processing (NLP), Named-Entity Recognition (NER).

Abstract: Companies need the data from multiple sources for analysis and to find meaning in deriving valuable business insights. Online news articles are one of the main data sources that present up-to-date business news offered by various companies in the market. Topic modeling, i.e. a key player in the Natural Language Processing (NLP) domain, helps businesses to drive value from news articles. Also, it supports the extraction of business insights from news articles to facilitate the identification of trends (topics) in the market and their evolution over time. These insights can help businesses not only automate routine tasks but also in building new marketing policies, decision-making, and customer support. It is also important to find the linked semantic information (i.e. key persons, organizations, or regions called named entities) involved in generating these topics for the identification of news sources. This paper presents the application of a hybrid approach based on dynamic topic modeling and Named-Entity Recognition (NER) for extracting business trends along with the related entities. To show the functionality of the proposed approach, the news articles collected from the websites that published the content related to company interests were from 2017 to 2021 inclusive. The proposed approach can serve as the foundation for future exploratory trend analysis to study the evolution of information not only in the business domain but also applicable in other domains.


1 INTRODUCTION


Digital data in the form of news articles is growing at a fast rate and becoming an asset concerning the information it carries for business users (Mann, 2021). The information retrieval from news articles provides an overview of the main topics in them (Chung, 2014). Such information can help in organizing a large collection of news articles for companies in structuring crucial market information. It will also make searching for similar documents in the collection a lot faster and more efficient. For organizing news articles, it is important to get an overall theme and the underlying narrative in the text (Kavvadias et al., 2022).

The most common technique in NLP for identifying the theme of documents in real-time is topic modeling (Kavvadias et al., 2022; Sestino and Mauro, 2022). The traditional topic modeling techniques like Latent Dirichlet Allocation (LDA) do

not output semantically meaningful topics as they do not consider semantic relationships among words (Guo and Diab, 2011; Kim and Kim, 2022). Hence, they may fail to precisely represent text documents.

The main motivation of this research is to use the latest pre-trained dynamic topic model (Xie et al., 2022) and apply it to finding semantically meaningful topics to improve the classification of the information resulting from the dynamic (i.e. changing) world news. Once the topics are identified, the next step is to map with them the semantic information (Li and Roth, 2006). regarding entities that will help to uncover key persons, organizations, or regions involved in generating these topics for the identification of news sources. The semantic information is extracted using the NER, i.e. a process of identifying pre-defined named entities in text, which are persons, organizations, or regions in our case (Shelar et al., 2020). Eventually, visualizing the trend analysis of these topics over the years will help

^a  <https://orcid.org/0000-0003-3682-7002>

^b  <https://orcid.org/0000-0002-5611-9479>

to understand business insights that can be used for decision-making in companies.

Up to now, the existing literature has a shortage of comprehensive studies which attempt to use state-of-the-art topic modeling techniques in conjunction with the NER to understand the evolution of topics about the corresponding semantic information (i.e. key persons, organizations, or regions). The semantic information provides the principal actors responsible for the generation of topics in the business market. To cover this gap, this article focuses on the application of Bidirectional Encoder Representations from Transformers (BERT)-based topic modeling that allows us to train the model using our latest news articles dataset to find emerging semantic topics from the changing business world. In addition, the exploitation of the NER technique will help us to uncover semantic information regarding the dominant topics in the domain of the business industry.

This paper is organized as follows. Section 2 reviews the background on topic modeling and NER. Section 3 introduces the proposed method. The discussion is mentioned in Section 4 and the conclusion is described in Section 5.

2 BACKGROUND

Topic modeling methods refer to the techniques which uncovers the underlying hidden latent semantic representations in the given set of documents (Mann, 2021; Grootendorst, 2022). Topic models process text documents to extract underlying themes (topics) it covers, and how these topics are linked with each other (Mann, 2021). Here, a topic is a set of the most probable words in the cluster. Dynamic topic models (Grootendorst, 2022) are the type of topic models that can be used to analyse the evolution of topics over time using a set of documents and quantifies their trends in real-time. The application of these techniques is applied to pieces of texts, e.g. online news articles. Topic detection produces two types of output, such as; 1) cluster output, and 2) term output (Milioris, 2015). In the first method, referred to as a document pivot, a cluster of documents is used for representing a topic. Whereas in the latter, referred to as feature-pivot, a cluster of terms is produced (Milioris, 2015).

In the literature (Mann, 2021), there are several popular methods of topic detection that fall into either of the two categories. Some of them are; Latent Dirichlet Allocation (LDA), Graph-Based Feature-Pivot Topic Detection (GFeat-p), and Frequent

Pattern Mining (FPM). Srivastava and Sahami, (2009) discussed LDA for understanding the textual data using summarization, classification, clustering, and trend analysis using the textual data. Hall et al. (2008) discussed trend analysis using temporal data and generated visualizations using LDA topic models. In addition, Vayansky and Kumar (2020) clustered the documents using topic modeling and found the topics for each cluster. Schofield et al. (2017) discussed the impact of pre-processing on topic modeling applications. They found that the removal of stop words from the text corpus has little impact on the performance of the inference of topics. Also, the process of stemming potentially reduces the efficiency of the resulting topic model. Moreover, Guo and Diab (2011) mentioned that traditional topic models treat words as strings without considering predefined knowledge about word sense. They perform inference to extract topics by calculating word co-occurrence. Though, the co-occurred words may not be semantically related to topics.

To capture the semantics, there have been numerous attempts regarding the latest developments in the field of topic modeling using machine learning techniques. For instance, Deng et al. (2020) proposed a semi-supervised learning method by applying topic modeling and deep learning to establish a better understanding of the customer's voice using textual data. In addition, Sahrawat et al. (2020) used BERT embeddings for extracting key phrases from textual scholarly articles. Their approach used a BERT-based classification algorithm with a Probabilistic and Semantic Hybrid Topic Inference (PSHTI) model to automate the process of recognizing main topics from the data. Moreover, Grootendorst (2022) presented a topic model that extracts coherent topic representation using a class-based variation of Term Frequency-Inverse Document Frequency (TF-IDF). More precisely, it generates document embeddings using a pre-trained transformer-based language model, performs clustering on the embeddings, and consequently generates topic representations with the class-based TF-IDF mechanism. In literature, BERTopic remains competitive among a variety of existing classical topic modeling architectures as it generates coherent topics. For this reason, it was selected for this research.

Once the topic modeling is achieved, NER (Shelar et al., 2020) is used to identify named entities in text. In our case, we have chosen three named entities, which are persons, organizations, and geographical regions. There are different Java and Python-based libraries available for executing NER on text. Some of them are SpaCy, Apache OpenNLP, and

TensorFlow (Shelar et al., 2020). The TensorFlow and SpaCy are Python-based libraries. The parameters such as execution and prediction time, quality of identification, and accuracy are used to compare the performance of these NER libraries. Shelar et al. (2020) found that SpaCy gives better and more accurate results as compared to Apache OpenNLP and TensorFlow to identify named entities in the text. The prediction time for the SpaCy model is also less compared to other libraries. Hence, SpaCy is selected for the proposed system.

3 PROPOSED SYSTEM

To understand the dynamicity of business topics in the market over time, a proof-of-concept system is proposed. The developed system is based on topic modeling in conjunction with the semantic information extracted using the NER technique. The applied state-of-the-art architectures used to execute topic modeling and NER are BERTopic (Grootendorst, 2022) and SpaCy NER (Honnibal et al., 2020). To apply BERTopic to create business topics, a dataset of 26,509 news articles is used (see Figure 1 and 2). These articles were fetched from more than 312 different news sources using their original websites, such as actu.fr, ladepeche.fr, fusacq.com, etc. during the five years from 2017 to 2021. The majority of the text files retrieved using these websites were in French and English. The French text files were translated into English using a Python library named Deep Translator (i.e. <https://pypi.org/project/deep-translator/>). Later, basic pre-processing processes are executed for stop-word removal, filtering of emails, numbers, websites, and special characters, and lemmatization. The Natural Language Toolkit (NLTK) Python library (Bird, 2006) is used to perform the pre-processing. As a

result, a corpus of around 8,128,981 words was obtained to execute the topic modeling.

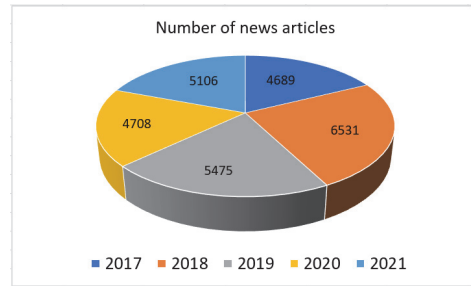


Figure 1: The number of news articles spans the last 5 years.

To apply BERTopic, we need to convert the pre-processed news articles to numerical data. For this, a python package i.e. “sentence-transformers” is used for capturing document-level embeddings by processing text data into 512-dimensional vectors (Grootendorst, 2022). For initiating BERTopic, we set the language to English since our corpus is in the English language. However, there is a possibility to use a multi-lingual model. Also, the calculation of topic probabilities is enabled while training the BERTopic model. Once the training of the model is completed, the model output 66 business topics, where each topic is represented by 10 most probable words as shown in Figure 3. Here, we can visualize the dominant topics based on their frequency over the last 5 years (see Figure 4). The aim of adding this figure here is to give an idea of the percentage of document from the corpus of news articles falling within a business topic. In addition, to analysis the semantic similarity of identified dominant topics with other topics, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm. It is used to cluster similar business topics together. In the literature, there are plenty of different

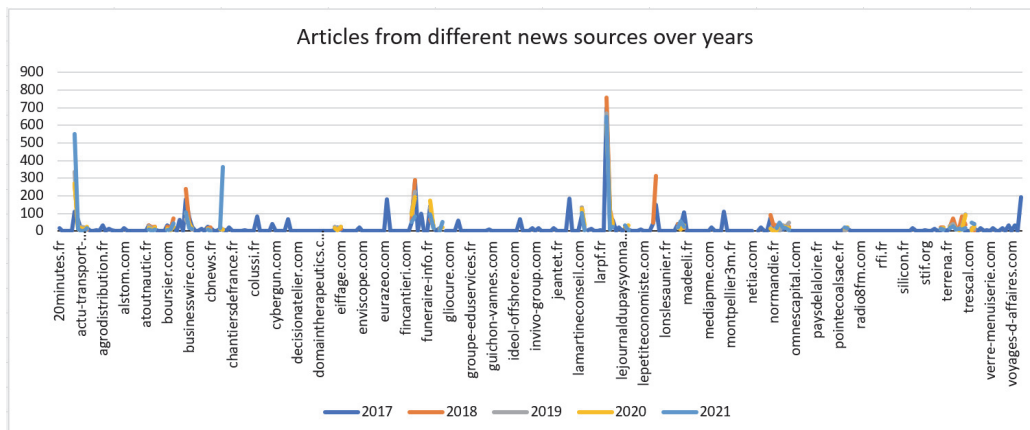


Figure 2: The number of articles from various news websites spans the last 5 years.

clustering algorithms available, but HDBSCAN is applied (see Figure 5) as it does not force less relevant data points into clusters (Grootendorst, 2022). This algorithm considers less relevant points as outliers. After the acquisition of business topics, we need to visualize how topics in news articles have evolved over time. For this, a function named “topic_model.topics_over_time (docs, topics, timestamps, global_tuning, evolution_tuning)” is used. Where, “docs” are the pre-processed news articles corpus, and “topics” are the topics that we have acquired before. “Timestamps” are the values of the timestamp of each news article. “Global_tuning”

is used for averaging the topic representation of a topic at a time ‘t’ with its global topic representation. “Evolution_tuning” is used for averaging the topic representation of a topic at time t with the topic representation of that topic at time t-1. Lastly, the parameter “nr_bins” specifies the number of bins to put our timestamps into. It is not recommended to extract the topics at thousands of different timestamps. Hence, it is suggested to keep this value below 20 (Grootendorst, 2022). After inputting the values to these parameters, Figure 6 is plotted to see the evolution of business topics over the last 5 years.

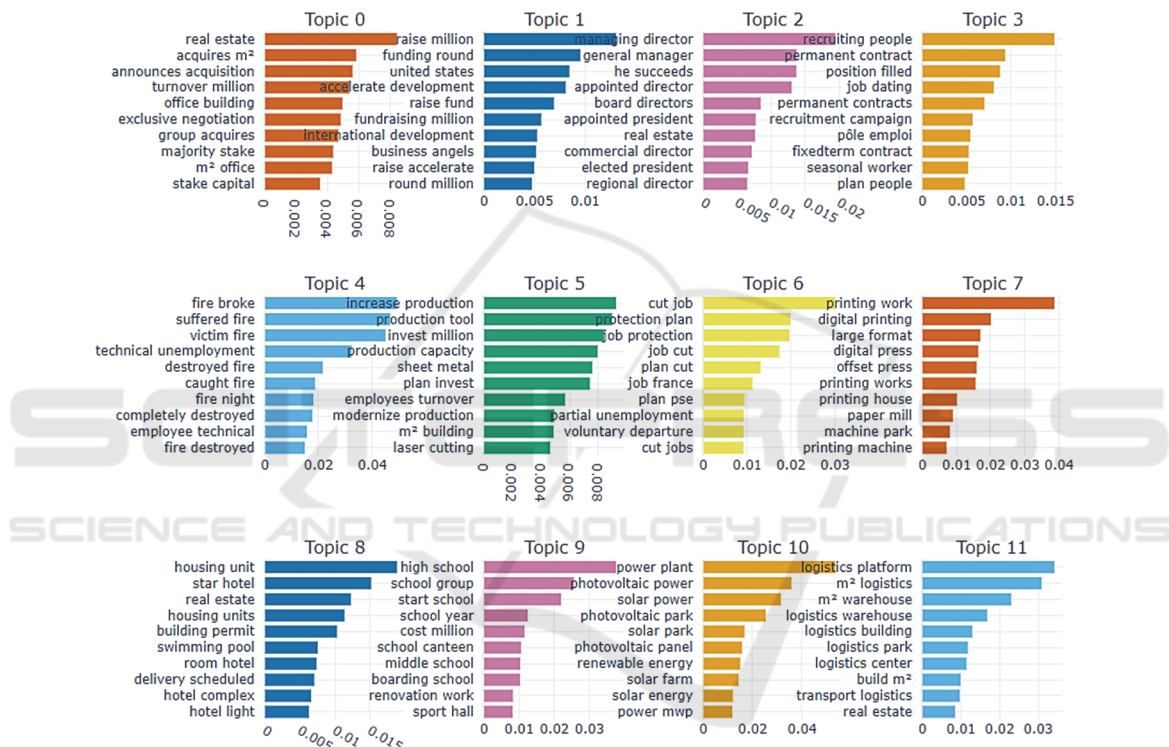


Figure 3: Extracted business topics.

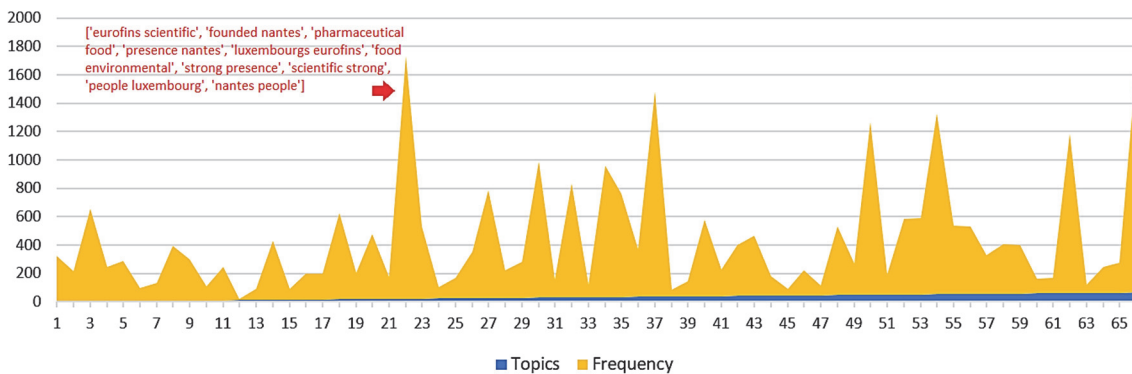


Figure 4: Frequency of topics over the last 5 years.

After identifying the business topics, the same text corpus is used for extracting named entities. For this, SpaCy Python library is used i.e., a free open-source library for NLP. The functionality of NER with SpaCy transformer (Roberta-base) English model is executed to identify key persons, organizations, and

geographical locations as shown in Figure 7. The process of NER is executed for each news article file presented in the corpus. Later, the identified entities are mapped with their corresponding business topics as extracted previously for analysis.

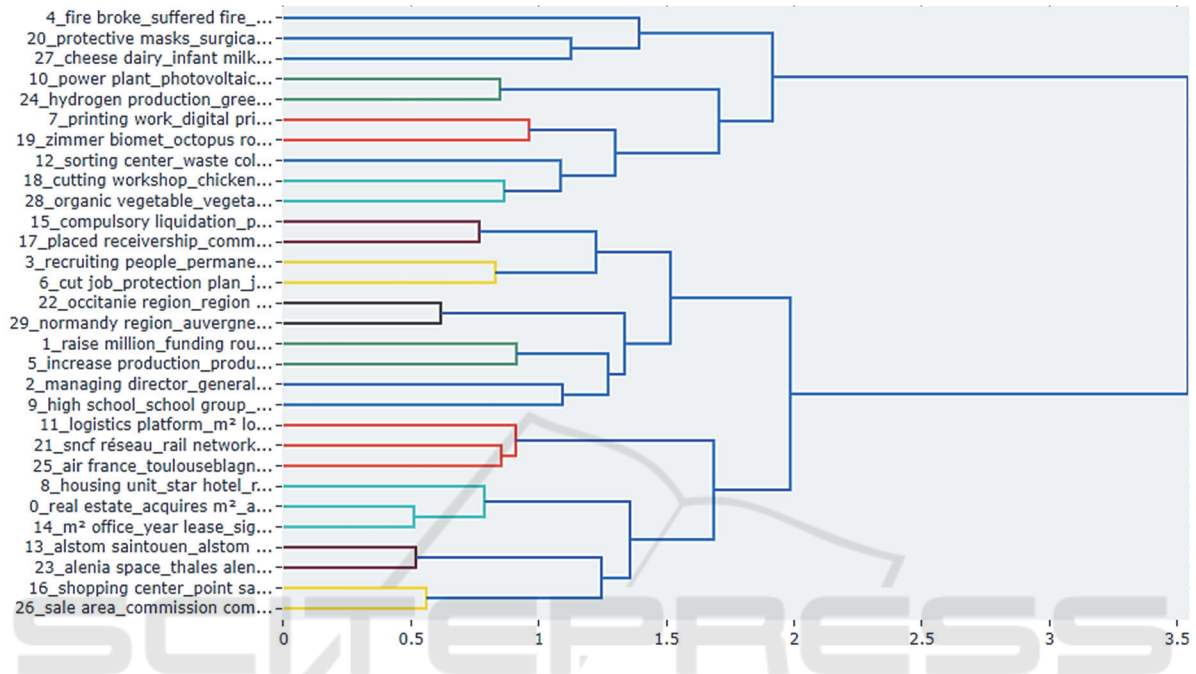


Figure 5: Hierarchical grouping of identified clusters having business topics.

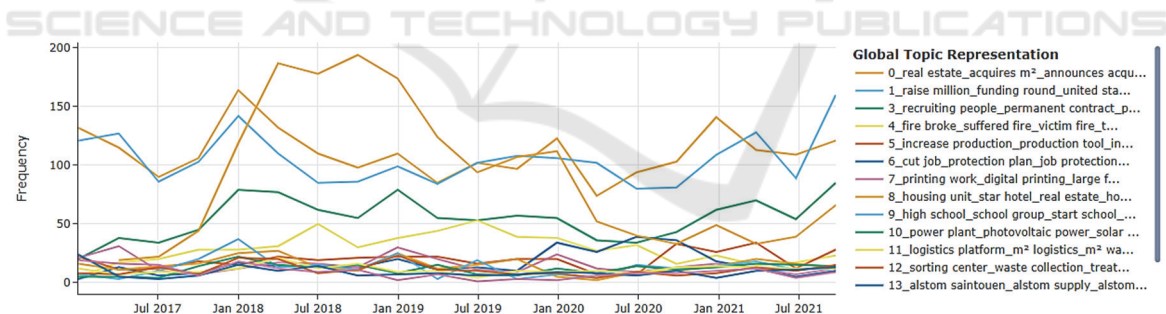


Figure 6: Evolution of business topics over time.

AJR Transports **ORG** wants to move from Brebières to Corbehem where a 50,000m2 building is going to be built, Based in Brebières **GPE** in the area of La Ventelle, the company AJR Transports **ORG** (25 employees), specializing in transport and logistics, plans to move to its second site, in Corbehem **GPE**, where a 50,000m2 building will be built.

Figure 7: Extracting named-entities in news articles for acquiring the semantic information.

4 DISCUSSION

After mapping the NER information with the topics, it will help us to study the evolution of different business topics with their associated semantic information to uncover the news sources in terms of key persons, organizations, and geographical regions involved in a specific topic. For instance, for the sake of demonstrating the type of information which can be inferred using the integrated data captured using dynamic topic modeling and NER is described using the formulation of the below-mentioned questions;

Q.1 What has been the evolution of a specific business topic across different geographical locations over the last years?

Q.2 Which business organizations were involved in a specific business topic over the last 5 years?

Q.3 Which people have been involved in a specific business topic over the last 5 years?

To find the answers to the above-mentioned questions, first, a business topic related to “recruiting people” is selected for the analysis.

Topic: [*'recruiting people', 'permanent contract', 'position filled', 'job dating', 'permanent contracts',*

'recruitment campaign', 'pole emploi', 'fixedterm contract', 'seasonal worker', 'plan people']

Then, the NER information linked to this topic is further analysed. Some of the information which can be inferred from this trend analysis is:

Bordeaux was the leading location for recruiting people in 2018 (see Figure 8). Whereas in 2021, it lost its position, and Paris is identified as the leading city for this topic. Another insight that can be noticed using this plot is that, in the world, France was in the leading position for recruiting people in 2018 and 2021 as compared to other countries. In terms of key organizations related to recruiting people, Renault, Finvens, and Verisure were the leading companies over the last 5 years (see Figure 9). Lastly, by analyzing Figure A, it can be observed that Bertrand Zimmer, Carlos Tavares, and Marianne Laigneau were the leading personalities who were quoted in the news articles concerning recruiting people (see Figure 10).

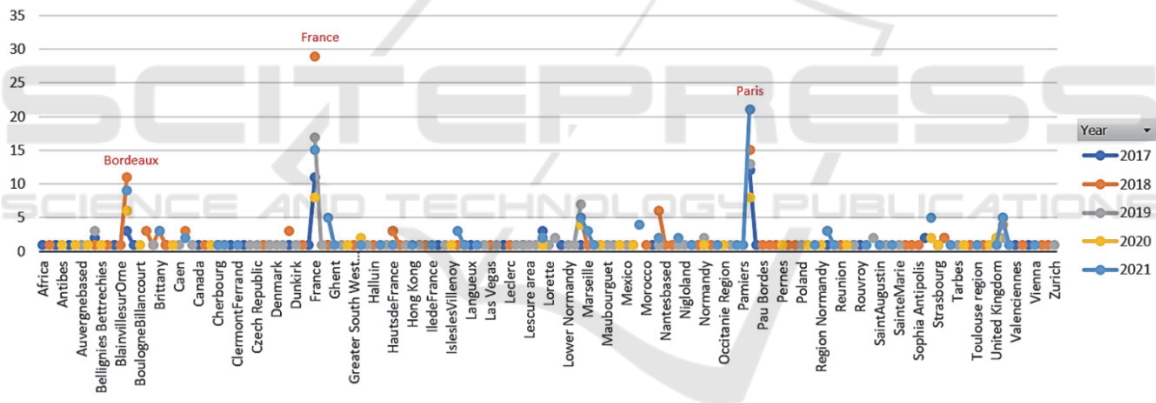


Figure 8: Identifying key regions related to a business topic over the years.

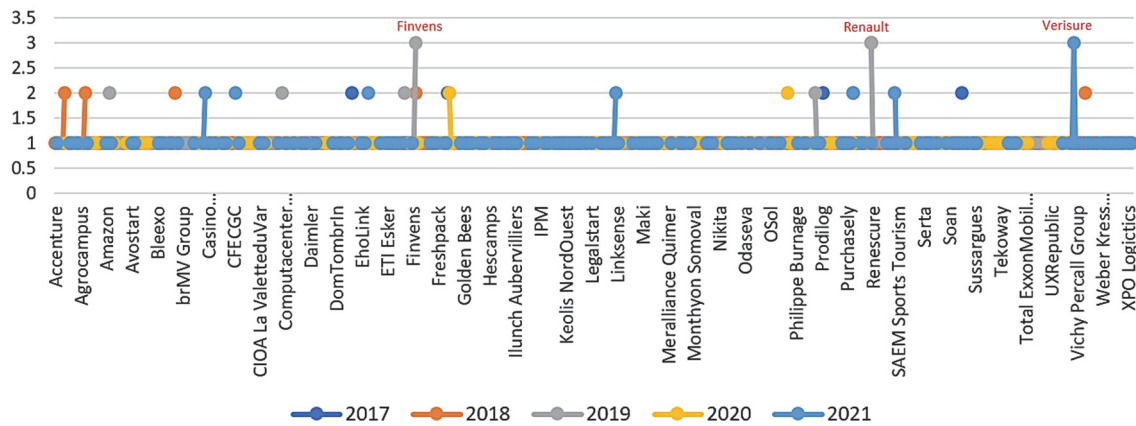


Figure 9: Identifying key organizations related to a business topic over the years.

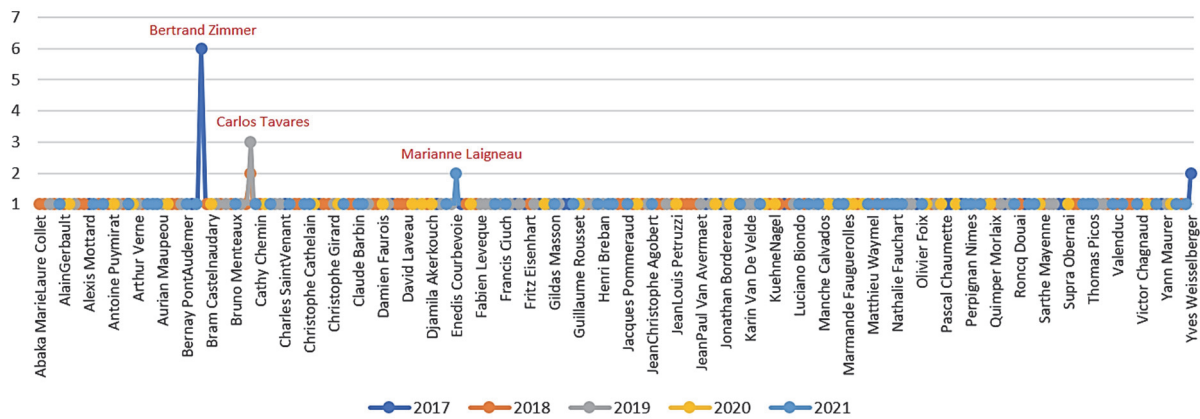


Figure 10: Identifying key persons related to a business topic over the years.

5 CONCLUSIONS

The paper presents a proof-of-concept system application that allows business analysts with minimum experience in programming to execute dynamic topic modeling and NER using built-in Python libraries for analyzing business trends of the market. The idea of comparing the acquired topics using up-to-date news articles leads to new business perspectives. There are several limitations to this research. First, the textual corpus of news articles that is exploited for the study was small. Also, there were limited publications about business in the news articles. Second, the identification of named entities such as key persons, organizations, or regions in text is not very accurate. During the data analysis, it was observed that the SpaCy NER detected a significant percentage of persons mentioned in the text as organizations. Whereas, there were also problems with the identification of geographical locations. Due to this issue, the system needs further evaluation and validation process to confirm its efficiency and quality of performance for business. For further studies, researchers can examine the application of the developed system in different domains other than business using enhanced NER techniques and with extended text corpus.

ACKNOWLEDGEMENTS

The authors thank the French company FirstECO (<https://www.firsteco.fr/>) for providing the taxonomy, the French government for the plan France Relance funding, and Cyril Nguyen Van for providing the technical advice.

REFERENCES

- Bird, S. (2006, July). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions (pp. 69-72).
- Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, 34(2), 272-284.
- Deng, X., Smith, R., & Quintin, G. (2020). Semi-supervised learning approach to discover enterprise user insights from feedback and support. *arXiv preprint arXiv:2007.09303*.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Guo, W., & Diab, M. (2011, July). Semantic topic models: Combining word distributional statistics and dictionary definitions. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 552-561).
- Hall, D., Jurafsky, D., & Manning, C. D. (2008, October). Studying the history of ideas using topic models. In Proceedings of the 2008 conference on empirical methods in natural language processing (pp. 363-371).
- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). SpaCy: Industrial-strength Natural Language Processing in Python. doi:"10.5281/zenodo.1212303
- Kavvadias, S., Drosatos, G., & Kaldoudi, E. (2020). Supporting topic modeling and trends analysis in biomedical literature. *Journal of Biomedical Informatics*, 110, 103574.
- Kim, M., & Kim, D. (2022). A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results. *Applied Sciences*, 12(6), 3118.
- Li, X., & Roth, D. (2006). Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3), 229-249.

- Mann, J. K. (2021). Semantic Topic Modeling and Trend Analysis. Master's thesis.
- Milioris, D. (2015). Trend Detection and Information Propagation in Dynamic Social Networks (Doctoral dissertation, Palaiseau, Ecole polytechnique).
- Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., ... & Zimmermann, R. (2020, April). Keyphrase extraction as sequence labeling using contextualized embeddings. In European Conference on Information Retrieval (pp. 328-335). Springer, Cham.
- Sestino, A., & De Mauro, A. (2022). Leveraging artificial intelligence in business: Implications, applications and methods. *Technology Analysis & Strategic Management*, 34(1), 16-29.
- Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding text pre-processing for latent Dirichlet allocation. In Proceedings of the 15th conference of the European chapter of the Association for Computational Linguistics (Vol. 2, pp. 432-436).
- Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named entity recognition approaches and their comparison for custom ner model. *Science & Technology Libraries*, 39(3), 324-337.
- Srivastava, A. N., & Sahami, M. (Eds.). (2009). Text mining: Classification, clustering, and applications. CRC press.
- Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582.
- Xie, Y., Ning, C., & Sun, L. (2022, January). The twenty-first century of structural engineering research: A topic modeling approach. In *Structures* (Vol. 35, pp. 577-590). Elsevier.

