World Patent Information 84 (2026) 102426

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Review

Clustering doc2vec output for topic-dimensionality reduction: A MITRE ATT&CK calibration☆,☆☆

Nathan Monnet [a], Loïc Maréchal [b],*

[a] Cyber-Defence Campus - armasuisse, Science and Technology, 1015 Lausanne, Switzerland
[b] Institute of Entrepreneurship & Management, HES-SO Valais-Wallis, 3960, Sierre, Switzerland

ARTICLE INFO

Dataset link: https://github.com/nmonnet

JEL classification:
C45
C58
G12.

Keywords:
Clustering
Machine learning
Natural language processing
Trends analysis
Cybersecurity

ABSTRACT

We introduce a novel approach to text classification by combining doc2vec embeddings with advanced clustering techniques to improve the analysis of specialized, high-dimensional textual data. We integrate unsupervised methods such as Louvain, K-means, and Spectral clustering with doc2vec to enhance the detection of semantic patterns across a large corpus. As a case study, we apply this methodology to cybersecurity risk analysis using the MITRE ATT&CK framework to structure and reduce the dimensionality of cyberattack tactics. Louvain clustering proved the most effective among the tested methods, achieving the best balance between cluster coherence and computational efficiency. Our approach identifies four "super tactics", demonstrating how clustering improves thematic coherence and risk attribution. The results validate the utility of combining doc2vec with clustering, particularly Louvain, for enhancing topic modelling and text classification.

Contents

1. Introduction 2
2. Related work 2
3. Data and methodology 3
   3.1. Cyberattacks knowledge database 3
   3.2. Textual and semantical training data 3
   3.3. Methodology 3
      3.3.1. Vector representation using doc2vec 3
      3.3.2. Cosine similarity 3
      3.3.3. Training of doc2vec 4
      3.3.4. Clustering methods 4
4. Results 5
   4.1. Clustering of MITRE ATT&CK 5
5. Conclusion 7
CRediT authorship contribution statement 8
Declaration of competing interest 10
Appendix 10
Data availability 10
References 10

☆ This article is part of a Special issue entitled: 'GTM 2025's Special Issue' published in World Patent Information.

☆☆ This document results from a research project funded by the Cyber-Defence Campus, armasuisse Science and Technology, Switzerland. We appreciate helpful comments from seminar participants at the Cyber Alp Retreat 2024.

* Corresponding author.
E-mail addresses: nathan.monnet@alumni.epfl.ch (N. Monnet), loic.marechal@hevs.ch (L. Maréchal).

https://doi.org/10.1016/j.wpi.2026.102426
Received 1 July 2025; Received in revised form 4 November 2025; Accepted 5 January 2026
Available online 14 January 2026

1. Introduction

Natural Language Processing (NLP) has become essential for analysing large volumes of textual data, particularly in domains such as finance, where identifying and interpreting risk-related information embedded in corporate documents is crucial. Over the years, text classification techniques have evolved significantly from traditional dictionary-based approaches to more advanced vector-based models. Dictionary-based methods, once prevalent, rely on predefined lists of words or n-grams to classify texts according to specific themes or sentiments (*e.g.*, positive or negative sentiment, or risk-related content).

In this paper, we make three main contributions in this domain. First, we provide a systematic comparison of classical and graph-based clustering methods (K-means, spectral clustering, and Louvain) applied to cybersecurity text data. Second, we propose a novel aggregation of the fourteen MITRE ATT&CK tactics into four coherent "super tactics," improving interpretability and reducing dimensionality. Third, we establish a methodological bridge between ATT&CK and financial disclosures (10-K filings), enabling risk attribution and potential financial applications.

While modern transformer-based embeddings such as BERT [1], SciBERT [2], RoBERTa [3], and BERTopic [4] have become state-of-the-art for semantic modelling and topic reduction, we deliberately retain doc2vec. This choice ensures computational efficiency across millions of paragraphs, transparent interpretability via cosine similarity, and methodological continuity with previous work [5], which demonstrated its validity for risk-related text analysis.

A key innovation in this approach is the use of clustering techniques to enhance the doc2vec model's ability to detect patterns across a sizable textual corpus. Inspired by previous studies on topic modelling and clustering in NLP (*e.g.*, [6,7]), we employ unsupervised clustering methods to group similar paragraphs, improving the identification of cybersecurity-related content. Specifically, integrating the Louvain clustering method, as outlined by Blondel et al. [8], allows us to partition paragraphs into clusters that reflect distinct aspects of cyber risk. This clustering technique is applied to cybersecurity descriptions from MITRE ATT&CK, enabling a more granular analysis of the text's semantic structure.

Our results reveal four critical super tactics: Preparation and Reconnaissance, Persistence and Evasion, Credential Movement, and Command and Data Manipulation. The Preparation and Reconnaissance super tactic encompasses tactics related to information gathering and preparatory actions before an attack. At the same time, Persistence and Evasion highlight techniques adversaries use to maintain access to a compromised network and avoid detection. The Credential Movement super tactic focuses on the theft and use of credentials for lateral movement within a network. Command and Data Manipulation involves controlling and manipulating compromised systems to achieve the attacker's objectives.

Our comprehensive clustering analysis uses K-means, Louvain, and Spectral clustering to derive insights from a cosine similarity matrix constructed from MITRE ATT&CK vectors. While the K-means method provides a foundational structure, it lacks exclusivity among super tactics, resulting in low balanced scores and high entropy. In contrast, the Louvain method improves cluster heterogeneity, particularly for tactics 5 (Resource Development) and 12 (Reconnaissance), while addressing overly similar paragraph structures by applying similarity thresholds. Spectral clustering is a robust alternative that produces results comparable to those of Louvain when hyperparameters are finely tuned. Ultimately, we select the Louvain method's output as our baseline for grouping tactics, balancing the need for coherent clusters with the requirement for minimal hyperparameter tuning. This comprehensive analysis contributes valuable insights into structuring textual data for practical application and contributes to the expanding literature on NLP-based analysis and clustering to reduce topic dimensionalities. The remainder of the paper is structured as follows. Section 2 reviews related work; Section 3 details the data and our methodology; Section 4 presents the results; and Section 5 concludes.

2. Related work

Recent advances in Natural Language Processing (NLP) have substantially improved how semantic information is captured and represented in text. Building upon earlier vector-based embeddings such as word2vec and doc2vec, transformer-based encoders—most prominently BERT [1], SciBERT [2], and RoBERTa [3]—represent a major methodological leap. Unlike static embeddings that assign a single vector to each word or document, transformer models dynamically contextualize meaning by accounting for surrounding words in both directions, allowing them to capture fine-grained semantic nuances and domain-specific terminology. These models have become the foundation for many state-of-the-art applications in text classification, topic modelling, and document clustering. In this context, BERTopic [4] combines transformer embeddings with a class-based TF–IDF representation and density-based clustering to produce interpretable topics. It has been successfully applied in large-scale scientific and bibliometric analyses (*e.g.*, [9–11]), illustrating how contextual embeddings can enhance the interpretability and scalability of text-based research.

To situate these developments within a broader historical perspective, it is useful to recall how earlier approaches to textual analysis emerged in the financial and economic literature. Given the structured availability of corporate filings and financial disclosures, researchers in finance were among the first to exploit textual data systematically. Pioneering studies, such as those by Antweiler and Frank [12], Garcia [13], Jegadeesh and Wu [14] and Arslan-Ayaydin et al. [15], employ dictionary approaches to analyse the impact of sentiment in financial news and stock forums, demonstrating their effectiveness in specific contexts. However, these methods are inherently limited, relying on static, human-curated dictionaries that may overlook essential nuances, especially in specialized fields such as cybersecurity. Early dictionary-based approaches (*e.g.*, [12,14]) offered interpretable but static methods for text classification. Vector-based embeddings, such as word2vec [16] and doc2vec [17], have improved representation quality by capturing semantic similarity beyond word frequencies.

Recent advancements in NLP, primarily utilizing word embeddings and vector-based models such as word2vec and doc2vec, have addressed many of these limitations. By embedding words or entire paragraphs into a vector space, these models capture semantic relationships between words, going beyond simple word frequency counts to consider word context and order. One of the critical advantages of vector-based models, such as those proposed by Mikolov et al. [16] and later extended in the paragraph-to-vector model by Le and Mikolov [17], is their ability to understand the semantic similarity between paragraphs even when synonymous or domain-specific terms are used. This feature makes them particularly well-suited for specialized fields, such as cybersecurity, where language is highly technical and rapidly evolving. Applications of word embedding methods are beneficial for the automated processing of financial documents. For instance, Sautner et al. [18] and Hassan et al. [7] explore text classification to measure climate and political risks, respectively. Similarly, Adosoglou et al. [19], inspired by the findings of Cohen et al. [20], use doc2vec to compare subsequent annual SEC filings from listed US firms and find support for the stylized fact that similar documents predict positive expected returns. Finally, and close to our extension, Celeny and Maréchal [5], and Maréchal and Monnet [21] use doc2vec to extract the cyber risk-related context of annual SEC filings and test whether the cyber risk exposure of firms commands positive expected returns.

Clustering methods, such as K-means, spectral, and Louvain (see [8]) remain widely used for structuring high-dimensional text data. In finance and policy analysis, similar combinations of embeddings and clustering have been applied to quantify risks such as political or environmental exposure. However, in cybersecurity research, applications remain limited. Our work contributes to this literature by systematically comparing clustering approaches and introducing a coherent "supertactic" aggregation of the MITRE ATT&CK framework.

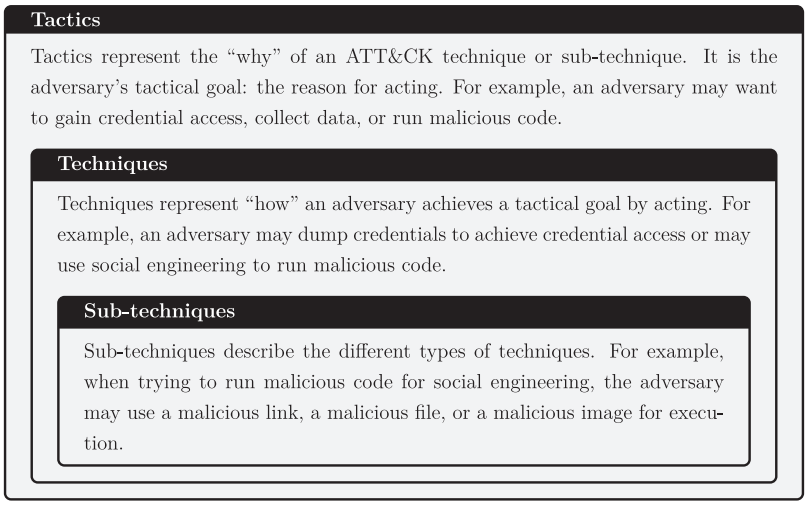

Fig. 1. Structure of MITRE ATT&CK.

Table 1
Examples of sub-techniques from MITRE ATT&CK.

		Description
Tactic	Credential Access	Adversaries may forge web cookies that can be used to gain access to web applications or Internet services. Web applications and services (hosted in cloud SaaS environments or on-premise servers) often use session cookies to authenticate and authorize user access.
Technique	Forge web credentials	
Sub-technique	Web cookies	
Tactic	Reconnaissance	Adversaries may gather employee names that can be used during targeting. Employee names can be used to derive email addresses and help guide other reconnaissance efforts and/or craft more believable lures.
Technique	Gather victim identity information	
Sub-technique	Employee names	

3. Data and methodology

For clarity, we define our terminology as follows:
(i) *tactic* refers to the 14 top-level MITRE ATT&CK categories,
(ii) *technique refers to the 245 lower-level MITRE ATT&CK categories*,
(iii) *sub-technique* refers to the 785 atomic descriptions under each tactic,
(iv) *sub-cluster* refers to the original ATT&CK groupings in the context of our clustering procedure, and
(iv) *supertactic* refers to our aggregated clusters.

3.1. Cyberattacks knowledge database

The MITRE ATT&CK[1] cybersecurity knowledge base is used as a reference for cyber attack descriptions. This knowledge base was established in 2013 to document cyberattack tactics, techniques, and procedures. It is structured by tactics, techniques, and sub-techniques as depicted in Fig. 1. There are 14 tactics: reconnaissance, resource development, initial access, execution, persistence, privilege escalation, defence evasion, credential access, discovery, lateral movement, collection, command and control, exfiltration, and impact. There are 785 sub-techniques across all tactics. Two examples of sub-techniques are given in Table 1.

3.2. Textual and semantical training data

The doc2vec method used in this paper was initially developed to identify cybersecurity-related paragraphs in the financial statements (10-K reports) of US-listed companies. We will briefly explain what a 10-K report is to clearly describe how the doc2vec model was initially designed and trained. 10-K statements are financial filings that publicly traded companies submit annually to the U.S. Securities and Exchange Commission (SEC). They contain textual information such as companies' financial statements, risk factors, and executive compensation. The SEC's Edgar archives index files were used to download and structure the 10-K reports. These files contain information about all the documents filed by firms for specific dates. From these files, a total of 64,988 10-K statements were identified.

3.3. Methodology

3.3.1. Vector representation using doc2vec

We use the paragraph-to-vector model proposed by Le and Mikolov [17], which is an extension of the word2vec model [16]. There are various advantages to working with this NLP approach compared to others, such as the dictionary approach. First, the comprehension of the method is semantical, meaning that it is not limited to a count of word frequencies. The word order affects the resulting vector, and paragraphs with similar or synonymous words will have closely related vector representations. Second, training the model with specific text that involves a particular vocabulary allows the incorporation of relatively unknown words. Finally, the resulting vectors have a dimension usually much smaller than the vectors resulting from the dictionary approach.

Two versions of the model exist: the distributed memory model (DM) and the distributed bag-of-words model (DBOW). In the DM, a neural network is trained as follows. First, a word is removed from a paragraph. Then, by inputting the paragraph vector representation and the context words (also in vector representation) surrounding the missing word, the neural network is optimized to guess the missing word. In the DBOW, the neural network is trained to predict a series of words sampled from a paragraph using only the vector representation of the paragraph as input. Fig. 2 illustrates the training process of the two models.

3.3.2. Cosine similarity

Using the doc2vec method, all paragraphs of interest can be embedded into vectors. A common way to attribute a similarity score to two

[1] https://attack.mitre.org/.

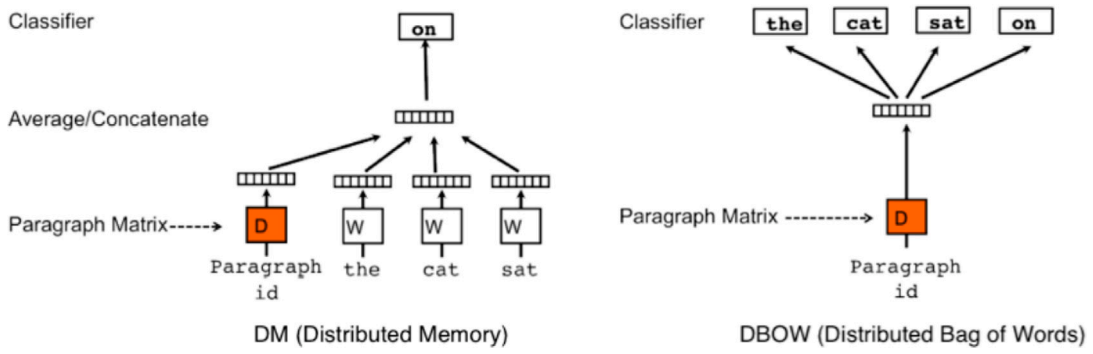

Fig. 2. Illustration of doc2vec training.
Illustration of the training of the neural network of the two versions of doc2vec, distributed memory model (DM) and distributed bag-of-words model (DBOW). *Source:* The figure is taken from [17].

Table 2
doc2vec parameters.

Method	Training size	Vector size	Window size	Min count	Sub-sampling	Negative sampling	Epoch
DBOW	1.7M	200	15	5	10^{-5}	5	50

Parameters of the chosen doc2vec model. DBOW stands for distributed bag-of-words.

paragraph vectors is to take the cosine of the angle they form. Other methods exist, but only measuring the angle has been proven more effective than considering the magnitude of the vectors (see [19]). This is because the latter is more susceptible to the random initialization of weights during training in the neural network that generates the vectors.

3.3.3. Training of doc2vec

The detailed training procedure of the doc2vec model follows Celeny and Maréchal [5] who provide full hyperparameter settings and validation. In this study, we reuse their trained model to ensure consistency with prior work, allowing our analysis to focus on the clustering and aggregation phases (also see [22]). In Table 2, we present the retained hyperparameters of the model and in Table 3, we present 11 top-scoring paragraphs from the doc2vec validation sample.

The doc2vec model was trained on a mixed corpus consisting of both MITRE ATT&CK sub-technique descriptions and corporate 10-K filings. This design choice ensures that the learned embeddings capture linguistic patterns relevant to both technical cybersecurity terminology and the broader language of corporate disclosures, which constitute the ultimate application domain of this framework. Incorporating 10-K text enhances the model's ability to generalize across contexts, allowing it to recognize cyber-related semantics even when expressed in non-technical business language. At the same time, the ATT&CK corpus anchors the embedding space in specialized cybersecurity concepts. This hybrid training strategy, therefore, provides a balanced representation that supports both methodological validity and cross-domain interpretability.

3.3.4. Clustering methods

A natural way of splitting the cybersecurity textual database into different categories comes from the written structure of MITRE ATT&CK, with 14 categories already established as tactics. In [21], the similarity of each tactic to 10-K paragraphs yield 14 distinct cyber-scores, which were too high to retain their explanatory power. Therefore, they argue for aggregating the 14 tactics into a few super-tactics. Following a similar methodology, we cluster textual data derived from the 785 MITRE ATT&CK sub-techniques.

Using the doc2vec method and cosine similarity, we transform each paragraph into a vector and compute a 785 × 785 similarity matrix. This matrix can be understood as a weighted network, where nodes represent paragraphs and edges correspond to pairwise similarities. On this matrix, several classical clustering methods can be applied.

We briefly tested two standard clustering approaches: K-Means (spherical, using cosine distances) and Spectral clustering (on a dimensionally reduced Laplacian matrix). Both methods allow partitioning the data into a pre-specified number of clusters, but they have limitations: K-Means depends on an initial choice of cluster number and initial centroids, while Spectral clustering requires careful dimensionality reduction. In our experiments, neither method produced a clustering that balanced sub-cluster homogeneity and overall cluster distribution as effectively as Louvain.

In the Louvain method, the cosine similarity matrix is converted into a weighted graph where each sub-technique is a node and edge weights correspond to cosine similarities. The resolution parameter of the Louvain algorithm, which controls community granularity, was tested across several values and found not to significantly affect the cluster structure. We report results with the default setting. To minimize noise from remarkably similar or dissimilar paragraphs, we impose lower and upper similarity thresholds of 0.25 and 0.85, respectively. These values were selected after examining the overall distribution of cosine similarities across the corpus. Pairs of documents with similarity scores below 0.25 typically exhibit negligible semantic overlap, while scores above 0.85 correspond to near-duplicate textual content. Setting these thresholds therefore filters out both noise and redundancy, ensuring that clustering reflects meaningful semantic relationships rather than trivial textual similarities. This choice provides a balanced trade-off between inclusiveness and interpretability and is consistent with practices commonly adopted in embedding-based text clustering. Finally, Appendix, Figs. A.1, A.2, and A.3 present a sensitivity analysis demonstrating that the entropy and balanced-score metrics remain stable across a range of initialization values.

The method relies on the concept of modularity, a measure of the density of links within communities relative to links between communities:

$$Q = \frac{1}{2m}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[S_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j),$$

Table 3
Top scoring paragraphs from the doc2vec validation sample.
Source: The table is taken from [5].

Score	Preprocessed paragraph	Firm ID (Ticker)	Tactic
0.593	Currently available internet browsers allow users modify browser settings remove cookies prevent cookies stored hard drives however third persons able penetrate network security gain access otherwise misappropriate users personal information subject liability include claims misuses personal information unauthorized marketing purposes unauthorized use credit cards	VSTY	Defence Evasion
0.590	Network security data recovery measures may adequate protect computer viruses break ins similar disruptions unauthorized tampering computer systems theft sabotage type security breach respect proprietary confidential information electronically stored including research clinical data material adverse impact business operating results financial condition	LXRX	Collection
0.583	Domain names derive value individual ability remember names therefore assurance domain name lose value example users begin rely mechanisms domain names access online resources government regulation internet regulation increasing number laws regulations pertaining internet	VSTY	Credential Access
0.577	Perceived actual unauthorized disclosure information collect breach security harm business factors beyond control cause interruptions operations may adversely affect reputation marketplace business financial condition results operations timely development implementation continuous uninterrupted performance hardware network applications internet systems including may provided third parties important facets delivery products services customers	MDAS	Credential Access
0.571	Unauthorized parties may attempt copy aspects products obtain use information regard proprietary others may independently develop otherwise acquire similar competing technologies methods design around patents cases rely trade secret laws confidentiality agreements protect confidential proprietary information processes technology	CSCD	Collection
0.570	Possible cookies may become subject laws limiting prohibiting use term cookies refers information keyed specific server file pathway directory location stored user hard drive possibly without user knowledge used among things track demographic information target advertising	VSTY	Discovery
0.561	Cannot certain advances computer capabilities discoveries field cryptography developments result compromise breach algorithms use protect content transactions website proprietary information databases anyone able circumvent security measures misappropriate proprietary confidential customer company information cause interruptions operations	VSTY	Impact
0.558	Ordering delivery customers ready place order proceed shopping cart function directly checkout page orders placed online website via toll free telephone number customer service agents available take orders customers access internet uncomfortable placing order online	VSTY	Credential Access
0.557	Process allows identify catalogue embryonic stem cell clone dna sequence trapped gene select embryonic stem cell clones dna sequence generation knockout mice used gene trapping technology automated process create omnibank library frozen gene knockout embryonic stem cell clones identified dna sequence relational database	LXRX	Persistence
0.556	Believe systematic biology driven approach technology platform makes possible provide substantial advantages alternative approaches drug target discovery particular believe comprehensive nature approach allows uncover potential drug targets within context mammalian physiology might missed narrowly focused efforts	LXRX	Discovery
0.554	Concerns security internet may reduce use website impede growth significant barrier confidential communications internet need security rely ssl encryption technology designed prevent customer credit card data transaction process current credit card practices merchant liable fraudulent credit card transactions case transactions process merchant obtain cardholder signature	VSTY	Credential Access

The paragraphs are shown after preprocessing. Tactic refers to the MITRE ATT&CK tactic to which the paragraph is most similar, as measured by cosine similarity.

where S_{ij} is the edge weight between nodes i and j, k_i the sum of edge weights for node i, N the total number of nodes, c_i the community assignment of node i, and δ the Kronecker delta. The Louvain algorithm iteratively optimizes modularity in two phases: (i) nodes are moved between communities to maximize local modularity; (ii) communities are aggregated into single nodes to form a new network, repeating the process until convergence.

We evaluate cluster quality using two scores. First, Shannon entropy measures sub-cluster heterogeneity:

$$H_{sub_j} = -\sum_{i=1}^{\text{clusters}} P(sub_j)_i \log P(sub_j)_i,$$

where $P(sub_j)_i$ is the proportion of paragraphs from sub-cluster j in cluster i. Low entropy indicates that sub-clusters remain coherent within clusters. Second, a Balanced score ensures clusters are not disproportionately sized, defined as the standard deviation of cluster label counts. The optimal clustering minimizes both entropy and imbalance.

The final classification assigns each paragraph to the cluster where most of its sub-cluster members reside, preserving the original MITRE ATT&CK structure while leveraging semantic similarities to refine super-tactics.

4. Results

4.1. Clustering of MITRE ATT&CK

We apply clustering methods on the cosine similarity matrix created from MITRE ATT&CK paragraphs' vector embeddings. This allows for identifying the relevant paragraphs tied to previously mentioned super tactics (command and data manipulation, credential movement, persistence and evasion, and preparation and reconnaissance). We report the results of various attempts with different clustering methods in Figs. 3, 4, and 5. The K-means methods provide a coherent but crude initial structure, which we report in Fig. 3. Indeed, the paragraphs tend to be well spread across the super tactics (clusters), but at the cost of heterogeneity, with the exclusivity of a tactic in a super tactic being nonexistent. This results in a low balanced score at the cost of entropy, as depicted in Fig. 6. Fig. 4 shows the performance of the Louvain method. This method dramatically improves heterogeneity,

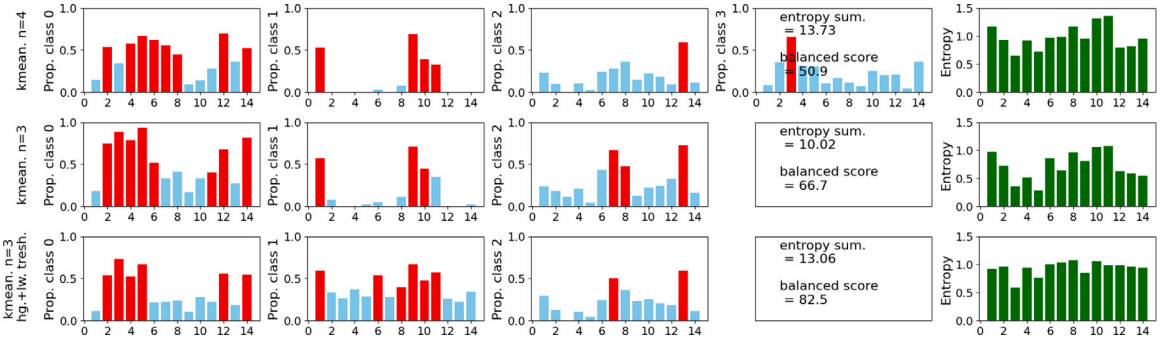

Fig. 3. Clustering results part.1.
This figure presents the results of each clustering method indicated on the left. The figure in red and blue represents $P(sub_j)_i$, the proportion of paragraphs of sub-cluster (tactic) j belonging to cluster (super tactic) i. The 14 sub-cluster labels are on the x-axis of each figure, and the cluster labels correspond to the columns (class 0 to 3, here). If the proportion is in red, it means it is the highest in the cluster (in other clusters/columns, the same sub-cluster will be in blue). We also report the entropy sum and the balanced score for each method in the figure. Finally, the individual Shannon entropy of each sub-cluster is reported in green in the last column. In the name of the method, we also indicate the hyperparameters of the method. Here, n corresponds to the number of clusters imposed by the k-means method. "hg. tresh." and "lw. tresh." corresponds to a change applied to the similarity matrix. If the value in the similarity matrix is lower than 0.25, it is set to 0 (lower threshold); if the similarity is higher than 0.85, it is set to 0.5 (higher threshold). In parts two and three, "egn" corresponds to the K eigenvectors in the spectral clustering. We also ensured that the output clusters of each method matched. Hence, the comparison is simpler (otherwise, what the Louvain method called cluster 2 is not necessarily cluster 2 for the k-means method). The following list shows the corresponding number of each tactic : 1: Persistence, 2: Command and Control, 3: Impact, 4: Initial Access, 5: Resource Development, 6: Collection, 7: Exfiltration, 8: Credential Access, 9: Privilege Escalation, 10: Execution, 11: Defence Evasion, 12: Reconnaissance, 13: Lateral Movement, 14: Discovery. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

especially with tactics 5 and 12 (resource development and reconnaissance), which are exclusive to cluster 1 (the super tactic: preparation and reconnaissance). However, not putting a threshold on the input similarity matrix component induces the Louvain method to create two superfluous clusters. Hence, we include those restrictions. Indeed, when comparing two paragraphs of MITRE ATT&CK, it is not uncommon to encounter sentences with similar structures for different semantic content. Thus, we tone down the similarity of a paragraph that is too similar with a higher threshold. Conversely, we define a lower threshold such that similarities that are too low, and therefore most likely to be noise reflecting no similarity, are set to zero.

The last method can be seen as a safeguard for the output of the Louvain method. Applying the spectral clustering method, we retrieve the structure previously encountered with higher heterogeneity than with K-means. If the hyperparameters are correctly tuned, the output is similar to the Louvain method, particularly for $n = 4$ and $egn = 6$. We report the results in Fig. 5. Including more dimensions (higher egn value) adds noise and decreases the clustering quality.

Finally, we select the output of the Louvain method as a baseline to group the tactics without splitting them across super tactics. Although Fig. 6 shows that outputs of other methods may be slightly better, we favour the Louvain method since no additional hyperparameter tuning is required.

As a robustness check, we retrained the doc2vec embeddings using only the MITRE ATT&CK corpus, excluding 10-K filings from the training set. The resulting clusters were consistent with the main model, confirming that the inclusion of 10-K data for calibration did not bias the ATT&CK-only clustering outcome. This indicates that the semantic structure learned from corporate disclosures does not distort the underlying relationships among ATT&CK sub-techniques. In addition, we present a sensitivity analysis of the similarity thresholds (lower bound = 0.25, upper bound = 0.85), as shown in the Appendix, Figs. A.1, A.2, and A.3. Entropy and balanced-score metrics remain stable under moderate variations in the threshold, supporting the robustness and reliability of the clustering framework (also see Fig. 7).

This yields the following super tactics, named after their content:

Preparation and Reconnaissance: This super tactic encompasses adversaries' tactics to prepare and gather information before launching an attack. **Impact** involves actions that disrupt, destroy, or manipulate systems and data to achieve the attacker's objectives. **Initial Access** includes techniques adversaries use to gain an initial foothold within a network, such as exploiting vulnerabilities or spear phishing. **Resource Development** entails the acquisition of resources like infrastructure, tools, and credentials necessary to support operations. **Reconnaissance** involves gathering information about the target environment to identify potential entry points and vulnerabilities. **Discovery** refers to techniques used to explore and map the target environment, such as network scanning and enumeration.

Persistence and Evasion: Once inside a target network, adversaries employ these tactics to maintain their foothold and avoid detection. **Persistence** ensures the attacker can maintain access even if the system is rebooted or credentials are changed by installing malware or creating rogue accounts. **Privilege Escalation** involves gaining higher-level permissions to access more sensitive information and critical systems. **Execution** refers to running malicious code on a victim system, often necessary to carry out the attacker's objectives. **Defence Evasion** includes a variety of methods to avoid detection and thwart defencive measures, such as disabling security software, obfuscating code, or using fileless malware.

Credential Movement: This group focuses on techniques to steal and use credentials to move within a network. **Credential Access** involves obtaining account names, passwords, and other secrets that allow attackers to authenticate themselves as legitimate users. Techniques include keylogging, credential dumping, and brute force attacks. **Lateral Movement** is moving through a network to find and access additional targets or more valuable data. This can be done using remote services, exploiting trust relationships, or leveraging legitimate credentials to access other systems and resources.

Command and Data Manipulation: In this phase, adversaries control compromised systems and manipulate data to achieve their goals. **Command and Control** involves establishing a communication channel with the compromised environment to issue commands and control malware. This can be achieved through web traffic, DNS, or custom communication protocols with command servers. **Collection** refers to gathering sensitive information from compromised systems, such as capturing screenshots, logging keystrokes, or accessing stored files. **Exfiltration** involves transferring the collected data from the target network to an external location controlled by the adversary, often using encrypted channels or covert methods to avoid detection.

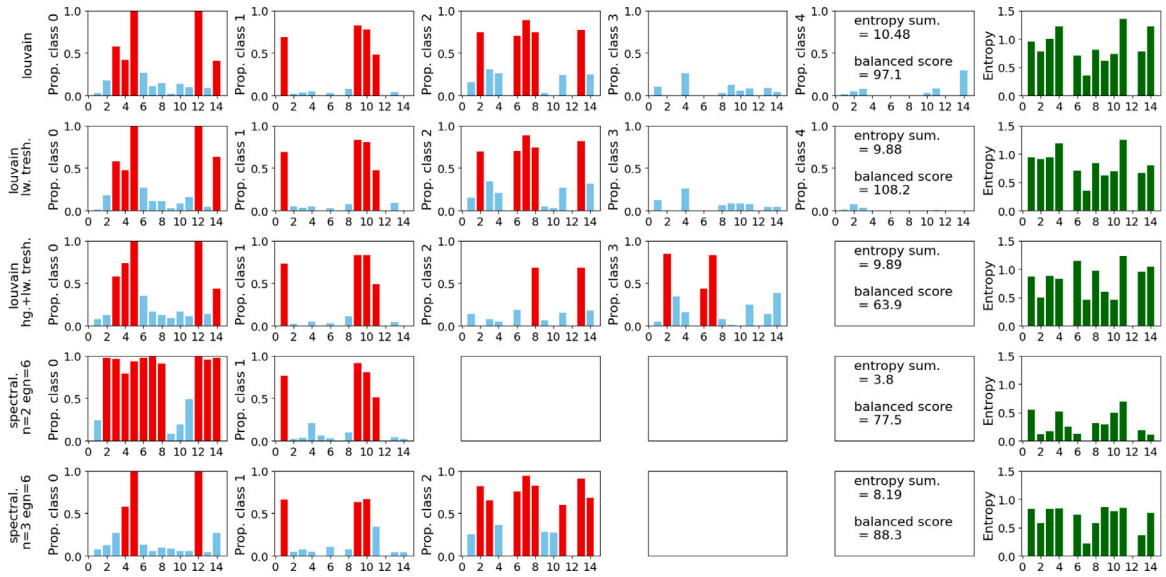

Fig. 4. Clustering results part.2.

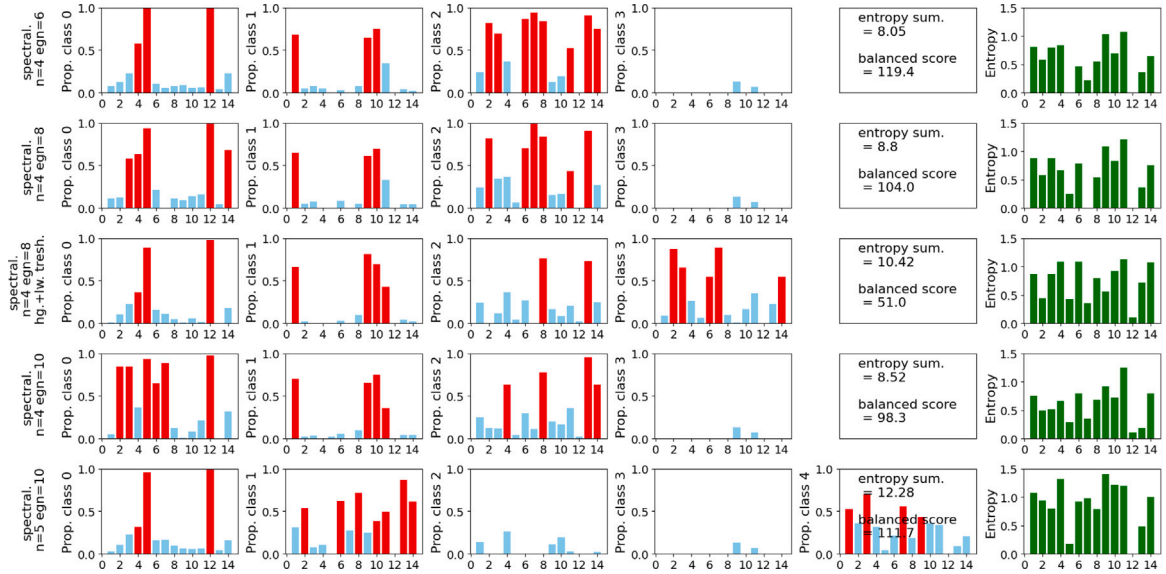

Fig. 5. Clustering results part.3.

5. Conclusion

This paper makes three main contributions. First, it systematically compares classical and graph-based clustering methods for cybersecurity textual data. Second, it aggregates the fourteen MITRE ATT&CK tactics into four interpretable "super tactics," improving coherence and dimensionality reduction. Third, it provides a methodological framework that links ATT&CK to corporate financial disclosures for risk quantification. While modern transformer-based embeddings such as BERT [1] and BERTopic [4] offer promising extensions, our choice of doc2vec ensures computational efficiency, interpretability, and methodological continuity. Future research could adapt this framework to transformer embeddings for even greater scalability and domain adaptation.

Our clustering analysis identified four critical super tactics (aggregation of tactics): Preparation and Reconnaissance, Persistence and Evasion, Credential Movement, and Command and Data Manipulation. Each super tactic encapsulates a range of specific techniques employed by adversaries. Preparation and Reconnaissance encompasses tactics such as initial access and resource development, focusing on how adversaries gather information and prepare for attacks. Persistence and Evasion encompass techniques that enable attackers to maintain access to compromised systems while evading detection, underscoring the importance of stealth in cyber operations. Credential Movement addresses methods attackers use to gain and exploit credentials, facilitating lateral movement within networks to access sensitive information. Command and Data Manipulation illustrates how adversaries control compromised systems and exfiltrate data, showcasing the ultimate goals of cyber intrusion. We observe that each super-tactic is a natural and causal step from the previous super-tactic, further highlighting the coherence of the aggregated structure.

We explore multiple methods to refine our clustering results, including K-means, Louvain, and Spectral clustering. Our analysis revealed that while K-means provided a coherent but somewhat crude initial structure, the Louvain method significantly enhanced heterogeneity, particularly in distinguishing tactics like Resource Development and Reconnaissance. However, we also identified the need to fine-tune similarity thresholds to prevent the creation of superfluous clusters due to syntactically similar but semantically distinct paragraphs. The

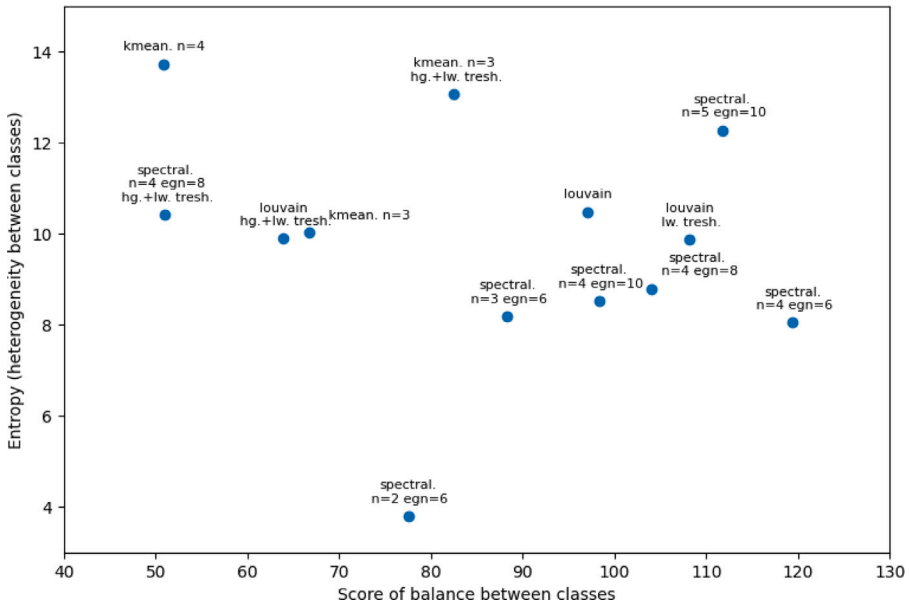

Fig. 6. Comparison of clustering scores: Entropy sum and Balanced score.
Each clustering method of Figs. 3, 4, and 5 is presented here using their respective entropy sum and balanced score. Recall that the aim was to reduce both scores to distinguish the best clustering method. Also, note that there is no guideline regarding the optimal additional amount to forfeit to the entropy sum to lower the balanced score and vice versa.

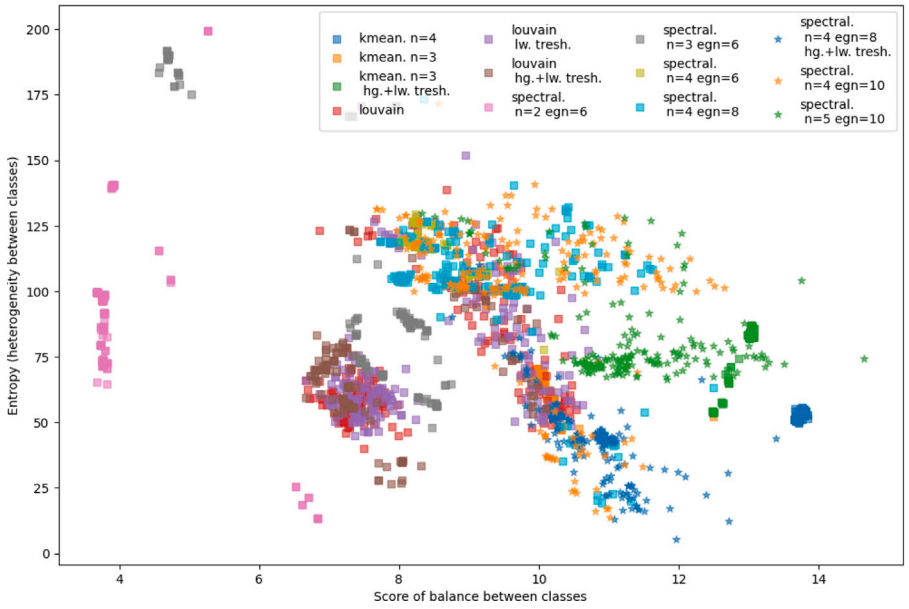

Fig. 7. Clustering scores: Entropy sum and balanced score N = 200.
The experiment of Fig. 6 is reproduced here by reapplying 200 times the involved clustering methods to avoid the bias of random initialization. Each clustering method is displayed using its respective entropy sum and balanced score.

Louvain method was chosen as our baseline clustering approach, not because of its slight performance advantages over other methods but because of its practicality. This method yields meaningful clusters without extensive hyperparameter tuning, making it a straightforward and practical choice.

While we use a case study focused on cybersecurity risks using the MITRE ATT&CK framework, our methodology has much broader potential. The combination of doc2vec embeddings with clustering techniques, particularly the Louvain method, can be readily adapted to analyse other forms of risk, such as legislative, geopolitical, or environmental risks. This approach could efficiently identify and categorize risk factors embedded in corporate or regulatory texts by substituting the MITRE ATT&CK knowledge base with a similarly structured dataset relevant to these domains. Maintaining a coherent and structured aggregation of textual information makes this method a powerful tool for exploring risk in various sectors, such as finance and public policy.

CRediT authorship contribution statement

Nathan Monnet: Writing – review & editing, Visualization, Software, Investigation, Formal analysis, Data curation. **Loïc Maréchal:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Conceptualization.

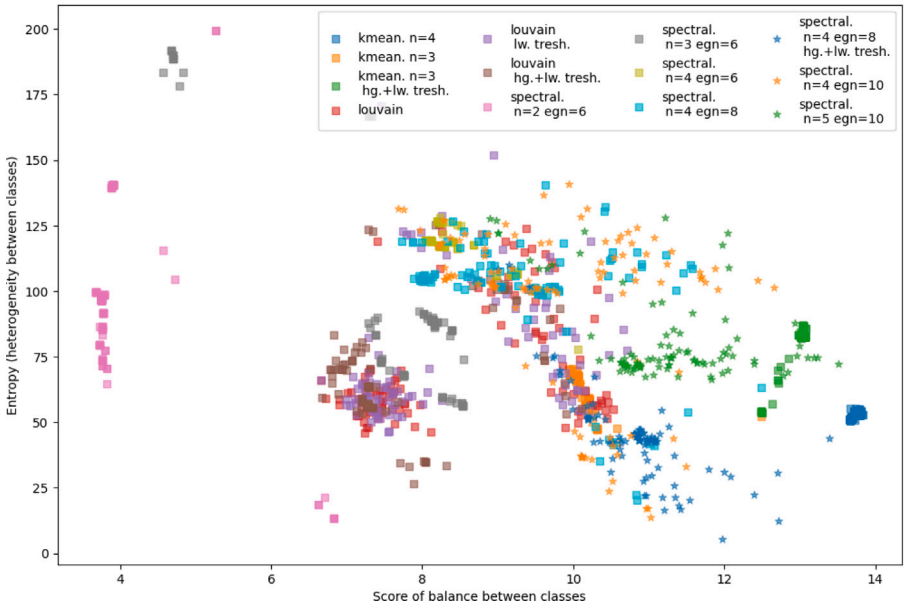

Fig. A.1. Clustering scores: Entropy sum and balanced score N = 100.
The experiment of Fig. 6 is reproduced here by reapplying 100 times the involved clustering methods to avoid the bias of random initialization. Each clustering method is displayed using their respective entropy sum and balanced score.

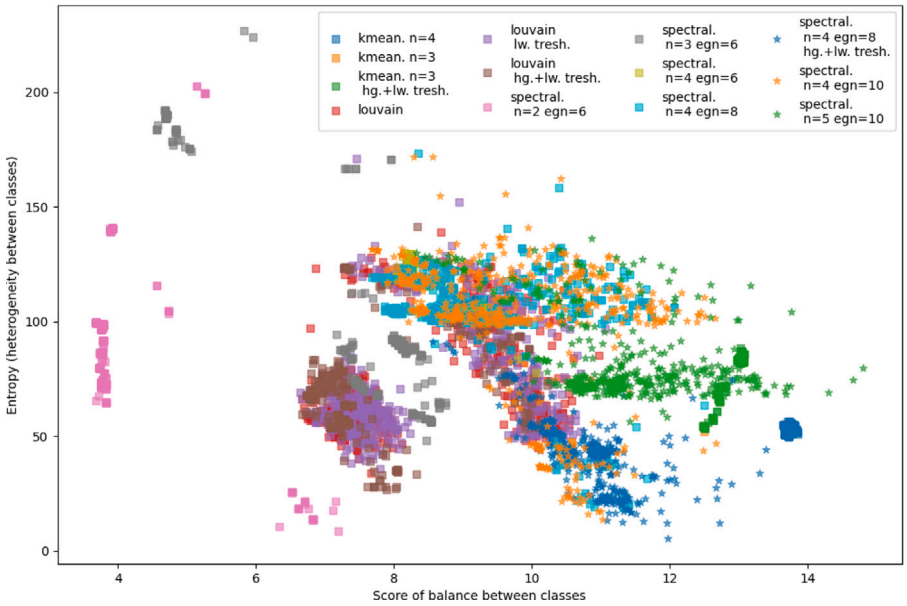

Fig. A.2. Clustering scores: Entropy sum and balanced score N = 500
The experiment of Fig. 6 is reproduced here by reapplying 500 times the involved clustering methods to avoid the bias of random initialization. Each clustering method is displayed using their respective entropy sum and balanced score.

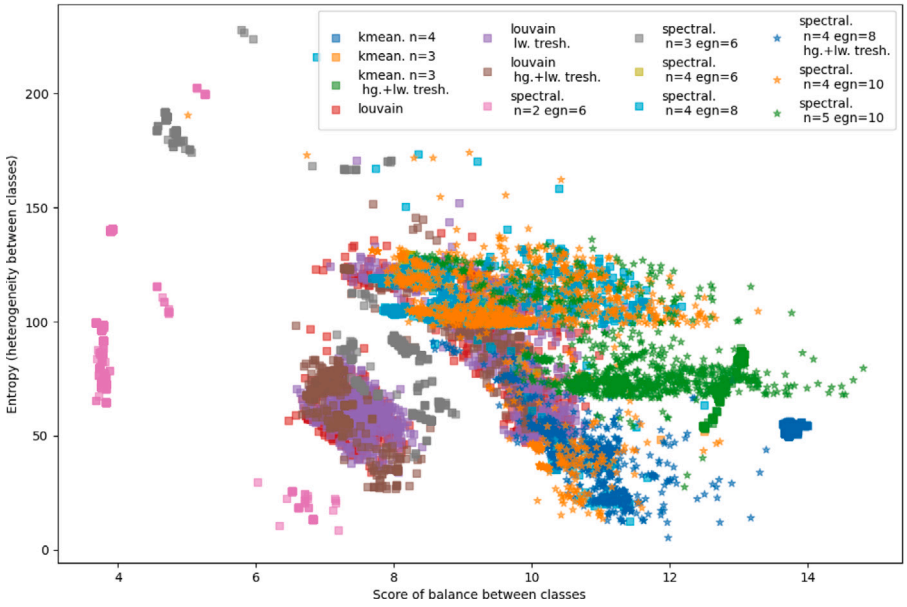

Fig. A.3. Clustering scores: Entropy sum and balanced score N = 1000.
The experiment of Fig. 6 is reproduced here by reapplying 1000 times the involved clustering methods to avoid the bias of random initialization. Each clustering method is displayed using its respective entropy sum and balanced score.

Declaration of competing interest

The authors have no conflict of interest and nothing to disclose.

Appendix

See Figs. A.1–A.3.

Data availability

Our code and data are publicly available on our github, see, e.g., https://github.com/nmonnet.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, Available at: https://doi.org/10.48550/arXiv.1810.04805.
[2] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, 2019, Available at: https://doi.org/10.48550/arXiv.1903.10676.
[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019, Available at: https://doi.org/10.48550/arXiv.1907.11692.
[4] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022, Available at: https://doi.org/10.48550/arXiv.2203.05794.
[5] D. Celeny, L. Maréchal, Cyber risk and the cross section of stock returns, 2023, Available at http://dx.doi.org/10.2139/ssrn.4587993.
[6] C.W. Calomiris, H. Mamaysky, How news and its context drive risk and returns around the world, J. Financial Econ. 133 (2019) 299–336.
[7] T.A. Hassan, S. Hollander, L. van Lent, A. Tahoun, Firm-level political risk: Measurement and effects, Q. J. Econ. 134 (2019) 2135–2202.
[8] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. (2008) P10008.
[9] Z. Wang, J. Chen, J. Chen, H. Chen, Identifying interdisciplinary topics and their evolution based on BERTopic, Scientometrics 129 (2024) 7359–7384.
[10] M. Wu, G. Siversten, L. Zhang, F. Qi, Y. Zhang, Scaling research aim identification: Language models for classifying scientific and societal-oriented studies, J. Assoc. Inf. Sci. Technol. 70004 (2025) 1–18.
[11] M. Wu, Y. Zhang, M. Markley, C. Cassidy, N. Newman, A. Porter, COVID-19 knowledge deconstruction and retrieval: An intelligent bibliometric solution, Scientometrics 129 (2024) 7229–7259.
[12] W. Antweiler, M.Z. Frank, Is all that talk just noise? The information content of internet stock message boards, J. Finance 59 (2004) 1259–1294.
[13] D. Garcia, Sentiment during recessions, J. Finance 68 (2013) 1267–1300.
[14] N. Jegadeesh, D. Wu, Word power: A new approach for content analysis, J. Financial Econ. 110 (2013) 712–729.
[15] O. Arslan-Ayaydin, K. Boudt, J. Thewissen, Managers set the tone: Equity incentives and the tone of earnings press releases, J. Bank. Finance 72 (2016) 132–147.
[16] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, Available at https://arxiv.org/abs/1301.3781.
[17] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: E.P. Xing, T. Jebara (Eds.), Proceedings of the 31st International Conference on Machine Learning, PMLR, Bejing, China, 2014, pp. 1188–1196.
[18] Z. Sautner, L. van Lent, G. Vilkov, R. Zhang, Firm-level climate change exposure, J. Finance 78 (2023) 1449–1498.
[19] G. Adosoglou, G. Lombardo, P.M. Pardalos, Neural network embeddings on corporate annual filings for portfolio selection, Expert Syst. Appl. 164 (2021) 114053.
[20] L. Cohen, C. Malloy, Q. Nguyen, Lazy prices, J. Finance 75 (2020) 1371–1415.
[21] L. Maréchal, N. Monnet, Cybersecurity tactics: A clustering approach with MITRE ATT&CK, 2024, Available at: https://arxiv.org/abs/2409.08728.
[22] J.H. Lau, T. Baldwin, An empirical evaluation of doc2vec with practical insights into document embedding generation, in: Proceedings of the 1st Workshop on Representation Learning for NLP, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 78–86.

Nathan Monnet received his B.Sc. and M.Sc. degrees in Physics and a Master's degree in Financial Engineering from the École Polytechnique Fédérale de Lausanne (EPFL). His research interests lie at the intersection of quantitative finance, natural language processing, and risk modelling. He recently completed an internship focused on developing NLP-based metrics to assess firms' exposure to cyber threats, a project that extends his thesis work in financial engineering. He is proficient in Python and has experience in collaborative, data-driven research environments. He is currently based in Switzerland.

Loïc Maréchal holds a M.Sc. degree in finance from the University Paris Nanterre and a Ph.D. from the University of Neuchâtel. He is a scientific collaborator at the Institute of Entrepreneurship and Management at the HES-SO Valais-Wallis. In this position, he utilizes financial methods to estimate the costs of cyberattacks and assess the value of cybersecurity-providing firms. He has been teaching finance at the Universities of Neuchâtel and Geneva, as well as at ESSEC Business School and Les Roches. He has over 10 years of experience in commodity markets, which includes working on trading desks and writing his Ph.D. dissertation. He also has a strong interest in machine learning and market microstructure. He is currently based in Switzerland.