



# City Research Online

## City St George's, University of London

**Citation:** Gad, M., Park, G., Rawsthorne, S. & Young, S. (2026). When methods matter: how implementation choices shape topic discovery in financial text. *Accounting and Business Research*, doi: 10.1080/00014788.2026.2625716

This is the published version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/37457/>

**Link to published version:** <https://doi.org/10.1080/00014788.2026.2625716>

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).



## When methods matter: how implementation choices shape topic discovery in financial text

Mahmoud Gad, Gitae Park, Sam Rawsthorne & Steven Young

To cite this article: Mahmoud Gad, Gitae Park, Sam Rawsthorne & Steven Young (20 Apr 2026): When methods matter: how implementation choices shape topic discovery in financial text, Accounting and Business Research, DOI: [10.1080/00014788.2026.2625716](https://doi.org/10.1080/00014788.2026.2625716)

To link to this article: <https://doi.org/10.1080/00014788.2026.2625716>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 20 Apr 2026.



[Submit your article to this journal](#)



Article views: 230



[View related articles](#)



[View Crossmark data](#)

# When methods matter: how implementation choices shape topic discovery in financial text

MAHMOUD GAD<sup>a</sup>, GITAE PARK<sup>b</sup>, SAM RAWSTHORNE<sup>a\*</sup> and STEVEN YOUNG<sup>a</sup>

<sup>a</sup>*Department of Accounting and Finance, Lancaster University Management School, Lancaster, UK;*

<sup>b</sup>*Faculty of Finance, Bayes Business School, City St George's, University of London, London, UK*

This paper examines the application of LDA topic modelling to risk disclosures in FTSE350 firms' annual reports. We show that LDA implementation choices significantly impact topic representations and subsequent inferences. Using a corpus of FTSE350 annual reports, we show that preprocessing decisions, multiword expressions and labelling strategies materially affect topic interpretability and granularity. Our analysis reveals that while risk reporting addresses key business risks at an aggregate level, the degree of firm-specific commentary is sensitive to topic granularity. Hierarchical linear modelling suggests that 27% of topic variation is within firms for broad topics, increasing to 75% for granular topics. We leverage GPT to enhance topic labelling, showcasing the potential of LLMs in financial text analysis. We also compare LDA to modern embedding-based topic models, finding that while they often generate more coherent topics, they introduce a new set of critical implementation choices and do not eliminate the need for researcher discretion. These findings challenge the claims of LDA objectivity and highlight the importance of domain expertise. We propose a practical checklist for LDA implementation in accounting and finance research emphasising transparency and robustness checks.

**Keywords:** textual analysis; topic modelling; risk disclosure; annual reports; Latent Dirichlet Allocation; GPT

**JEL Classification:** C55, M41, G14

## 1. Introduction

As the volume of narrative disclosure in financial markets grows exponentially, researchers face a critical challenge: how to reliably extract meaningful themes from vast corpora of text by combining methodological rigour with domain expertise. Unsupervised topic modelling using Latent

---

\*Corresponding author. Email: s.rawsthorne1@lancaster.ac.uk

Paper accepted by Edward Lee

Dirichlet Allocation (Blei et al. 2003) is gaining significant traction in the accounting and finance literature as researchers seek to understand and exploit the huge volumes of narrative disclosure. Over the last decade, Latent Dirichlet Allocation (LDA) has been applied to a variety of financial corpora including 10-K annual reports (Ball et al. 2015, Dyer et al., 2017), analysts' reports (Huang et al. 2018), tax disclosures (Bozanic et al. 2018), conference calls (Donovon et al. 2021), and financial news (Bybee et al. 2024). However, despite LDA's growing popularity, a fundamental tension remains; while it aims to reduce subjectivity in theme identification, implementation choices can nevertheless affect inferences. Although researcher discretion and its effects have been studied in the computational linguistics field, LDA applications in accounting are often portrayed as an objective, replicable process. Using risk disclosures from FTSE350 firms, we show how basic LDA methodological choices produce systematically different conclusions about disclosure quality and informativeness.

We focus on a rich corpus of risk disclosures published between 2018 and 2022 and shaped by UK regulations including the Companies Act 2006, Guidance on the Strategic Report (2018), the UK Corporate Governance Code (2018), and guidance from the Financial Reporting Council (2014, 2021). The setting supports our research objective in two ways. First, guidance and analysis published by the Financial Reporting Council (FRC) (FRC 2018a, 2018b, 2022) and the FRC Lab (2021) provide a comprehensive and objective benchmark for evaluating LDA outputs and variation therein caused by research design choices. Second, despite FRC and FRC Lab guidance, management retain substantial discretion in determining which risks are material and how to report on them. The resulting variation in topic coverage serves to foreground the effect of LDA implementation choices on findings and conclusions.<sup>1</sup> While our primary goal lies in demonstrating the impact of researcher discretion in LDA applications, our topic analysis also provides insights on risk reporting among FTSE350 firms that speak to FRC concerns about boilerplating and limited coverage of key risks in areas such as human rights and cybersecurity threats.

We scratch beneath the veneer of LDA algorithmic objectivity to highlight how implementation choices can enhance the method's effectiveness for extracting meaningful insights. Specifically, we review five considerations in LDA model implementation and discuss the choices facing researchers in each case. We consider: (1) text preprocessing, (2) hyperparameter tuning, including optimising the number of LDA topics, (3) evaluating LDA model performance, (4) moving beyond the bag-of-words (BoW) model on which LDA is based, and (5) topic labelling. Despite widespread adoption of topic modelling in finance and accounting research, the field lacks a comprehensive understanding of how these choices impact discovery and interpretation of latent themes in corporate disclosures even though the tripwires are clearly articulated in the computational linguistics literature. For each consideration, we outline the issue(s) that researchers need to consider and the options available, and we illustrate applications and their impact using our corpus of risk disclosures. We draw on these discussions to build a checklist to which accounting researchers can refer when implementing and describing their LDA model to ensure their work is robust, transparent, and replicable. We do not recommend a specific 'solution' in most cases because no universally best option typically exists. Instead, the final choice is contingent on the research setting and therefore rests with the researcher to justify transparently. We also seek to promote informed and transparent application of LDA models in accounting research by providing a comprehensive set of resources in the form of further technical discussion, examples, and (python) script snippets in an accompanying Appendix.

---

<sup>1</sup>While we focus on UK risk disclosures, our main findings regarding LDA implementation generalise to other regulatory settings, as we demonstrate in Internet Appendix Table A9 where we replicate our main analysis using Item 1A: Risk Factors disclosures from US 10-K filings.

We conduct five experiments to illustrate how choices in modelling LDA topics, whether implicit or explicit, impact findings and conclusions about the content of annual report risk narratives. Results reveal how text preprocessing choices, careful treatment of important multiword expressions (phrases), and use of part-of-speech (POS) tagging to leverage the basic rules of grammar can overcome some of the more egregious limitations of LDA's BoW assumption. We also show how different approaches to selecting LDA hyperparameters including the optimal number of topics ( $K$ ) impacts topic representation, and how methods such as visualisation can help to fine-tune model hyperparameters and reveal topics that are more interpretable. Critically, we demonstrate that a mechanical approach of setting  $K$  equal to the maximum coherence score (or any of the alternative evaluation metrics that are available) is unlikely to generate the most interpretable set of topics. Finally, we demonstrate the critical role of labelling strategy for interpreting results and informing subsequent analyses. Holding the number of topics constant, we show how labels (and therefore insights) vary between a single labeller, multiple labelers working as team, and generative AI labelling using GPT. In the latter case, we show how more refined prompting strategies can generate more descriptive labels.

We use insights from our experiments utilising FTSE350 firms' annual report risk disclosures to illustrate the consequences of these LDA methodological choices. Our analysis reveals how seemingly minor implementation decisions can lead to fundamentally different conclusions about the properties of corporate disclosure. For instance, conclusions about whether risk reports contain generic 'boilerplate' commentary or firm-specific analysis depend on the level of topic granularity set by the researcher. We show that a broad, 35-topic model, which scores highly on standard evaluation metrics, leads to the conclusion that risk disclosures are largely stale and generic. Conversely, using a more granular 100-topic model suggests that firms provide distinctive tailored commentary. The evidence supports the view that the 'correct' conclusion is not an objective output of the LDA algorithm but is instead contingent on researcher choice during the modelling process. The very definition of what constitutes boilerplating versus discussion of a meaningful risk is not a pre-existing fact to be discovered, but a conclusion shaped by the decisions made during the analytical process.

While our primary analyses focus on LDA, we also acknowledge recent advances in transformer architectures that support semantic text embeddings and therefore help to relax the BoW assumption. We repeat a subset of our analyses using topic models that leverage two embedding-based approaches: BERTopic (Grootendorst 2022) based on Google's BERT language model (Devlin et al., 2019) and Alibaba Group's GTE lightweight model built on BERT (Li et al. 2025). Results yield two important insights. First, embedding-based pipelines often (but not always) generate more coherent and interpretable topic representations, meaning that accounting researchers should not view LDA as their default topic modelling method. Second, adopting this new generation of embedding-based methods still requires researchers to make active, informed, and transparent implementation choices.

A key takeaway from our experiments is that accounting researchers must not shy away from acknowledging and embracing discretion when applying topic modelling. At a minimum, recognising and justifying available choices is essential for delivering rigorous analysis that yields reliable and replicable inferences. Additionally, discretion provides the opportunity for accounting and finance researchers to leverage their specialist knowledge and elevate the work from a purely statistical exercise that any data scientist could perform, to a rich, context-informed analysis where the incremental value of domain expertise is clear.

Our conclusions parallel recent methodological insights in other areas of accounting and finance research. Breuer and DeHaan (2024) demonstrate that while fixed effects regression is powerful for eliminating unwanted variation, its use requires careful justification, as it can transform samples and variables in unintended ways. In addition, Menkveld et al. (2024) document

substantial ‘nonstandard errors’ arising from researcher discretion in implementing statistical tests. In both cases, the insights are new to many accounting researchers despite being common knowledge among econometricians. In a similar vein, we show that topic modelling implementation requires careful consideration of methodological choices that despite being well known to computational linguists are not fully appreciated by accounting researchers. Many critical topic modelling implementation choices often go undocumented in accounting research, creating hidden variation in how the same textual data can be analysed. We contribute to research on corporate disclosure by illustrating the need to acknowledge, document and carefully justify choices when applying topic modelling.

Our study also makes two practical contributions to the rapidly expanding literature in accounting and finance that uses NLP methods to extract information from textual disclosures. Our first practical contribution takes the form of an accessible guide to the pitfalls and opportunities of applying topic modelling. We use this guide to develop a simple checklist of implementation considerations for authors, reviewers, and research consumers to follow when constructing and interpreting topics. The checklist serves to highlight the choices that researchers must address and, critically, the opportunities for applying their domain expertise to enhance the richness of results and insights. Our second practical contribution is a suite of resources designed to lower the costs and increase the robustness of topic modelling for the typical accounting and finance researcher who has limited formal NLP training. Reducing costs in the form of barriers to entry and time is important because it allows accounting and finance researchers to invest proportionately more energy leveraging their comparative advantage in the form of domain expertise.

## 2. Background, literature review and research questions

### 2.1 *Critique of topic modelling in accounting and finance research*

LDA topic modelling provides researchers with a powerful means of organising, summarising and understanding very large archives of documents such as annual reports, analyst reports, earnings press releases, and prospectuses. The method is a form of unsupervised categorisation that relies on a statistical algorithm to ‘discover’ hidden groupings of tokens in the document collection.<sup>2</sup> Having identified a fixed number of clusters (or topics) in the entire corpus, researchers are then able to represent documents by a distribution of discovered themes. For example, Document 1 may contain 10% Topic Y and 90% Topic Z, while Document 2 comprises 40% Topic Y and 60% Topic Z. The unsupervised approach is appealing because there is no need to train the model beforehand or even pre-specify what topics to look for in the data. The algorithmic nature of the method also leads many empiricists to conclude that the results are more objective and replicable than manual approaches to scoring text. For example, Huang et al. (2018, p. 2835) state that ‘LDA offers several advantages over manual coding ... [it] provides a reliable and replicable classification of topics’; Dyer et al. (2017, p. 223) claim the method provides ‘... an approach for evaluating topical coverage for large samples of lengthy documents on a consistent and objective basis over time.’; and Ball et al. (2015, p. 4) argue that ‘... LDA generates a rich audit trail of

---

<sup>2</sup>LDA is one of several families of methods for modelling latent structure in text corpora. Alternatives include probabilistic models such as Probabilistic Latent Semantic Analysis (PLSA); algebraic/matrix-factorisation methods such as Non-Negative Matrix Factorisation (NMF); and embedding-based/neural approaches that cluster transformer sentence or document embeddings and derive sparse topic descriptors (e.g., Top2Vec, BERTopic, and contextualised topic models). See Hofmann (1999) for PLSA, Lee and Seung (1999) for NMF, Angelov (2020) for Top2Vec, Grootendorst (2022) for BERTopic, and Bianchi et al. (2021) for contextualised topic models.

topic vocabularies that can be compared across firms. These vocabularies are computer generated and are not susceptible to researcher prejudice or data mining.’

While there is little doubt that LDA represents a valuable text mining tool for accounting and finance researchers, claims regarding its objectivity and ability to produce meaningful insight deserve closer scrutiny. Many limitations of LDA are well documented in the NLP literature (e.g. Mu et al. 2024b, Schofield and Mimno 2016):

- (1). Although LDA models rely exclusively on a statistical clustering algorithm, results are highly conditional on a suite of parameters that the researcher must select. These include the optimal number of topics ( $K$ ), the Dirichlet prior for document-topic distribution, the Dirichlet prior for topic-word distribution, and the learning method used for inference. While applications in accounting typically discuss the approach to selecting  $K$ , other parameter choices are rarely mentioned. This lack of transparency impedes replicability.
- (2). A variety of metrics are available to inform the choice of optimal parameters, with different evaluation metrics producing different outcomes. Further, none of the evaluation metrics guarantee that topics in the ‘best’ model are optimal from an interpretability perspective. Careful tuning is required to generate interpretable topic representations and this tuning process requires researchers to exercise discretion.
- (3). Even with full transparency over parameter choices, replicability is questionable because the LDA algorithm is stochastic and so models generated with the same parameters over the same data will produce different results.
- (4). LDA is a BoW method and consequently there is no guarantee that the topics it generates will be semantically meaningful. Topics are merely probabilistic relations between tokens; the LDA algorithm does not discover *meaning*.
- (5). The task of attributing meaning to LDA topic representations requires the researcher to assign labels to topics and this process is unavoidably subjective.

Accordingly, LDA topics are not necessarily replicable, semantically meaningful, and free from researcher bias. Similar to a manual coding approach, researchers must apply significant amounts of judgement. Contrary to studies applying manual coding procedures (e.g. Beattie et al. 2004, Comyns and Figge 2015), however, the judgements that researchers make when applying LDA are often hidden. Moreover, the view that researcher judgement is something to avoid when modelling topics in financial text overlooks the value of domain expertise. Expert knowledge of the research phenomenon is the main competitive advantage for accounting and finance researchers over a pure data science approach that treats the setting as a black box. Our view is that the application of theoretically informed and logically consistent judgement is something we should be promoting as a discipline, not restricting. We aim to shed light on key areas of judgement in LDA modelling, the impact these choices can have on results, and how careful application of judgement can lead to more informative insights. The context in which we explore these issues is risk reporting.

## 2.2 Empirical setting: UK risk reporting

To illustrate the impact of methodological choices, we use the setting of risk disclosures in UK annual reports. This context is ideal for three reasons. First, the regulatory environment, guided by the UK Corporate Governance Code and the FRC, mandates detailed discussion of risks, creating a rich corpus for analysis. Second, despite this regulation, a persistent debate exists regarding disclosure quality, with regulators and researchers frequently citing concerns about generic, ‘boilerplate’ narratives that lack firm-specific insight (FRC 2022, Abraham and Shrivs 2014). This tension makes it a suitable setting to test how different analytical choices can lead to different

conclusions about disclosure quality. Third, guidance from the FRC provides an objective, external benchmark for evaluating the topics our models discover, which we leverage in our analysis.

We extract principal risks and uncertainties (hereinafter risk) commentary from PDF annual reports with fiscal year-ends in 2018–2022 published by FTSE350 firms. We tag a firm as a FTSE350 constituent if it is an index constituent for at least one quarter at any point during our sample window. We include all reports from tagged firms during the sample period, regardless of whether they are a FTSE350 constituent in a particular calendar year. Our FTSE350 sample comprises a maximum of 2,268 reports. We use the tool from El Haj et al. (2020) to extract annual report text and classify sections using headers in the report table of contents. We classify principal risks and uncertainties sections of the report as those whose header contains the text string ‘risk\*’ and ‘uncertain\*’. For reports that contain multiple principal risks-related sections, we aggregate text from the relevant sections.<sup>3</sup> Our risk corpus consists of 1,992 reports from 478 unique firms after removing cases where we are unable to detect risk commentary automatically. The risk corpus contains 7.1 million tokens after removing punctuation, stop words, and single-letter words.

### 3. Areas of judgement in LDA topic modelling

Just how objective, replicable, and reliable is LDA topic modelling in terms of generating semantically meaningful insights? This section reviews implementation decisions where significant researcher judgement is unavoidable. The implementation decisions on which we focus are text preprocessing; ways to relax the BoW constraint; hyperparameter tuning; evaluating model performance and topic interpretability; and topic labelling.<sup>4</sup> Further detail regarding each aspect is available in the appendix.

#### 3.1. Text preprocessing

Careful preparation of the document corpus is a key step for LDA topic modelling. The properties of the input corpus can have a dramatic effect on topic interpretability and processing time. We distinguish between two levels of preprocessing. The first level, which we view as nondiscretionary involves basic clearing tasks such as lowercasing, removing HTML tags, non-ascii characters, symbols and elements of punctuation, ensuring consistency in spelling, abbreviations and hyphenation, and removing core stop words (i.e. unigrams that carry little or no useful information). While these steps are nondiscretionary, a degree of judgement nevertheless rests with the researcher for tasks such as choosing the stop word list. The second level of preprocessing involves considerably more researcher discretion, both in terms of what to do and how to do it. Examples in this group include:

- Choice of tokenizer for breaking a stream of textual data into words, terms, sentences, symbols, or some other meaningful elements. Different tokenizers provide different functionality, which in turn can impact LDA outcomes.

<sup>3</sup>Risk-related disclosures that are not identified separately in the report table of contents are not included in the analysis. We also exclude sections that discuss specific risks such as climate change on the grounds that our primary interest is discussion of principal risks as defined by the reporting entity.

<sup>4</sup>The list of implementation decisions is not exhaustive. Other hyperparameters in the LDA algorithm that permit or require researcher judgement include chunk size, passes (epochs), and iterations. However, we focus on the highlighted aspects as they tend to have the most substantial impact on topic interpretability and semantic coherence, while also being the most commonly reported parameters in applied LDA research.

- Lemmatising and stemming to convert inflected words to a common base root. Stemmers eliminate affixes from words (e.g. runs, runner, running are mapped to run). Lemmatizers reduce derivationally related forms of word to a single dictionary base form (e.g. democracy, democratic, democratisation are mapped to their root form). These processes help to reduce dimensionality and improve clustering accuracy. However, outputs vary depending on the choice of stemming and lemmatising algorithm.
- Removing context-specific stop words to further reduce dimensionality.
- Removing very frequent and very rare words. High frequency words provide little or no discriminatory power across topics and can bias scoring functions. Rare words are candidates for removal because their association with other words is typically dominated by noise. Various statistics are available for identifying high and low frequency words but there is no accepted norm for setting cut-off levels.

The optimal combination of preprocessing steps is contingent on the text domain, the source and size of the input text, and the research purpose; simply following the steps in previous work may not be appropriate for a different dataset and research question. Careful judgement is therefore necessary to ensure internal and external consistency, while full transparency regarding steps in the processing pipeline is critical to ensure replicability.

### 3.2. *Multiword expressions*

Most LDA applications work with unigrams by default and therefore treat common multiword expressions (MWEs) such as ‘earnings per share’ as separate tokens rather than a single term. Meanwhile, the BoW assumption ignores context and meaning. For example, ‘bank’ is more likely to load on a financial services topic when it appears before ‘borrowing’, but on a trading performance topic when it appears before ‘holiday’. Researchers face the choice to operate exclusively at the unigram level or instead to capture semantically meaningful MWEs, with the final decision having a potentially significant impact on LDA results and topic interpretability.

### 3.3. *Hyperparameters*

The LDA algorithm requires researchers to define a number of parameters including:

- The optimal number of topics ( $K$ ).
- The Dirichlet prior for the document-topic distribution ( $\alpha$ ).
- The Dirichlet prior for the topic-word distribution ( $\beta$ ).
- The inference algorithm or learning method (e.g. Gibbs sampling with the *mallet* algorithm or variational inference with the *gensim* (Hoffman et al. 2010) algorithm).

Since default values for  $\alpha$  and  $\beta$  apply for each learning method, and choice of LDA package (e.g. *scikit-learn* and *gensim* in python or *mallet* with python wrapper) fixes the learning method, researchers can easily make (implicit) choices without appreciating the effects. Relying on default settings does not provide a sound basis for implementation.

### 3.4. *Evaluation*

Selecting the ‘best’ LDA model from a grid of hyperparameter combinations involves choosing an evaluation metric. Popular measures for selecting the optimal combination of hyperparameters include perplexity, coherence, diversity, and granularity (e.g. Chang et al. 2009, Röder

et al. 2015). Some measures such as coherence and diversity favour models with fewer topics whereas granularity tends to favour richer topic representation. There is no magic formula for determining the best evaluation metric to adopt.

Topic modelling is an inherently interpretive task. A purely metric-based approach to hyperparameter optimisation cannot therefore guarantee that the topic representation with the highest score will yield the cleanest (most interpretable) set of topics. Interpretation is one of the steps in the LDA pipeline where accounting and finance researchers enjoy a substantial competitive advantage over data scientists. Exercising this judgement is therefore a process to be encouraged, not downplayed or avoided entirely. Best practice encourages manual intervention when selecting the preferred topic representation to maximise interpretability (Chang et al. 2009). Several options for manual input are available including Word Intrusion Tests (WIT) (Dyer et al. 2017) and graphical representation of the topic space based on topic distance scores (e.g. pyLDAvis library in python). Finally, researchers can pre-determine a set of topics based on their domain expertise and then examine the extent to which the trained model covers these topics. These interpretative methods are complements rather than substitutes. Using multiple methods provides accounting and finance researchers with the opportunity to leverage their domain expertise to elevate the analysis beyond a pure text mining exercise. We add more detail on evaluation in section 1.2.4 in the appendix.

### 3.5. Labelling

The final step in the LDA topic model pipeline involves labelling topics. Although some black-box tests rely on unlabelled topics as inputs, most empirical analyses aim to discover meaning at some level and therefore labelling is a central part of the modelling process. Labelling is a wholly interpretative task that is entirely separate from the LDA algorithm and not amenable to quantification. Researcher discretion is therefore unavoidable regardless of the particular labelling strategy applied. Indeed, labelling is another step in the LDA pipeline where domain expertise, when applied rigorously, provides a comparative advantage for accounting and finance researchers.

Table 1 summarises the state-of-the-art for LDA topic modelling in accounting and finance research. Several observations emerge from this review. First, the text preprocessing pipelines vary considerably across studies, both in terms of transparency and specific steps. Some studies incorporate advanced features such as MWEs and lemmatisation (e.g. Fedyk 2024, Bybee et al. 2024), while others apply very basic steps such as stop word removal. The treatment of rare and frequent tokens is inconsistently reported. Further, where processing steps are discussed, the level of detail is typically not sufficient to support precise replication. Second, many studies do not report specific  $\alpha$  and  $\beta$  values, suggesting possible use of default parameters. Where parameter values are disclosed, they are typically fixed; we found no mention of hyperparameter tuning using a grid search.<sup>5</sup> Third, a variety of evaluation metrics are used including perplexity, coherence, and qualitative assessments. Very few studies report visualisations of the latent topic space as part of the model tuning process, and no studies of which we are aware work with multiple topic representations despite plausible alternatives almost certainly existing. Fourth, strategies for topic labelling, where relevant, attract little rigorous discussion or critical analysis. Finally, the opportunity for limiting researcher bias is a consistent theme that authors highlight when justifying the LDA method.

---

<sup>5</sup>The default parameters for popular LDA implementations can significantly impact results. For instance, Mallet's default  $\alpha$  is  $50/K$  (where  $K$  is the number of topics) and  $\beta$  is 0.01, while Gensim's defaults are  $\alpha = 1/K$  and  $\beta = 1/K$ .

Table 1. Representative examples of LDA topic model application in accounting and finance research.

Study	Summary	Text processing pipeline	Model hyperparameters	Topics/Evaluation method	LDA Visualisation
Ball et al. (2015)	Examines whether MD&A disclosures, which can be forward looking, complement backward looking financial statements when the business environment is changing rapidly.	Does not detail text processing pipeline explicitly.	Specific hyperparameters are not detailed,	Initially select 100 topics for parsimony. Removes 25 topics judged as boilerplate or non-economic industry-specific, resulting in 75 'economic' topics.	No visualisation.
Bozanic et al. (2018)	Examines how disclosure rules influence IRS enforcement and disclosure behaviours. Uses LDA topic model to analyse changes in the tax-related disclosures in response to IRS regulations.	Does not provide details of text processing pipeline.	Specific hyperparameters are not detailed,	Not available,	No visualisation.
Dyer et al. (2017)	Examines thematic content in a large sample of 10-K filings.	Refers to removing common stopwords along with words that do not occur in at least 100 documents, and terms in the top 0.1% of most common words.	Specific hyperparameters are not detailed, although mentions that a (prominence of topics) varies across the corpus.	Used perplexity to narrow down to models with 150, 200, and 250 topics. Performed Word Intrusion Task on these three models. Selected the 150-topic model based on best performance in the Word Intrusion Task.	Perplexity of LDA model for different numbers of topics
Hoberg and Lewis (2017)	Investigates whether fraudulent firms' 10-K MD&A disclosures exhibit abnormalities in text that can predict fraud.	Refers to removing stop-words defined as any word appearing in more than 25% of MD&A filings in first year of sample; and rare words appearing in less than 100 MD&As in the first year of the sample.	Specific hyperparameters are not detailed,	Follows Ball et al. (2015) and uses 75 topics. Topics generated using MetaHeuristica.	No visualisation.
Huang et al. (2018)	Compare the thematic content of analyst reports issued after earnings conference calls to the content of the calls themselves.	Refers to removing stop words and 'venue-specific language'. Consolidates financial terms into single words or common abbreviations and excludes company names and tickers to prepare for LDA analysis.	Sets $\alpha = 1.0$ and $\beta = 0.01$ following standard practice in the literature.	Perplexity. Computed perplexity scores for models with 2–120 topics. Selected 60 topics based on the point where improvement in perplexity score diminishes significantly.	Perplexity of LDA Model for Different Numbers of Topics

(Continued)

Table 1. Continued.

Study	Summary	Text processing pipeline	Model hyperparameters	Topics/Evaluation method	LDA Visualisation
Hanley and Hoberg (2019)	Identify key risk topics discussed by banks, then use Semantic Vector Analysis to convert these topics into interpretable risk factors.	Does not detail text processing pipeline explicitly.	Specific hyperparameters are not explicitly detailed,	The paper focus on 25 topics for parsimony.	No visualisation.
Brown et al. (2020)	Identify linguistic patterns in 10-K filings that predict future financial misstatements.	The pipeline removes non-narrative content (e.g. XBRL data, HTML tags) and special characters. Then it extracts the text by identifying MD&A sections and removing short paragraphs under 80 characters. Also, removed common and rare words.	While $\alpha$ and $\beta$ are not detailed in the paper, the online code were used: <code>olda = onlineLDA(vocab, K, D, 1./K, 1024., 0.7)</code>	31 topics were selected based on simulated results where the paper optimised the LDA model by choosing parameters that maximised the predictive power of topics for detecting financial misreporting in the training data. Also, used WIT as a secondary validation.	No visualisation.
Bellstam et al. (2021)	Develops a measure of corporate innovation using the text of analyst reports and examines its relationship with firm performance, patenting outcomes, R&D intensity, and general manager ability.	Removes common stopwords. Stems words to their root. Drops reports with under 100 or over 5,847 words.	Specific hyperparameters are not explicitly detailed,	15 topics ( $K = 15$ ). Identifies "innovation topic" using Kullback-Leibler divergence from an innovation textbook. The authors experimented with $K = 10$ and $K = 50$ and found that $K = 15$ produced the most useful and meaningful results.	Word cloud of the "innovation topic".
Lowry et al. (2020)	Analyses SEC-firm communications prior to IPOs using LDA and KL divergence.	Removes non-English words, stopwords, rare words. Applies stemming using PorterStemmer.	Specific hyperparameters are not explicitly detailed,	In determining the optimal number of topics for SEC letters, two methods were employed, both suggesting that eight topics were most appropriate. Selection methods: (1) LDA topic saliency (Goldsmith-Pinkham et al. 2016, p. 2) Hierarchical Dirichlet Process on 5% sample (Jegadeesh and Wu 2017).	No visualisation.

Donovan et al. (2021)	Develop measures of credit risk based on qualitative disclosures from earnings conference calls and MD&A disclosures.	Removes common stopwords, and words/phrases used in fewer than 10 conference calls. Converts high-frequency terms to single words. Constructs bigrams. Applies stemming using PorterStemmer.	Specific hyperparameters are not detailed.	Not available.	No visualisation.
Brown et al. (2022)	Use due diligence meeting notes to identify key topics discussed during the hedge fund manager evaluation process and then use Kullback-Leibler to measure information acquisition over time.	Omits documents with fewer than 3 content words. Standardises language. Removes stop words, rare words (<15 documents), and frequent words (>50% of documents). Applies lemmatizer but no further details provided.	Specific hyperparameters are not detailed.	Topic coherence UMass coherence score. Chose K = 30 topics, which maximises the average topic coherence.	No visualisation.
Fedyk (2024)	Analyse news articles to examine the effects of news positioning on the speed of price discovery in financial markets.	Removes stop words, rare words (unique to one document), and frequent words (>70% of documents). Construct bigrams but no details provided.	Specific hyperparameters are not detailed.	Varied the number of topics (k) from 10 to 40. For each k, used collapsed Gibbs sampling with 500 iterations. Selected k = 15 topics based on the highest log likelihood.	No visualisation.
Bybee et al. (2024)	Analyses ~800,000 Wall Street Journal articles (1984-2017), using a topic modelling approach to identify key themes in business news and quantify the proportion of attention each theme receives over time.	Excludes articles containing fewer than 100 words. Refers to removing common stop words and URLs. Applies 'light lemmatisation'. Generates bi-grams from adjacent unigrams to capture two-word phrases but does not filter. Removes unigrams and bigrams appearing in less than 0.1% of articles.	Sets $\alpha$ and $\beta$ equal to 1,	Estimates a variety of models with K ranging from 50 to 250 topics in increments of ten, and then select model with the highest Bayes factor. Also use tenfold cross-validation.	Cross-validation and Bayes factor visualisation. Also, topic hierarchy dendrogram,

## 4. How important are the choices in LDA topic modelling?

### 4.1 Experimental evidence

This section reports the results of five empirical experiments designed to assess the sensitivity of LDA topic model outcomes to implementation and interpretation decisions over which researchers exercise control and where judgement is therefore necessary.

#### 4.1.1. Experiment 1: hyperparameter search and evaluation

Our first experiment tests how topic representation varies with choice of LDA model hyperparameters and evaluation method. We set discrete values for various parameters to assess the impact of parameter choices but limit complexity of the grid search. We restrict the optimal number of topics ( $K$ ) to three options that reflect a broad topic representation ( $K = 25, 35$  and  $50$ ) and three options that generate a granular representation ( $K = 100, 200$  and  $400$ ). We choose high values for alpha ( $\alpha = 0.1, 0.5$ ) and beta ( $\beta = 0.01, 0.1$ ) to capture the broad topic structure, and low values for alpha ( $\alpha = 0.005, 0.01$ ) and beta ( $\beta = 0.005, 0.01$ ) to capture the granular structure. We also allow the ‘auto’ option for  $\alpha$  and  $\beta$  so the models can revise these parameters automatically to reflect the empirical Bayesian update.<sup>6</sup> Choices for  $\alpha$ ,  $\beta$ , and inference algorithm vary within our broad and granular topic structures as follows:

Parameters	Broad level topics	Granular level topics
K	[25, 35, 50]	[100, 200, 400]
a	[0.1, 0.5, auto]	[0.005, 0.01, auto]
b	[0.01, 0.1, auto]	[0.005, 0.01, auto]
Training algorithm	[gensim, mallet]	[gensim, mallet]

Our parsimonious text preprocessing pipeline involves lowercasing text and removing numerical digits, non-ascii characters, single letter words, and stop words as defined by Loughran and McDonald’s [genericLong](#) plus domain specific stop words that do not make significant contribution to meaning in the analysis of UK annual report (see Appendix for details). We do not apply a stemmer or lemmatizer, or generate N-grams, at this point. We filter words based on document frequency, defined as the proportion of documents including the target word, and then flex our low frequency word filters for our broad and granular topic representations. Specifically, we apply a five percent low frequency filter for our broad topic representations, whereas we allow a one percent filter for the granular topic representations as words appearing in less than five percent of reports may serve as critical words for granular level topics. We allow our high frequency word filter to vary between 50 and 70 percent in both groups. The least aggressive filtering (i.e. one percent and 70 percent filtering) yields a corpus of 3.7 million tokens from 7,228 unique words while the most aggressive filtering (i.e. five percent and 50 percent filtering) yields

<sup>6</sup>The "auto" option can help capture the underlying structure of documents and topics. However, this option may limit the researcher’s control over the model, hinder specific goals (such as obtaining very granular topics), and potentially lead to overfitting to the training data, which could reduce the model’s effectiveness on unseen documents. When  $\alpha$  and  $\beta$  are set to be estimated automatically, the model updates these parameters based on how well the current topic model fits the data. In Gensim, this process occurs during the M-step of the Expectation-Maximisation (EM) algorithm. In MALLET, the updates are performed using an Empirical Bayes approach, which involves updating the prior distribution of alpha and beta based on the observed data. The "auto" option can help models better capture the underlying structure of documents and topics.

2.2 million tokens from 3,029 unique words. Applying these hyperparameter combinations and word filters leads to a final loose grid search across 120 LDA models.

We assess the relative performance rankings of our 120 LDA models to the following five evaluation metrics: perplexity, coherence, diversity, WIT, and granularity. We discard for ranking purposes all trained models that contain one or more empty topics on the grounds the model fails to generate meaningful semantic groups. An empty topic is one where all words are evenly distributed within the topic (near uniform distribution). Our proxy for a uniform topic distribution is where the difference between the sum of top 10-word probabilities and the sum of bottom 10-word probabilities is less than 0.001. Fifty out of the 60 gensim models contain at least one unidentified topic. Results for Experiment 1 therefore involve a final set of 70 trained models, comprising 60 Mallet models and 10 Gensim models.

Table 2 contains results for Experiment 1. Panel A summarises model rankings by each of our five evaluation metrics. Our tabulation is restricted to the top five and bottom five models for each metric for parsimony and to ensure readability. Results are striking on several counts. First, no single combination of hyperparameter choices ranks among the top five models on all five evaluation metrics. Indeed, only three models rank in the top five for more than one evaluation metric: (G\_0.05\_0.5\_25\_auto\_auto) for coherence and diversity; (M\_0.05\_0.7\_25\_auto\_ -\_auto) for coherence and WIT; and (M\_0.01\_0.5\_400\_auto\_auto) for granularity and topic coverage. Notably, while M\_0.01\_0.5\_400\_auto\_auto ranks in the top five for both granularity and topic coverage, and M\_0.01\_0.5\_400\_0.01\_0.01 ranks in the top five for granularity, both models still appear among the bottom five under the WIT metric.

Focusing on the top-ranking model for each evaluation metric, our second key insight from the analysis is that the number of topics ( $K$ ) varies from 25 using the diversity metric to 400 using the granularity and topic coverage metrics. The notion of an unambiguously optimal value for  $K$  that the LDA algorithm ‘discovers’ objectively without researcher input is therefore a myth. Instead, a wide range of feasible topic representations of the corpus exist, with the final choice likely to depend to a large degree on the precise nature of the research question(s) under investigation. Collectively, results in Panel A highlight the sensitivity of LDA topic model outcomes to choice of hyperparameters and evaluation approach, both of which require significant researcher judgement.<sup>7,8</sup>

---

<sup>7</sup>We test Experiment 1 and remaining experiments in the paper by running the LDA configuration with a single random seed. However, we recognise that some differences in evaluation metrics could be due to stochastic variation in model fitting rather than differences in key parameters. To address this concern, we examine the variation in key metrics across 46 independent runs using different random seeds. Specifically, we select two LDA models that represent opposite ends of the performance spectrum in terms of coherence: M\_0.05\_0.5\_35\_0.1\_0.01 (the best-performing model) and M\_0.01\_0.7\_400\_0.005\_0.005 (the worst-performing model). In untabulated results, we find that the evaluation metrics are highly consistent across runs and that the observed differences between the two models are unlikely to be a result of random chance.

<sup>8</sup>We also assess the stability of topic content across random seeds to determine whether differences in topic content across parameter choices exceed stochastic variation. Specifically, we train a broad-based model (M\_0.05\_0.5\_35\_0.1\_0.01) and a narrow-based model (M\_0.01\_0.7\_400\_0.005\_0.005) 46 times each. This yields 1,035 unique model pairs per specification, across which we evaluate topic overlap. For each pair, we calculate the number of topics in Model A that match at least one topic in Model B, and vice versa. Two topics are defined as overlapping if they share at least 6 out of their top 10 keywords. We then scale the number of overlapping topics by the total number of topics in the model. Despite the inherent stochasticity of Latent Dirichlet Allocation (LDA), we observe a high degree of semantic consistency across random seeds: on average, approximately 71% (25 out of 35) of topics in the broad-based model and 67% (267 out of 400) of topics in the narrow-based model are matched across model runs.

Panel B reports correlations in model ranking by evaluation metrics. Correlations in Panel A reveal the key role that evaluation scoring methods can have on model choice. While evaluations using coherence, diversity and WIT are most closely aligned, the metrics are not perfect substitutes with correlations ranging from a high of 0.75 (diversity and WIT) to a low of 0.55 (coherence and WIT for raw scores). Further, these three evaluation metrics display *negative* correlations with granularity and topic coverage that range from  $-0.15$  (coherence and granularity for raw scores) to  $-0.94$  (diversity and topic coverage for raw scores). Results illustrate how metrics such as coherence and diversity favour coarser topic representations whereas granularity and topic coverage favour narrower topic representations. In the absence of any theoretically preferred evaluation method or combination thereof, the final choice rests with the researcher and the nature of the research question, and it is possible that the question may necessitate multiple representations.<sup>9</sup>

#### 4.1.2. Experiment 2: interpretation using visualisation

Inter-topic distance maps visualise the gap between topics in a latent space. These visualisations supplement traditional quantitative evaluation metrics and provide a valuable (subjective) tool for selecting between LDA models. A promising model features well-separated topics, indicating each one is capturing a distinct theme. Evidence of significant overlap suggests redundancy or that the topics are not sufficiently distinct (i.e. undercooked). Our second experiment provides a simple illustration of this complementary visualisation tool in action. We proceed by choosing at random one of the top performing models from Table 2. The model we select has parameters [M\_0.05\_0.5\_35\_auto\_auto] and ranks first on coherence score among all 120 models. Panel A on the left of Figure 1 reproduces the corresponding inter-topic distance map for the 35 topics using pyLDavis.

Topics 8, 19, 20, and 28 highlighted in red in Panel A display significant overlap. The accompanying keyword table provides the top 10 keywords for these topics. Keyword lists include common unigrams such as ‘banking’, ‘loan’, and ‘stage’, making the four topics hard to interpret and distinguish. The visualisation confirms that a high coherence score does not automatically guarantee topics that are interpretationally distinctive. One strategy open to researchers where significant overlaps occur in a high scoring model is to consider alternative models with parameter combinations that generate fewer overlapping clusters. Panel B on the right of Figure 1 contains the inter-topic distance map for an alternative model with 35 topics that ranks fifteenth by coherence score [M\_0.05\_0.5\_35\_0.1\_0.01]. Despite its lower evaluation score, the topics in this model display less overlap and may therefore be practically more interpretable. An alternative strategy is to aggregate similar topics by summing probabilities across the overlapping topics (Hu et al. 2014). For example, one could sum probabilities for topics 8, 19, 20, and 28 in Panel A to form a combined topic that nests the four overlapping topics. Ultimately, however, the final decision over which approach yields the best topic representation will require researchers to exercise judgement based on the research question and their domain expertise.

#### 4.1.3. Experiment 3: multiword expressions

This experiment examines the impact of MWEs on topic representations. We assess impact by first creating a new version of our baseline corpus that captures two – and three-word phrases

---

<sup>9</sup>Internet Appendix Table A9 reports results from our replication analysis using Item 1A (Risk factors disclosures) from US 10-K filings. We follow the same experimental design and evaluation framework. Findings from this supplementary analysis show that our core results regarding LDA model sensitivity to hyperparameter choices and evaluation approaches generalise beyond the UK setting.

Table 2: Results for Experiment 1 exploring the effect of hyperparameter tuning and different topic evaluation strategies.

Panel A: Model parameters by evaluation strategy

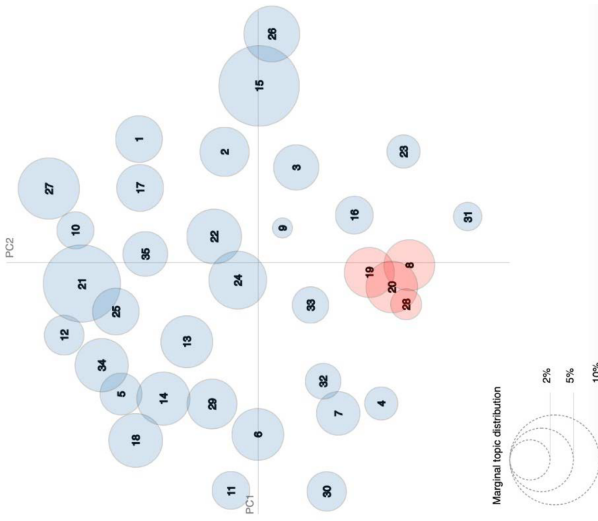
Ranking	Evaluation method used for ranking topic models:					Topic coverage
	Coherence	Diversity	Granularity	WIT accuracy		
1	M_0.05_0.5_35_auto_auto	G_0.05_0.5_25_0.5_0.01	M_0.01_0.5_400_auto_auto	G_0.05_0.7_50_0.5_0.01	M_0.01_0.7_400_auto_auto	
2	M_0.05_0.5_25_auto_auto	G_0.05_0.5_25_0.1_0.1	M_0.01_0.5_200_auto_auto	G_0.05_0.7_25_auto_auto	M_0.01_0.7_400_0.01_0.01	
3	M_0.05_0.7_25_auto_auto	G_0.05_0.5_35_0.5_0.01	M_0.01_0.5_100_auto_auto	G_0.05_0.7_35_0.5_0.01	M_0.01_0.5_400_auto_auto	
4	G_0.05_0.5_25_auto_auto	G_0.05_0.5_25_auto_auto	M_0.01_0.5_200_0.01_0.01	M_0.05_0.7_35_auto_auto	M_0.01_0.5_400_0.005_0.00	
5	M_0.05_0.5_50_auto_auto	M_0.05_0.5_25_0.5_0.01	M_0.01_0.5_400_0.01_0.01	M_0.05_0.7_25_auto_auto	M_0.01_0.7_200_0.01_0.01	
66	M_0.01_0.5_400_0.005_0.01	M_0.01_0.7_400_0.01_0.005	M_0.05_0.7_35_0.5_0.1	M_0.01_0.5_200_0.01_0.005	G_0.05_0.7_25_auto_auto	
67	M_0.01_0.7_400_0.01_0.005	M_0.01_0.7_400_0.005_0.00	M_0.05_0.7_25_auto_auto	M_0.01_0.7_400_0.005_0.00	M_0.05_0.7_25_0.1_0.01	
68	M_0.01_0.7_400_0.01_0.01	M_0.01_0.7_400_0.005_0.01	M_0.05_0.7_25_0.5_0.1	M_0.01_0.5_400_0.01_0.01	M_0.05_0.5_25_0.1_0.1	
69	M_0.01_0.7_400_0.005_0.01	M_0.01_0.7_400_0.01_0.01	M_0.05_0.7_25_0.1_0.1	M_0.01_0.5_400_auto_auto	M_0.05_0.5_25_0.5_0.1	
70	M_0.01_0.7_400_0.005_0.00	M_0.01_0.7_400_auto_auto	M_0.05_0.7_25_0.5_0.01	M_0.01_0.5_400_0.005_0.01	G_0.05_0.5_25_auto_auto	

Panel B: Pearson (Spearman) correlations between evaluation strategies used for ranking topic models

	Coherence	Diversity	Granularity	WIT accuracy	Topic coverage
Coherence	1.000				
Diversity	0.725 (0.688)	1.000			
Granularity	-0.150 (-0.269)	-0.120 (-0.233)	1.000		
WIT accuracy	0.566 (0.554)	0.754 (0.754)	-0.466 (-0.522)	1.000	
Topic coverage	-0.779 (-0.789)	-0.935 (-0.928)	0.206 (0.346)	-0.681 (-0.705)	1.000

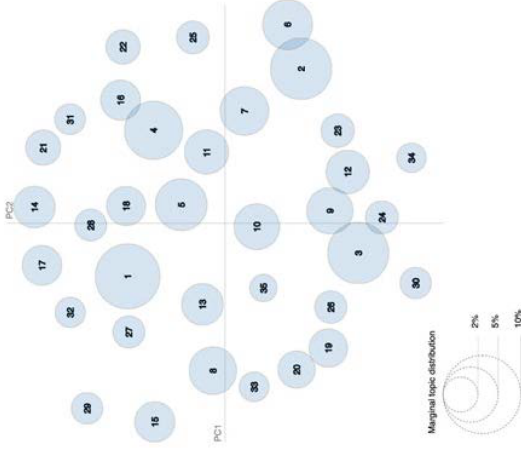
Panel A reports the top five and bottom five models in rank order of evaluation metrics: coherence score, diversity, granularity, WIT accuracy, and topic coverage. Coherence score evaluates the semantic consistency of topics by assessing how well the words grouped together in a topic make sense when they appear together in the actual text. We use the  $C_v$  coherence scores for gensim models. Diversity score evaluates the variety and distinctiveness of the topics generated by a topic model by calculating the ratio of the number of unique top words to the number of all top words. Granularity is calculated as one minus the mean document frequency of topic keywords divided by the total number of documents. This measure favours granular topics that capture keywords appearing less frequently across documents. Word intrusion task evaluates the interpretability of topics by asking evaluators to identify an intruder word mixed with top words from a topic. Topic coverage evaluates how many identified topics are matched with the predefined topics in Table 7. We use GPT4o to measure Word intrusion task (WIT) accuracy and Topic coverage. Models are named for the following parameter choices: training algorithms (l), infrequent word filtering (i), frequent word filtering (f), topic numbers (k), alpha (a), and beta (b): t\_i\_f\_k\_a\_b. The initial training set contains 120 models, 60 of which employ the mallet learning algorithm (Gibbs sampling) and 60 or which use the gensim learning algorithm (variational inference). We drop 50 models (all gensim) from this initial set as they include at least one unidentified topic whose word distribution is a uniform distribution. Panel B reports Spearman (Pearson) correlation coefficients between evaluation metrics for 70 trained models. Spearman correlations are for model rankings. Pearson correlations are for evaluation scores.

Panel A: Inter-topic distance map of M\_0.05\_0.5\_35\_auto\_auto



Topic	Top 10 unigrams
Topic 8	ps, bn, loans, total, impairment, bank, home, wholesale, stage, retail
Topic 19	ps, stage, banking, total, exposures, loans, stress, bank, lending, ratio
Topic 20	ps, lending, loans, total, stage, loan, bank, collateral, impairment, advances
Topic 28	stage, banking, loans, total, collateral, stress, exposures, impairment, clients, advances

Panel B: Inter-topic distance map of M\_0.05\_0.5\_35\_0.1\_0.01



Topic	Top 10 unigrams
Topic 3	stage, loans, lending, total, loan, collateral, ifrs, advances, stress, ratio
Topic 9	banking, stage, total, loans, exposures, stress, clients, collateral, country, chartered
Topic 12	ps, banking, total, exposures, equity, sheet, retail, bank, income, stress
Topic 24	ps, bn, loans, home, impairment, total, wholesale, bank, retail, leverage

Figure 1. Inter-topic distance maps for Experiment 2.

as an additional step in the text preprocessing pipeline. We automate the process of identifying n-grams using genism's 'phrases' function, which follows Mikolov's et al. (2013) approach to generating MWEs based on the collocation of two consecutive words.<sup>10</sup> We recursively apply this method to our corpus to generate bigrams and trigrams. Next, we apply spaCy POS tagging to help retain meaningful collocations by filtering on interpretable language structures (Bhalla and Klmcikova 2019). Specifically, we retain bigrams with the structure (Noun, Noun) or (Adjective, Noun), and trigrams with the structure (Adjective/Noun, Anything, Adjective/Noun). We replace white spaces with underscores ( ) in the set of filtered n-grams to convert them to unigrams. For example, instances of the bigram 'risk management' in the original corpus are replaced by 'risk\_management' in the updated corpus. This process generates 533 MWEs that we represent as unigrams. The list of MWEs is available in the Appendix and a cursory review suggests they reflect meaningful phrases in risk disclosures. Finally, we remove high frequency tokens (appearing in >50 percent of documents) and low frequency tokens (appearing in < five percent of documents).

Our experimental design evaluates whether inclusion of MWEs affects the number of topics the LDA model discovers and the interpretation of discovered topics. We start by identifying a baseline model. We follow Experiment 2 and use the top-ranking model from Experiment 1 by coherence score [M\_0.05\_0.5\_35\_auto\_auto]. Next, we train six new versions of this model using the new MWE corpus, allowing the number of topics to vary ( $K=25, 35, 50, 100, 200, 400$ ) while fixing all other parameters.<sup>11</sup> Finally, we use three approaches to assess whether inclusion of MWEs affects the topic representation: we compare evaluation scores for the six new models with the evaluation score for the baseline model; we compute the proportion of topics containing at least one MWE; and we compare the semantic similarity of the 35-topic baseline model with the comparable 35-topic version estimated with the MWE corpus. Table 3 presents a summary of the results.

The first row of Table 3 replicates evaluation scores for our baseline model from Table 2, while the remaining rows report comparable scores for our six new models trained on the MWE corpus. With the exception of coherence, where the baseline model continues to display the highest score, all other metrics in columns 3–6 display higher evaluation scores for at least one of the six MWE models. Overall, 19 of the 30 evaluation scores for models estimated with the MWE corpus (63%) exceed the relevant baseline model. The scale of the improvement in scores is also material, ranging from 4% for granularity to 84% for topic coverage. Note, however, that all five evaluation metrics also display a materially *lower* score for at least one MWE model, suggesting that the performance benefits of including MWEs are not universally positive. Results nevertheless support our core thesis that the decision on whether or not work with n-grams is likely to matter, and that the final choice regarding the optimal approach is likely to be context-specific.

As further evidence on the potential importance of MWEs for LDA topic representations, the final column of Table 3 reports the fraction of topics with at least one MWE in the top 10 keywords. More than 50 percent of topics for all six models contain at least one MWE among the top

---

<sup>10</sup>Our automated approach is motivated primarily on the grounds of cost minimisation given the scale and scope of our experiments. We encourage researchers to supplement automated approaches with manual intervention that captures domain expertise. As automated approaches are expected to generate more noise, we view our experiment as providing a lower bound on the potential impact of including MWEs.

<sup>11</sup>Our baseline model is optimised for unigrams not MWEs. Optimising the LDA model on an input corpus containing MWEs is likely to produce a different combination of hyperparameters, including a change to the optimal number of topics.

Table 3. Summary results for Experiment 3 examining the impact of multiword expressions (MWEs) on LDA topic model representation.

Models	Evaluation metric:						Percentage of topics including at least one MWE
	Coherence	Diversity	Granularity	WIT accuracy	Topic coverage		
Baseline unigram model							
M_0.05_0.5_35_auto_auto	0.648	0.660	0.681	0.571	0.528		NA
MWE models:							
M_0.05_0.5_25_auto_auto_MWE	0.629	0.744	0.684	0.640	0.472		56.0%
M_0.05_0.5_35_auto_auto_MWE	0.626	0.666	0.690	0.714	0.500		51.4%
M_0.05_0.5_50_auto_auto_MWE	0.602	0.598	0.696	0.600	0.583		60.0%
M_0.05_0.5_100_auto_auto_MWE	0.545	0.461	0.702	0.450	0.750		63.0%
M_0.05_0.5_200_auto_auto_MWE	0.497	0.335	0.707	0.370	0.888		64.0%
M_0.05_0.5_400_auto_auto_MWE	0.473	0.252	0.711	0.335	0.972		65.3%

Table 3 provides the model evaluation metrics for the baseline unigram model (M\_0.05\_0.5\_35\_auto\_auto) and six MWE models trained for Experiment 3. Coherence score evaluates the semantic consistency of topics by assessing how well the words grouped together in a topic make sense when they appear together in the actual text. We use the C\_v coherence scores for gensim models. Diversity score evaluates the variety and distinctiveness of the topics generated by a topic model by calculating the ratio of the number of unique top words to the number of all top words. Granularity is calculated as one minus the mean document frequency of topic keywords divided by the total number of documents. This measure favours granular topics that capture keywords appearing less frequently across documents. Word intrusion task evaluates the interpretability of topics by asking evaluators to identify an intruder word mixed with top words from a topic. Topic coverage evaluates how many identified topics are matched with the predefined topics in Table 7. We use GPT4o to measure Word intrusion task (WIT) accuracy and Topic coverage.

10 top keywords. Findings provide further evidence that restricting LDA models to consider only unigrams is likely to impact topic properties.

Our final test assesses whether inclusion of MWEs leads to a material change in topic labels. Specifically, we hold all model parameters constant at [M\_0.05\_0.5\_35\_auto\_auto] and test whether topic labels change conditional on the treatment of MWEs. We interpret evidence of a material change in labels as an indication that the treatment of MWEs matters for interpretability. We assign topic labels using GPT4o to minimise implementation costs and the risk of bias from human coders in labelling process.<sup>12</sup> Finally, we use GPT to identify semantically similar topic labels between the 35-topic baseline model and the equivalent 35-topic model estimated with the MWE corpus. Findings (not tabulated for parsimony) reveal that GPT clusters 26 topics (74%) from the baseline model and 27 topics (77%) from the MWE model together as being semantically similar based on their labels. Eight semantically different topics (23%) are nevertheless evident in the MWE model, while nine topics (26%) in the baseline model disappear. While some of these differences may be attributable to the inherent stochastic nature of topic modelling (e.g. rerunning the same model with a different random seed), the emergence of new multiword topics provides additional evidence that explicitly capturing MWEs can alter topic representations in material ways.

#### 4.1.4. Experiment 4: stemming

Stemming and lemmatisation help to reduce dimensionality in the term document matrix and potentially improve interpretability. We follow the same approach as in Experiment 3 but with stemming replacing MWE construction in the preprocessing pipeline. Specifically, use the top ranking model from Experiment 1 by coherence score [M\_0.05\_0.5\_35\_auto\_auto] as our baseline and then we train the same six LDA models as in Experiment 3 ( $K = 25, 35, 50, 100, 200, 400$ ) on a new version of the baseline corpus where we apply the Porter stemmer (NLTK Porter-Stemmer) as an additional step in the text preprocessing pipeline. We then repeat the same analysis as described in Experiment 3 after substituting the stemmed corpus for the MWE corpus. The baseline corpus includes 2.2 million tokens from 3,029 unique words, whereas the stemmed corpus contains 1.5 million tokens from 1,709 unique words. Table 4 summarises the results.

Results provide a mixed picture. Scores for coherence and WIT are lower than the baseline model across all six versions using the stemmed corpus. In addition, holding the number of topics constant at 35, evaluation scores for all metrics with the exception of granularity are lower for the model using the stemmed corpus than for the baseline model estimated with the unstemmed corpus. The decrease in performance suggests that stemming may merge semantically distinct words, leading to noisier topics. On the other hand, 11 of the 30 evaluation scores for models estimated with the stemmed corpus (37%) exceed the relevant baseline model. Results therefore provide some evidence that stemming can increase model performance, although the potential benefits appear to be less pronounced than the case for MWEs. The overall message is nevertheless consistent with that for MWEs; the decision whether or not to include stemming in the text preprocessing pipeline is likely to have a material impact on LDA topic representation. Researchers therefore need to approach the choice with care and transparency. Conducting robustness tests with and without stemming may also be helpful given the mixed results.

---

<sup>12</sup>Here, as in all other uses of GPT throughout the paper, we use the OpenAI API available in python. We enhance the replicability of our analysis by setting the temperature equal to zero. We use the latest available version of GPT4o (gpt-4o-2024-08-06) at the time of the analysis.

Table 4. Summary results for Experiment 4 examining impact of stemming on LDA topic model representation.

Models	Evaluation metric:				
	Coherence	Diversity	Granularity	WIT accuracy	Topic coverage
Baseline unigram model					
M_0.05_0.5_35_auto_auto	0.648	0.660	0.681	0.571	0.528
Stemmed models:					
M_0.05_0.5_25_auto_auto_stem	0.625	0.716	0.679	0.560	0.333
M_0.05_0.5_35_auto_auto_stem	0.621	0.646	0.684	0.429	0.472
M_0.05_0.5_50_auto_auto_stem	0.606	0.578	0.683	0.480	0.583
M_0.05_0.5_100_auto_auto_stem	0.553	0.417	0.678	0.280	0.750
M_0.05_0.5_200_auto_auto_stem	0.491	0.296	0.687	0.315	0.750
M_0.05_0.5_400_auto_auto_stem	0.459	0.196	0.687	0.245	0.806

Table 4 provides the model evaluation metrics for the baseline unigram model (M\_0.05\_0.5\_35\_auto\_auto) and six stemming models trained for Experiment 4. Coherence score evaluates the semantic consistency of topics by assessing how well the words grouped together in a topic make sense when they appear together in the actual text. We use the C\_v coherence scores for gensim models. Diversity score evaluates the variety and distinctiveness of the topics generated by a topic model by calculating the ratio of the number of unique top words to the number of all top words. Granularity is calculated as one minus the mean document frequency of topic keywords divided by the total number of documents. This measure favours granular topics that capture keywords appearing less frequently across documents. Word intrusion task evaluates the interpretability of topics by asking evaluators to identify an intruder word mixed with top words from a topic. Topic coverage evaluates how many of the identified topics are matched with the predefined topics in Table 7. We use GPT4o to measure Word intrusion task (WIT) accuracy and Topic coverage.

We also assess whether stemming leads to a material change in topic labels. Our test follows the approach described in Experiment 3. We hold all model parameters constant at [M\_0.05\_0.5\_35\_auto\_auto] and test whether topic labels change conditional on stemming strategy. We assign topic labels using GPT4o and we also use GPT to identify semantically similar topic labels between the 35-topic baseline model and the equivalent 35-topic model estimated with the stemmed corpus. We interpret evidence of a material change in labels as an indication that the treatment of stemming matters for interpretability. Findings (not tabulated) reveal that GPT clusters 31 topics (89%) from the baseline model and 32 topics (91%) from the stemmed model together as being semantically similar based on their labels. Three semantically different topics (9%) are nevertheless evident in the stemmed model, while four topics (11%) in the baseline model disappear. We note that some portion of the observed differences may be due to the stochastic nature of LDA. Nevertheless, the consistent emergence (or disappearance) of specific topics when stemming is (or is not) applied suggests that stemming decisions can influence topic labels, although the effect appears weaker than that documented for MWEs. Accordingly, we encourage researchers to explore the effects in their own corpus rather than relying mechanically on our findings.

#### 4.1.5. Experiment 5: labelling strategy

Our final experiment evaluates the sensitivity of topic labels to the labelling strategy. The labelling process is independent of the LDA algorithm and is entirely reliant on researcher judgement. Our test involves selecting a baseline LDA model and then assessing if and how labels vary for the following four labelling strategies: (1) manual labelling by one researcher working independently; (2) manual labelling by three researchers working as a group; (3) GPT4o labelling using a naïve prompt lacking domain expertise; and (4) GPT4o labelling with a chain-of-thought (CoT) prompt that incorporates significant domain expertise. Two of the co-authors working together developed the CoT prompt

through several iterations based on their knowledge of the UK risk reporting guidance and risk categories. Table 5 reproduces the naïve and CoT prompts.

Our baseline LDA model is [M\_0.05\_0.5\_35\_0.1\_0.01]. We select this model because we have FRC input on manually assigned labels from a separate project.<sup>13</sup> We assess whether labels assigned using the four methods differ materially, and whether method (3) that limits researcher discretion generates unambiguously more informative labels than the remaining three methods that permit subjectivity in varying degrees.

Table 6 reports the results. Columns 2 and 3 contain manually assigned labels. A number of labels vary materially between those assigned by a single co-author and those assigned by three co-authors working as a team. Examples include topic 4 (*Scenario analysis vs. Banking*), topic 12 (*Insurance liability management vs. Insurance*), topic 14 (*Auditing & compliance vs. Reporting & audit*), topic 20 (*Legal risk & compliance vs. Regulations*), topic 22 (*Client solutions vs. Undefined*), topic 28 (*Credit management vs. Banking*), topic 32 (*Loan portfolio management vs. Banking*), and topic 33 (*Trading exposures and asset limits vs. Security trading*). Results reveal that different manual labelling strategies generate different topic labels, which in turn may affect inferences and conclusions.

The final two columns in Table 6 contain labels assigned using GPT. Several points are worthy of note. First, GPT labels appear credible judged against the top 10 keywords and the labels assigned by human experts. This is potentially important for granular representations of the topic space (e.g. >100) where human coding may prove prohibitively costly. Second, the labels that GPT assigns tend to provide richer topic descriptions, which may lead to more refined insights and conclusions. Building on this point, we also see material semantic differences between labels that humans assign and those that GPT assigns. For example, while human coders label topic 9 as *Property portfolio management*, GPT CoT assigns the label *Debt covenants & property risks*. Meanwhile, human coders label topic 2 as *Banking* whereas both GPT prompts reference *impairments*; and human coders agree that topic 20 reflects *Risk appetite* whereas both GPT prompts generate labels that reference *cybersecurity* risks directly.

Finally, comparing the GPT naïve and CoT prompts highlights material differences.<sup>14</sup> For example, the CoT label for topic 5 references *culture risks* whereas the naïve label does not. Similarly, the CoT label for topic 15 references *remuneration* risks whereas the naïve label does not. Results demonstrate that prompting strategy can impact topic semantics materially when using LLMs for labelling, and more specifically that building prompts that exploit domain expertise can lead to richer topic descriptions.<sup>15</sup> Our findings provide no support for the view that restricting researcher judgement leads to more reliable labelling. Overall, results from Experiment 5 highlight the sensitivity of topic labels and associated economic insights

<sup>13</sup>Financial Reporting Council project FRC2023-0131, *Analysing trends in annual report language and content*. Project tendered 24 July 2023, awarded 11 September 2023, completed May 2024.

<sup>14</sup>One caveat to this analysis is that GPT is a stochastic model meaning there is a degree of randomness even when the temperature is set to zero (Atil et al. 2024). To assess this, we rerun the prompt to provide topic labels with 30 different random seeds. Manually inspecting the proposed labels reveals labels are very similar semantically. GPT therefore appears to produce consistent labelling outcomes.

<sup>15</sup>Note that there are limitations to applying GPT (or other LLMs) to label topics. Although pre-training equips LLMs with vast amounts of general knowledge, they may fail to label topics where interpretation requires specialised domain-specific knowledge and insights (Guo et al. 2023). A further concern is hallucination (Li et al. 2024, Mu et al. 2024a). LLMs may invent plausible-sounding labels even when the underlying keyword list is semantically incoherent. Therefore, it is important for researchers to consider LLM-generated labels as candidates that need to be carefully reviewed by human experts before they are accepted as final.

Table 5. GPT prompts for automated LDA topic model labelling.

---

*Panel A: Naïve prompt*

---

You are provided with the results of a topic modelling analysis.  
 You are provided with the index of the topic (e.g. Topic1) and top 10 words for the topic.  
 Please provide a label which best describes the topic and the rationale for the label.  
 Return only the label and rationale for the topic.  
 Do not use quotation marks in the label or rationale.  
 Provide your label and rationale as a string separated by a semicolon like in the following example: [Label; Rationale]

---

*Panel B: Chain-of-thought approach*

---

For context, in a research project, a corpus of ‘Principal Risks and Uncertainties’ sections from UK annual reports has been created.

An LDA topic model has then been constructed based on the text. The analysis generated 35 topics.

You are a research associate who is tasked with interpreting the output of topic models generated from corporate disclosures.

Your objective is to provide a label which best represents the semantic meaning of the topic.

Your goal is to review these keywords, generate specific labels for each topic, and ensure that the labels are mutually exclusive.

You are provided with the top 10 words for each of the 35 topics.

Instructions:

1. Read and Analyse Keywords:

- (a) Carefully read the list of keywords for each of the 35 topics.
- (b) Identify semantic links between the keywords within each topic.

2. Generate Initial Labels:

- (a) Based on your analysis, generate a specific and descriptive label for each topic.
- (b) Labels should reflect the specific nature of the risks and uncertainties rather than generic terms like ‘risk management’ or ‘risks.’

3. Ensure Mutually Exclusive Labels:

- (a) Compare topics with similar labels.
- (b) Identify subtle differences between the topics.
- (c) Modify the labels to ensure each one is unique and mutually exclusive.

Example:

Input word lists

Topic 1: [‘regulation’, ‘compliance’, ‘legal’, ‘policy’, ‘law’, ‘rules’, ‘regulatory’, ‘governance’, ‘legislation’, ‘audit’]

Topic 2: [‘market’, ‘competition’, ‘demand’, ‘consumer’, ‘price’, ‘sales’, ‘industry’, ‘growth’, ‘revenue’, ‘trend’]

Topic 1:

Label: Regulatory Compliance Risks

Rationale: The keywords are centred around legal and regulatory aspects, indicating risks associated with compliance with laws and regulations.

Topic 2:

Label: Market Competition Risks

Rationale: The keywords suggest risks related to market dynamics, competition, and consumer behaviour affecting the company’s performance.

Return only the labels and rationales for each topic.

Provide your labels and rationales in JSON format like in the following example ...

---

This table provides the prompt used for automated labelling by GPT. We utilise the GPT Python API to automate the process by providing the above prompt as a system message and the actual topic keywords as a user message. Since GPT’s performance may decline with longer input (e.g. asking 10 questions at a time versus all questions at once), we divide the task into chunks of 10 topics and then combine the responses. As chunking the task into smaller parts may limit the use of the chain-of-thought (CoT) approach to ensure mutual exclusivity between topics across different chunks, we address this in the final step by asking GPT to review the aggregated topic labels and revise them to ensure they capture mutually exclusive themes.

Table 6. Summary of results for Experiment 5 comparing findings for human and LLM labelling strategies.

Topic	Top 10 LDA keywords	Human labelling			GPT labelling		
		Single Reviewer	Team	Naive	Naive	Chain-of-thought	
1	ps, pension, scheme, funding, benefit, schemes, defined, liabilities, deficit, net	Pension	Pension	Pension schemes & funding	Pension risks	Pension scheme funding risks	
2	loans, total, impaired, advances, lending, society, retail, impairment, loan, funding	Banking	Banking	Retail lending & loan impairment	Retail lending & loan impairment risks	Retail lending & loan impairment risks	
3	department, measures, implementation, implemented, accordance, approved, accounting, units, principles, fraud	Regulatory compliance & implementation	Operations & compliance	Compliance & implementation of accounting measures	Accounting fraud risk management	Accounting fraud risk management	
4	stage, bank, ifrs, scenarios, scenario, total, gross, models, appetite, audited	Scenario analysis	Banking	Banking risk assessment & IFRS compliance	Banking risk scenario analysis	Banking risk scenario analysis	
5	people, deliver, delivery, leadership, talent, culture, teams, improve, progress, programmes	Human resource management	Human resource management	Leadership & talent development	Talent management & culture risks	Talent management & culture risks	
6	climate, sustainability, carbon, emerging, esg, emissions, transition, energy, impacts, physical	Climate change & sustainability	Climate change	Climate change & sustainability	Climate change & sustainability risks	Climate change & sustainability risks	
7	energy, power, network, fuel, electricity, prices, infrastructure, national, government, generation	Energy	Energy	Energy infrastructure & policy	Energy infrastructure risks	Energy infrastructure risks	
8	businesses, acquisition, acquisitions, integration, acquired, diligence, competition, profitability, organic, adversely	M&A	M&A	Business acquisitions & integration	Mergers & acquisitions risks	Mergers & acquisitions risks	
9	property, debt, properties, portfolio, covenants, asset, income, low, rental, estate	Property Portfolio Management	Property portfolio management	Real estate & property management	Debt covenants & property risks	Debt covenants & property risks	
10	banking, ps, appetite, stress, funding, loans, exposures, lending, total, portfolio	Banking	Banking	Banking & loan portfolio management	Banking stress testing & loan risks	Banking stress testing & loan risks	
11	project, contract, projects, divisional, delivery, monthly, health, commercial, people, suppliers	Project execution management	Project management	Project management & contract delivery	Project delivery & contract risks	Project delivery & contract risks	
12	insurance, claims, prudential, solvency, life, underwriting, liabilities, losses, limits, pricing	Insurance liability management	Insurance	Insurance & risk management	Insurance underwriting & liability risks	Insurance underwriting & liability risks	
13	firm, loans, portfolio, net, securities, trading, total, losses, fair, derivatives	Financial portfolio management	Financial portfolio management	Financial trading & portfolio management	Trading securities & derivatives risks	Trading securities & derivatives risks	

(Continued)

Table 6. Continued.

Topic	Top 10 LDA keywords	Human labelling			GPT labelling	
		Single Reviewer	Team	Naive	Chain-of-thought	
14	auditor, auditors, accounting, independence, matters, meetings, reviewing, independent, function, satisfied	Auditing & compliance	Reporting & audit	Auditing & accounting independence	Auditor independence & review risks	
15	director, chief, remuneration, officer, chairman, meetings, meeting, members, shareholders, code	Board & corporate governance	Board & corporate governance	Executive leadership & shareholder relations	Board governance & remuneration risks	
16	tax, income, total, fair, share, net, profit, shares, liabilities, note	Taxation	Taxation	Taxation & financial reporting	Tax compliance & financial reporting risks	
17	eu, european, ireland, restructuring, implementation, developments, sector, measures, businesses, resolution	EU regulations	EU	European union regulatory measures & business restructuring	European union restructuring risks	
18	tax, global, laws, local, anti, international, political, conduct, countries, bribery	Tax compliance	International tax compliance	Global taxation & anti-bribery laws	Global anti-bribery & corruption risks	
19	manager, portfolio, investments, shares, providers, shareholders, trust, share, income, meeting	Investment management	Investment management	Portfolio management & shareholder trust	Investment portfolio management risks	
20	viability, statement, concern, reasonable, scenarios, prospects, assessed, longer, fall, brexit	Viability analysis	Viability	Business viability & risk assessment	Viability statement & risk assessment	
21	adversely, condition, laws, affected, united, claims, materially, government, states, substantial	Legal risk & compliance	Regulations	Legal and regulatory risks in the US	Lawsuits & legal claims risks	
22	currency, foreign, debt, instruments, currencies, hedge, hedging, denominated, borrowings, fixed	Foreign currency	Foreign currency	Foreign currency & debt hedging	Foreign currency hedging risks	
23	clients, client, software, platform, revenues, solutions, content, description, recruitment, retain	Client solutions	Undefined	Client management & software solutions	Client relationship & revenue risks	
24	covid, pandemic, cyber, measures, government, disruption, response, restrictions, emerging, brexit	External disruptions	External disruptions	Pandemic & cybersecurity disruptions	Pandemic response & business disruption risks	
25	intellectual, property, sales, research, commercial, manufacturing, rights, healthcare, ip, marketing	Intellectual property	Intellectual property	Intellectual property & commercialisation in healthcare	Intellectual property & rights risks	
26	health, water, environmental, land, site, sites, local, construction, government, people	Environmental, health & safety	Environmental, health & safety	Environmental health & land management	Environmental & management risks	

to the choice of labelling strategy. Labelling is a critical step in the topic modelling process that researchers must approach with appropriate thought and care.

## 4.2 Summary and implications

Results from our five experiments reveal the impact of researcher choice, whether implicit or explicit, on the LDA topic modelling process. Contrary to popular claims, discovering topics using the LDA algorithm is a nuanced process that requires researchers to exercise careful and informed judgement, and provide full transparency on all aspects of the modelling pipeline. In this respect, the process is similar in many ways to a manual coding exercise. Both approaches are capable of generating important insights when they are applied correctly; but both approaches are also capable of providing spurious findings when the implementation lacks sufficient care, understanding, and transparency. The added concern with quantitative methods like LDA is that there is greater scope for obscuring researcher discretion and its impact behind a veil of pseudo-rigour (El Haj et al. 2019).

Unfortunately, findings from our experiments do not provide simple, mechanical solutions for researchers to apply as part of an ‘analysis-by-numbers’ approach to text mining. As a result, the task of discovering topics in a corpus of financial documents is a process of ‘art’ as much as ‘science’. This challenge mirrors broader issues in empirical research, where even seemingly straightforward analyses can produce varying results across different researchers working with identical data (Menkveld et al. 2024). Our experiments do, however, shine a light on some of the key decisions that researchers need to address and explain when applying LDA. We draw on these insights to propose a checklist for authors and readers to follow when implementing and interpreting LDA topic modelling. Table A1 in the Appendix presents our checklist. Failure by researchers to reference and explain each item in the checklist casts a cloud over the reliability of results and conclusions in our view.

## 5. Beyond LDA: leveraging language models

While our primary analysis focuses on LDA, recent advances in transformer architectures have enabled language-model-based topic modelling approaches that use semantic text embeddings rather than bag-of-words counts.<sup>16</sup> Transformer/embedding-based approaches such as BERTopic (Grootendorst 2022) represent a new generation of topic modelling that leverages transformer architectures (Vaswani et al. 2017) to generate semantic text embeddings that contextualise words and provide meaning.<sup>17</sup> The semantic nature of approaches such as BERTopic represents a significant conceptual advancement over the BoW approach and probabilistic nature of LDA.

In these new generation models, documents are first transformed into high-dimensional embedding vectors using models such as BERT (Devlin et al. 2019) or other text transformers (Reimers and Gurevych 2019). The resulting high-dimensional embeddings are then subjected to a dimensionality reduction process such as UMAP (McInnes et al. 2018), followed by density-based clustering using an algorithm such as HDBSCAN (Campbell et al. 2014) to identify coherent groups of semantically similar documents. Unlike LDA, which assigns probabilities of multiple topics to each document, BERTopic assigns each document to a single cluster and therefore a single topic (Grootendorst

---

<sup>16</sup>We use ‘language-model-based’ and ‘embedding-based’ interchangeably to refer to transformer encoder-based embedding pipelines (i.e., methods that compute semantic document embeddings with transformer models, then apply dimensionality reduction and clustering).

<sup>17</sup>Text embeddings convert text into numerical vectors that allows algorithms to understand and process the semantic meaning of words and phrases. These vectors, often termed embedding vectors, represent text in a high-dimensional space where similar text segments are positioned closer together, facilitating semantic comparisons and analysis that extend beyond simple keyword matching.

2022). However, this conceptual limitation may be circumvented for topic-rich documents such as annual reports, earnings announcements, analyst reports, etc. by segmenting documents into sentences or paragraphs before the embedding step to permit multiple density clusters to occur within a document. Each density cluster is then named by extracting distinguishing terms (words and phrases) as topic representations using a ranking process like class-based TF-IDF (c-TF-IDF). As a final step, these topic representations may then be further enhanced or replaced entirely using a generative LLM (e.g. GPT).

The appeal of embedding-based approaches to modelling topics lies in their potential to capture more subtle semantic relationships and contextual understanding inherent in the text, yielding topics that are more coherent or interpretable than those derived from count-based models such as LDA (Bianchi et al. 2021, Grootendorst 2022, Wu et al. 2024). El-Haj et al. (2020) discuss the dearth of semantic approaches to text processing in accounting and finance research, and the limitations this imposes on analyses and insights. Methods such as BERTopic therefore offer significant potential for our discipline. A small group of papers in accounting and finance use BERT but its application to topic modelling remains unexplored in the domain (Siano and Wysocki 2021, Bingler et al. 2022, Huang et al. 2023, Koelbel et al. 2024).

The key question in the context of our study is whether advanced language-model techniques for modelling topics reduce or remove the need for researcher judgement inherent in LDA. The answer is no. Instead, these methods introduce a new suite of methodological decision points that can influence the resultant thematic representations. In the case of BERTopic, these critical choices include (but are not limited to): (a) choice of the foundational embedding model, as different models are trained on diverse corpora and capture varying semantic nuances (Reimers and Gurevych 2019, Jehnen et al. 2025); (b) parameterisation of the dimensionality reduction stage (e.g. parameters such as `n_components`, `min_dist`, and `n_neighbors` in UMAP) (McInnes et al. 2018); and (c) settings for the clustering algorithm (e.g. `min_cluster_size` and `min_samples` in HDBSCAN) (Campbell et al. 2014). Each of these decisions, like the preprocessing steps and hyperparameter tuning choices in LDA, require careful consideration and justification as collectively they determine the granularity, coherence, and ultimate interpretation of the discovered topics. Also akin to the LDA case, the importance of these choices can be easily missed, ignored, or obscured by researchers when presenting their findings.

We provide evidence on the sensitivity of BERTopic results to key parameter choices by repeating Experiment 1 (section 4.1). Our focus is on understanding the impact on results of researcher discretion over key parameter choices in the modelling pipeline. To understand the impact on evaluation metric, we tabulate the top and bottom five BERTopic models based on rankings of coherence score, diversity, granularity, WIT accuracy, and topic coverage.

## 5.1 BERTopic parameters

Table A4 in the appendix provides a (non-exhaustive) list of key parameter choices when applying BERTopic. We demonstrate the importance of mindful parameter choices by selecting a parsimonious set of values for a subset of key parameters. The five key parameters whose values we flex are the embedding model, number of topics, `n_neighbors` in the UMAP dimensionality reduction algorithm, and `min_cluster_size` and `min_samples` in the HDBSCAN algorithm.<sup>18</sup>

---

<sup>18</sup>Providing pre-trained language model such as BERT with input data that closely matches the text on which they were trained is essential. Since BERT is trained on raw human-readable text, common text preprocessing steps for LDA topic modelling such as removing stop words and applying stemming are unnecessary and may even lead to suboptimal performance caused by deviations from the training text.

We explore two language models: BERT (Devlin et al. 2019), the most widely-known language model introduced by Google, and GTE (Li et al. 2023), a recent and relatively lightweight language model built on BERT with advanced training techniques developed by Alibaba Group.<sup>19</sup> We limit choice of the optimal number of topics ( $K$ ) to four options that deliver a broad topic structure ( $K = \text{Auto}, 23, 35, \text{and } 50$ ), and four options that generate a granular topic structure ( $K = \text{Auto}, 100, 200 \text{ and } 400$ ). For the BERT embedding model, we choose higher values of  $n\_neighbors$  (30, 50),  $min\_cluster\_size$  (50, 100) and  $min\_samples$  (50, 100) to capture broad-level topic structures, and lower values of  $n\_neighbors$  (10, 15),  $min\_cluster\_size$  (25, 50) and  $min\_samples$  (1, 15) to capture granular level structures. For the GTE embedding model, we choose higher values of  $n\_neighbors$  (50, 100),  $min\_cluster\_size$  (200, 400) and  $min\_samples$  (100, 200) to capture broad-level topic structures, and lower values of  $n\_neighbors$  (10, 50),  $min\_cluster\_size$  (100, 200) and  $min\_samples$  (1, 50) to capture granular level structures. These parameter options result in 159 model combinations. The naming convention for our topic models follows the format [*LanguageModel\_Nneighbors\_MinClusterSize\_MinSamples\_TopicNumber*].

Table A5 in the appendix presents the performance of the top and bottom five BERTopic models across five evaluation metrics: coherence, diversity, granularity, WIT accuracy, and topic coverage. The results clearly demonstrate the critical role of parameter selection in shaping the final topic model. Even with modern embedding-based approaches like BERTopic, the process of parameter tuning remains as crucial as it is for traditional BoW methods like LDA. The substantial variance in scores between the best and worst performing models for each metric reinforces this conclusion. For example, coherence scores range from 0.764 down to 0.486, while WIT accuracy reveal even larger variation, from 0.880 to 0.239. This sensitivity shows that choices regarding the embedding model, UMAP neighbours, and HDBSCAN cluster size materially shape the quality and nature of the topic representation.

Consistent with findings for LDA, results also show that no single set of parameters is universally superior. A model specification that performs well on one metric can perform poorly on another. For instance, the [GTE\_10\_400\_10\_auto] model ranks in the top five for both coherence and diversity but fails to place in the top tier for any other metric. In contrast, several models achieve a perfect topic coverage score of 1.000 but do not rank highly on coherence or WIT accuracy. This trade-off across evaluation metrics highlights how the optimal parameter choices depend on the specific goals of the research project.

## 5.2 BERTopic vs LDA

To shed light on the performance of BERTopic relative to LDA, we tabulate the top five BERTopic and LDA topic models based on rankings of coherence score, diversity, granularity, WIT accuracy, and topic coverage. Columns 1 and 2 in Table A6 in the appendix rank the top five BERTopic and LDA topic models, respectively, according to coherence score (Panel A), diversity (Panel B), granularity (Panel C), WIT accuracy (Panel D), and topic coverage (Panel E).

In Panel A, coherence scores of embedding-based models consistently outperform LDA models, with coherence scores reaching up to 0.764 compared to LDA's highest score of 0.649. Results indicate that language-based models generate topics with more semantically consistent keywords, likely due to their ability to capture more nuanced contextual relationships

<sup>19</sup>We do not use generative language models, such as GPT and Gemini. While generative LLMs are powerful tools, their main purpose is to generate new text rather than converting input text into embedding vector for classification or clustering. We use the encoder LLMs, such as BERT and GTE, because they are specifically designed convert input text into embeddings.

between words. In Panel B, diversity scores for both LDA and language-based models are similar (0.832 for LDA vs. 0.812 for BERT-based models). Our evidence suggests that LDA performs marginally better at generating a variety of distinct topics. In Panel C, granularity scores show LDA models outperforming language-based models, with the highest LDA score at 0.730 compared to 0.619 for LLM-based models. Conversely, WIT accuracy of LLM models shows superior performance in Panel D (0.880 for GTE-based models versus 0.700 for LDA), indicating that topics generated by language-based models are more interpretable since GPT finds it easier to identify intruder words. Finally, all top LLM-based models achieve perfect scores of 1.000 for topic coverage in Panel E, confirming their effectiveness in covering predefined topics comprehensively. The highest score for LDA models is 0.972.

Results show how replacing the BoW assumption inherent in LDA with semantic analysis can lead to more coherent and interpretable topic representations for embedding-based models. Nevertheless, our analysis indicates that embedding-based topic models are not universally superior. Embedding-based pipelines demand substantial computational resources and may underperform if the pretrained embedding model is poorly matched to a domain. These new generation models also introduce many hyperparameters (UMAP, HDBSCAN, etc.) where the tuning materially affects outcomes. By contrast, LDA is computationally lighter, captures document-level topic mixtures by construction, and yields document-topic probabilities that are straightforward to use in downstream regression analyses (Blei et al. 2003). Overall, our analysis yields two important takeaways. First, accounting and finance researchers should not treat LDA as the default topic modelling method for their analyses, although in some scenarios the approach may still represent a viable choice. Second, moving to the new generation of embedding-based models does not obviate the need for researchers to make active and informed implementation choices. Table A7 in the appendix provides a more comprehensive comparison of factors that researchers should consider when choosing between LDA and transformer/embedding-based approaches, such as BERTopic. We complement this with a checklist in Table A8 for authors and readers to follow when implementing and interpreting embedding-based topic models.

## 6. Applied evidence for risk reporting

We conclude our analysis with a demonstration of the impact of LDA implementation choices on insights for FTSE350 firms’ annual report risk disclosures. Reflecting FRC concerns over reporting quality, we focus on the themes that risk-related disclosures cover and the propensity for generic ‘boilerplate’ commentary.

### 6.1 Topic coverage and alignment with FRC expectations

We evaluate risk reporting content using two high-scoring LDA models from Experiment 1: a parsimonious 50-topic model (M\_0.05\_0.5\_50\_auto\_auto) and a granular 100-topic model (M\_0.01\_0.7\_100\_auto\_auto).<sup>20</sup> We apply GPT CoT labelling and map outputs from both models against 36 risk categories derived from FRC guidance (Table 7).<sup>21</sup>

<sup>20</sup>In previous experiments, we use the 35-topic model as the baseline. Here, there problem is the FRC guidance specifies 36 topics, meaning a 35-topic model cannot capture all of them. To address this, we select the topic model that allows for at least 36 topics and achieves the highest coherence score – specifically, the M\_0.05\_0.5\_50\_auto\_auto model.

<sup>21</sup>The FRC identify 36 coherent topics plus an additional ‘Other’ category that pools remaining unallocated risks.

Table 7. Externally derived list of principal risks based on regulatory publications and guidance.

Risk theme	Details
Market risks	
Economic	(including inflation, recession, and interest rate changes)
Commodity price	(including fluctuations in prices of raw materials and commodities)
Political & geopolitical	(including risks related to trade restrictions and other economic sanctions)
Systemic	(including pandemic, financial crisis, and Brexit. Does not include climate and energy security, which are treated as a separate risk category under ESG)
Legal & compliance risks	
Regulatory	(including risks related to changes in law, regulations, and guidelines)
Legal disputes	(including risks related to contract, intellectual property, employment law, product liability)
Compliance	(including risks related to industry-level regulations, environmental regulations, data protection, and anti-corruption and bribery laws)
Supply chain risks	
Supply chain disruptions	(disruptions, delays, and failures in the supply chain, as well as risks related to inventory management and logistics)
Supplier due diligence	(including risks related to human rights and environmental issues of suppliers)
Human resources risks	
Cultural	(risks related to the organisation's work culture, including employee engagement, values alignment, and workplace environment)
Talent retention	(including risks related to retaining skilled and talented employees)
Health & Safety	(including risks related to accidents, injuries, and illnesses among employees)
Human rights	(freedoms of individuals within and impacted by the organisation, including employee diversity and inclusion, non-discrimination, and ethical labour practices)
Operational risks	
Production	(including risks related to product or service failures, quality control, and production capacity)
Natural disaster & catastrophe	(including risks related to service failure arising from earthquakes, hurricanes, floods, and other natural disasters)
Financial risks	
Liquidity risks	(including risks related to cash flow, funding, and access to capital)
Credit	(including risks related to borrower default, creditworthiness, and credit ratings)
Interest rate	(including risks related to changes in interest rates, inflation, and monetary policy)
Foreign currency	(including risks related to changes in foreign currency exchange rates)
Retirement benefit	(including risks related to the management and funding of employee retirement and pension plans)
Tax	(risks related to tax regulations and enforcement influencing financial results)
Strategic risks	
Business model	(including risks related to industry and business model)
Mergers & acquisitions	(including risks related to integration, cultural fit, and financial performance)
Market entry	(including risks related to market research, competition, and regulatory requirements)
Partner relationship	(including risks related to developing a strong network with customers and suppliers)
Technology risks	
Cybersecurity threats	(including risks related to data breaches, hacking, and cyber-attacks)

*(Continued)*

Table 7. Continued.

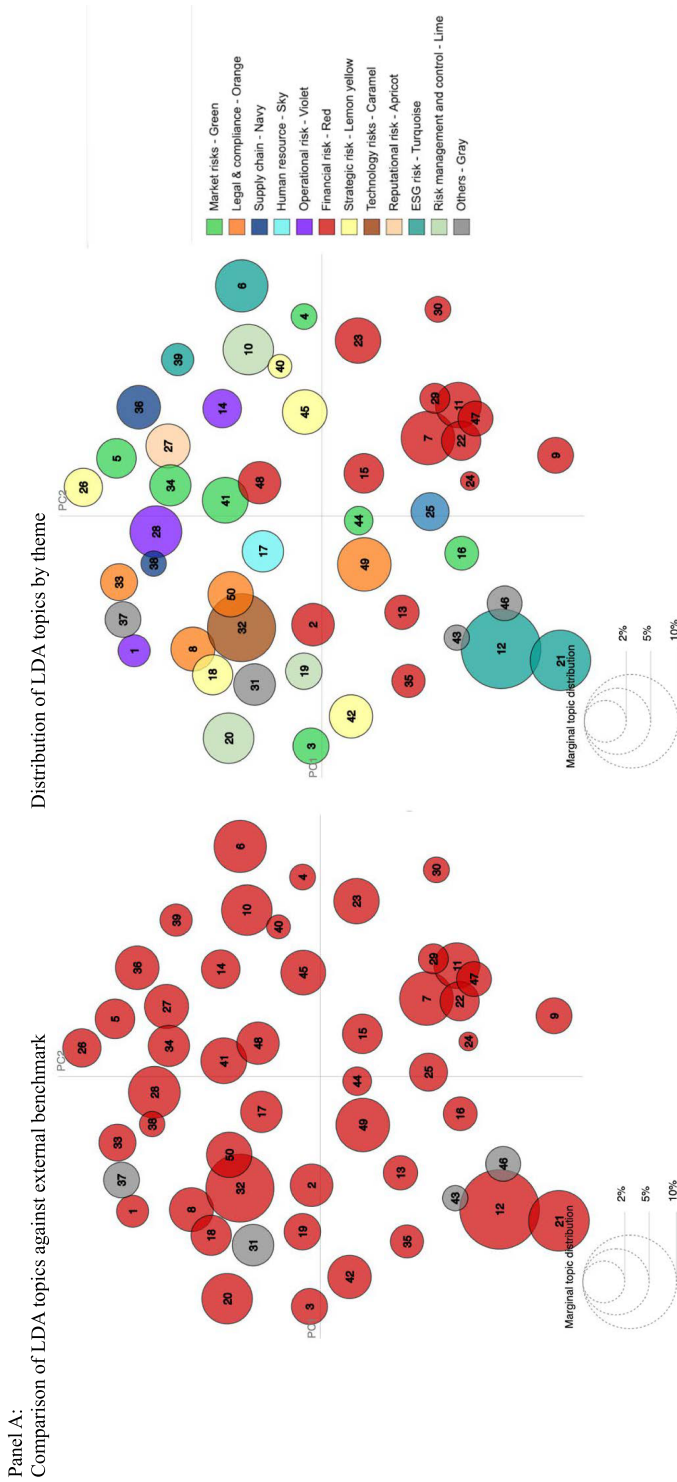
Risk theme	Details
Technological changes	(including risks related to technological disruptions, obsolescence, and innovation)
Reputational risks	
Brand damage	(including risks related to reputation, brand dilution, and loss of customer trust)
ESG risks	
Environmental	(including risks related to climate change, pollution, and environmental degradation)
Social	(including risks related to unethical business practices, social issues, and human rights issues)
Governance	(including risks related to ineffective corporate governance such as board members' conflicts of interest, misconduct, and corruption, misalignment of executive compensation)
Risk management & control	
Risk management structures	(including overview of risk categories and measures, lines of responsibility and accountability, risk management, and mitigation processes)
Risk profile	(including discussion of risk magnitudes and importance, time horizon, and year-over-year changes in risk exposure)
Risk appetite	(level of risk that an organisation is comfortable with and willing to accept)
Resilience & continuity	(ability to withstand, recover from, and adapt to disruptions and adverse events. Examples include risk mitigation plans, strong leadership, culture of adaptability and learning, and strong relationships with stakeholders)
Scenario planning & stress testing	(including the development and analysis of potential scenarios and the testing of the organisation's responses to these scenarios)
Other risks	
Others	(risks that do not fit any of the above categories)

Panel A of [Figure 2](#) visualises the latent topic space for the 50-topic model, revealing that 92% of the LDA topics map directly to expected regulatory categories. However, only 24 of the 36 risks from the FRC-defined list (67%) emerge as distinct topics, with notable gaps in supply chain risks (particularly supplier due diligence), human resource risks (particularly cultural risks and human rights issues), and operational risks related to natural disasters and catastrophes. The dominant risk categories include Market Risks, Compliance Risks, and Financial Risks (Panel A, [Figure 2](#)), with substantial industry clustering that reflects the FTSE350's sectoral composition rather than universal risk concerns.

The 100-topic model visualised in [Figure 3](#) exposes nuanced risk differentiation not visible in the broader 50-topic model, resulting in 75% coverage of FRC-defined topics (27 of 36). Where the 50-topic model identifies a single 'Health & Safety' risk category (Topic 46), the 100-topic model distinguishes between student safety (Topic 4), public transport safety (Topic 33), workplace safety (Topic 51), and multiple other contexts. Similarly, credit risks expand from five general topics to 14 specialised areas including corporate solvency (Topic 75), interest rate exposure (Topic 15), and commercial lending practices (Topic 96). Nevertheless, even this more granular representation fails to capture key risks like Human Rights and Cybersecurity Threats adequately. Findings point to systematic gaps in UK risk reporting regardless of analytical granularity.

## 6.2 Are risk disclosures characterised by boilerplating?

The FRC perceives bland commentary discussing generic risks as a key threat to the usefulness of risk reporting for investors and other stakeholders (FRC 2022, FRC Lab 2021). We use the



Panel A provides inter-topic distance maps for our parsimonious model (M\_0.05\_0.5\_50\_auto). They are coloured in two ways. The first approach uses red for groups matched with any of the 50 pre-defined topics and grey for unexpected topics. The second approach uses a palette of 12 colours: Green for Market risk topics (economic, commodity price, political & geographic, and systemic), Orange for Legal & compliance topics (regulatory, legal disputes, and compliance), Navy for Supply chain risk topics (supply chain disruptions, supplier due diligence), Sky for Human resource risks (cultural, talent retention, health & safety, human rights), Violet for Operational risk topics (production, natural disaster & catastrophe), Red for Financial risk topics (liquidity, credit, interest rate, foreign currency, retirement benefits, tax), Lemon yellow for Strategic risk topics (business model, mergers & acquisitions, market entry, partner relationship), Caramel for Technology risks (cybersecurity, technological changes), Apricot for Reputational risk topics (brand damage), Turquoise for ESG risk topics (environment, social, governance), Lime for Risk management and control (risk management structure, risk profile, risk appetite, resilience and continuity, scenario planning & stress testing), and Gray for the others.

Figure 2. M\_0.05\_0.5\_50\_auto.

## Panel B:

Topic	Top five LDA topic words	GPT mapping to themes in benchmark list
1	packaging, water, materials, energy, integrated	Production Risks
2	contract, pension, defence, scheme, contractual	Retirement Benefit Risks
3	ukraine, russia, sanctions, gas, oil	Political and Geopolitical Risks
4	energy, prices, commodity, network, mitigations	Commodity Price Risks
5	oil, gas, production, prices, energy	Commodity Price Risks
6	carbon, emissions, energy, physical, transition	Environmental Risks
7	ps, stage, banking, total, exposures	Liquidity Risks
8	privacy, network, consumer, digital, tolerance	Compliance Risks
9	ps, shares, total, impairment, profit	Credit Risks
10	ps, stress, severe, plausible, covenants	Scenario Planning and Stress Testing
11	ps, bn, loans, total, impairment	Credit Risks
12	auditor, independence, judgements, auditors, chair	Governance Risks
13	fund, equity, funds, returns, valuation	Liquidity Risks
14	land, construction, build, homes, housing	Production Risks
15	clients, client, money, limits, icaap	Credit Risks
16	adviser, valuation, manager, energy, inflation	Economic Risks
17	recruitment, profit, software, owner, register	Talent Retention
18	acquisition, integration, acquisitions, description, acquired	Mergers and Acquisitions Risks
19	trend, description, defence, register, inherent	Risk Management Structures
20	link, fy, movement, kpis, mitigations	Risk Profile
21	auditor, auditors, independence, ended, appointment	Governance Risks
22	stage, banking, loans, total, collateral	Liquidity Risks
23	property, income, tenants, covenants, properties	Credit Risks
24	bank, loans, loan, limits, currency	Foreign Currency Risks
25	claims, underwriting, solvency, losses, life	Insurance Risks
26	manager, trust, providers, discount, managers	Partner Relationship Risks
27	consumer, brand, consumers, marketing, online	Brand Damage
28	manufacturing, production, materials, raw, sites	Production Risks
29	loans, exposures, ps, counterparty, total	Credit Risks
30	currency, foreign, hedge, fair, instruments	Foreign Currency Risks
31	travel, vehicle, air, fleet, transport	Others
32	transformation, colleagues, drive, digital, esg	Technological Changes
33	manufacturing, intellectual, healthcare, privacy, ip	Legal Disputes Risks
34	food, consumer, sites, labour, movement	Economic Risks
35	condition, currency, income, funds, securities	Foreign Currency Risks
36	retail, store, colleagues, stores, brand	Supply Chain Disruptions
37	divisional, divisions, division, content, direction	Others
38	mining, communities, exploration, community, water	Supplier Due Diligence
39	water, south, priorities, west, erm	Environmental Risks
40	london, property, properties, building, buildings	Business Model Risks
41	production, mining, commodity, prices, communities	Commodity Price Risks
42	category, joint, venture, care, links	Mergers and Acquisitions Risks
43	man, firm, fund, head, funds	Others
44	prudential, esg, transition, restrictions, china	Political and Geopolitical Risks
45	property, retail, ps, income, rental	Business Model Risks
46	healthcare, auditor, fy, africa, owner	Others
47	lending, ps, loans, total, stage	Credit Risks
48	acquisition, currency, foreign, acquisitions, affected	Foreign Currency Risks
49	defence, chair, stress, outcomes, lines	Compliance Risks
50	bribery, register, anti, corruption, whistleblowing	Compliance Risks

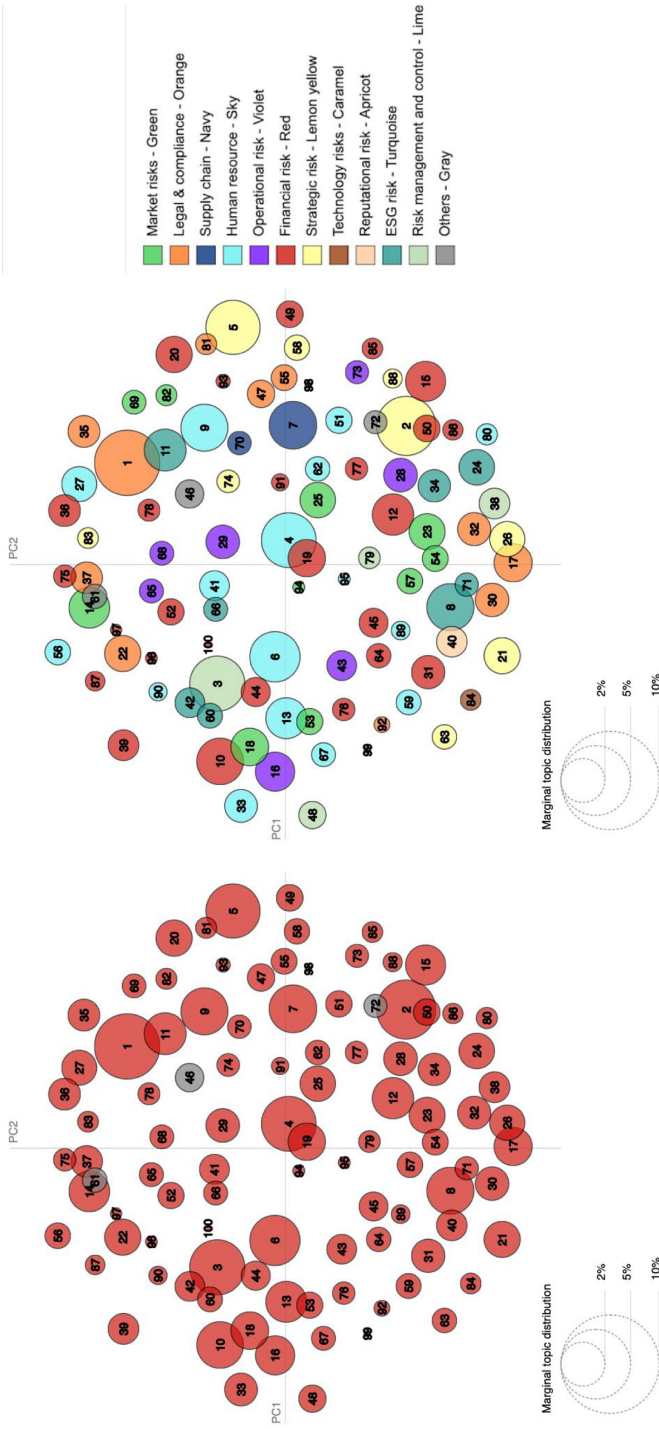
Panel B provides top five keywords of each topic with GPT's classification of the topic into 37 predefined benchmark themes in Table 7. We use GPT4o for this classification.

Figure 2 *Continued*

Panel A:

Comparison of LDA topics against external benchmark

Distribution of LDA topics by theme



Panel A provides inter-topic distance maps for our granular model (M 0.01 0.7 100 auto auto). They are coloured in two ways. The first approach uses red for groups matched with any of the 36 pre-defined topics and grey for unexpected topics. The second approach uses a palette of 12 colours: Green for Market risk topics (economic, commodity price, political & geographic, and systemic), Orange for Legal & compliance topics (regulatory, legal disputes, and compliance), Navy for Supply chain risk topics (supply chain disruptions, supplier due diligence), Sky for Human resource risks (cultural, talent retention, health & safety, human rights), Violet for Operational risk topics (production, natural disaster & catastrophe), Red for Financial risk topics (liquidity, credit, interest rate, foreign currency, retirement benefits, tax), Lemon yellow for Strategic risk topics (business model, mergers & acquisitions, market entry, partner relationship), Caramel for Technology risks (cybersecurity, technological changes), Apricot for Reputational risk topics (brand damage), Turquoise for ESG risk topics (environment, social, governance), Lime for Risk management and control (risk management structure, risk profile, risk appetite, resilience and continuity, scenario planning & stress testing), and Gray for the others.

Figure 3. M 0.01\_0.7\_100\_auto\_auto.

Panel B:

Topic	top words	GPT mapping to themes in benchmark list
1	consumer, businesses, products, privacy, tax	Compliance Risks
2	man, firm, funds, fund, ey	Business Model Risks
3	client, clients, resilience, chief, stress	Scenario Planning and Stress Testing
4	property, students, read, student, safety	Health and Safety
5	portfolio, investments, valuation, valuations, returns	Business Model Risks
6	production, adjusted, health, culture, margin	Cultural Risks
7	colleagues, safety, product, suppliers, colleague	Supply Chain Disruptions
8	climate, carbon, emissions, energy, targets	Environmental Risks
9	businesses, safety, health, suppliers, colleagues	Health and Safety
10	psm, contracts, interest, rate, hedge	Foreign Currency Risks
11	sustainability, environmental, health, diversity, social	Environmental Risks
12	psm, loans, stage, ps, credit	Credit Risks
13	water, south, west, safety, health	Health and Safety
14	brexit, eu, increasing, competitive, relationships	Political and Geopolitical Risks
15	credit, interest, rate, exposures, stress	Credit Risks
16	ar, indd, pm, proof, product	Production Risks
17	clients, integrity, privacy, code, elt	Compliance Risks
18	digital, political, trend, tax, transformation	Political and Geopolitical Risks
19	ps, psm, banking, insurance, lloyds	Liquidity Risks
20	fy, auditor, clinical, international, mr	Tax Risks
21	trend, description, acquisitions, tax, acquisition	Mergers and Acquisitions Risks
22	waste, contracts, safety, environmental, health	Compliance Risks
23	effect, oil, gas, earnings, condition	Commodity Price Risks
24	officer, chief, chair, defence, conduct	Governance Risks
25	covid, pandemic, climate, government, response	Systemic Risks
26	fy, businesses, products, local, home	Mergers and Acquisitions Risks
27	culture, delivery, teams, progress, leadership	Cultural Risks
28	products, safety, health, demand, product	Production Risks
29	content, advertising, safety, product, direction	Production Risks
30	products, safety, tax, product, privacy	Compliance Risks
31	credit, stage, total, loans, standard	Credit Risks
32	prudential, insurance, products, businesses, adversely	Regulatory Risks
33	rail, bus, transport, safety, demand	Health and Safety
34	mining, safety, project, projects, environmental	Environmental Risks
35	contract, contracts, safety, health, divisional	Legal Disputes Risks
36	client, clients, trading, credit, limits	Credit Risks
37	energy, pwc, safety, health, project	Compliance Risks
38	scenarios, debt, scenario, ps, concern	Scenario Planning and Stress Testing
39	manager, portfolio, trust, investments, providers	Liquidity Risks
40	tolerance, owner, brand, products, chain	Brand Damage
41	aa, protection, revenue, details, health	Health and Safety
42	auditor, accounting, independence, meetings, matters	Governance Risks
43	products, product, clinical, commercial, manufacturing	Production Risks
44	bank, credit, loans, loan, currency	Credit Risks
45	ps, tax, impairment, psm, net	Tax Risks
46	insurance, claims, underwriting, reinsurance, losses	Others
47	royal, auditor, pwc, ofcom, trade	Regulatory Risks

Figure 3 *Continued*

48	life, standard, rmf, holdings, outcomes	Resilience and Continuity
49	credit, interest, currency, tax, foreign	Credit Risks
50	adversely, condition, products, income, tax	Tax Risks
51	london, health, safety, property, staff	Health and Safety
52	hsbc, holdings, bn, transition, rate	Interest Rate Risks
53	energy, infrastructure, gas, network, commodity	Commodity Price Risks
54	portfolio, investments, manager, valuation, adviser	Economic Risks
55	low, register, director, engineering, fy	Compliance Risks
56	land, homes, safety, build, government	Health and Safety
57	oil, gas, production, price, prices	Commodity Price Risks
58	retail, store, product, stores, brand	Business Model Risks
59	airlines, travel, aviation, air, aircraft	Health and Safety
60	production, oil, safety, likelihood, medium	Environmental Risks
61	natwest, stage, ps, ecl, psm	Others
62	project, projects, wood, safety, execution	Health and Safety
63	consumer, brand, revenue, online, marketing	Business Model Risks
64	ps, credit, loans, core, trading	Credit Risks
65	product, products, safety, manufacturing, chain	Production Risks
66	auditor, manager, independence, ended, valuation	Governance Risks
67	portfolio, retail, property, ps, safety	Health and Safety
68	property, portfolio, income, asset, tenants	Production Risks
69	ukraine, iron, ore, steel, production	Commodity Price Risks
70	project, chain, construction, safety, contract	Supply Chain Disruptions
71	remuneration, share, shares, director, bonus	Governance Risks
72	client, auditors, head, clients, fund	Others
73	mining, production, mine, safety, exploration	Production Risks
74	vehicle, vehicles, ti, oem, demand	Business Model Risks
75	credit, solvency, businesses, longevity, climate	Credit Risks
76	link, tax, profile, chain, product	Tax Risks
77	manager, property, probability, low, portfolio	Liquidity Risks
78	banking, credit, stage, loans, impairment	Credit Risks
79	residual, low, likelihood, rating, moderate	Risk Management Structures
80	healthcare, clinical, patient, nhs, movement	Health and Safety
81	packaging, water, safety, products, plastic	Compliance Risks
82	climate, esg, inflation, ukraine, covid	Systemic Risks
83	store, stores, practices, venture, care	Business Model Risks
84	product, products, chief, officer, colleagues	Technological Changes
85	psm, ps, credit, total, money	Credit Risks
86	arc, gambling, tax, gaming, fi	Tax Risks
87	barclays, ps, bn, psm, loans	Liquidity Risks
88	products, businesses, acquisition, product, revenue	Mergers and Acquisitions Risks
89	food, safety, consumer, products, health	Health and Safety
90	clients, recruitment, regional, client, resource	Talent Retention
91	funding, credit, lending, loan, loans	Credit Risks
92	product, products, consumer, regulation, categories	Regulatory Risks
93	pension, scheme, businesses, defence, contracts	Retirement Benefit Risks
94	commodity, safety, demand, communities, production	Commodity Price Risks
95	divisional, businesses, safety, products, hse	Health and Safety

Figure 3 *Continued*

96	bank, ps, lending, commercial, credit	Credit Risks
97	credit, bank, funding, conduct, lending	Credit Risks
98	chief, officer, clearing, profile, transaction	Liquidity Risks
99	bp, liability, safety, laws, relationships	Legal Disputes Risks
100	debt, rating, medium, royalty, income	Credit Risks

Panel B provides top five keywords of each topic with GPT's classification of the topic into 36 predefined benchmark themes in Table 7. We use GPT4o for this classification.

Figure 3 *Continued*

Table 8. Summary statistics for Hierarchical Linear Models examining the source of variation in topic intensity.

Model	Within firm	Across firms within industry	Across industries
M_0.05_0.5_35_auto_auto	27.20%	47.95%	24.85%
M_0.01_0.7_100_auto_auto	74.57%	20.30%	5.13%

Table 8 provides the result of Hierarchical Linear Model analysis decomposing the variance in a parsimonious model (M\_0.05\_0.5\_35\_auto\_auto) and a granular model (M\_0.01\_0.7\_100\_auto\_auto). For each model-topic, we obtain interclass correlation coefficients (ICCs) within firms, across firms within industry, and across industries. Then, we calculate the mean value of each type of ICCs across topics. Our sample for HML analysis includes 1,992 annual reports.

distribution of LDA topic intensity across annual reports to assess whether the content and focus of FTSE350 firms' risk disclosures vary in ways consistent with informative commentary.<sup>22</sup> We use Hierarchical Linear Modelling [HLM] (Raudenbush and Bryk 2002) to measure variation in topic intensity.<sup>23</sup> We use this method to test whether risk reporting contains meaningful firm-, sector – and time-specific content versus generic and invariant boilerplate commentary.

We use the 35-topic model as the baseline for this analysis. Table 8 (row 1) reveals that the 35-topic model suggests risk reporting is largely static: permanent industry differences explain 25% of variation, time-invariant firm characteristics account for 48%, while dynamic firm-level variation that arguably offers most value to investors, account for only 27%. This pattern suggests that FRC concern about boilerplating is valid (FRC 2022, FRC Lab 2021). The 100-topic model (Table 8, row 2) nevertheless reverses this conclusion entirely. Dynamic firm-level variation accounts for 73% of topic intensity variation, with permanent industry effects dropping to just 5%. These results suggest firms actively tailor risk commentary to their evolving circumstances, contradicting the boilerplate conclusion apparent with our coarser model.

Taken together, these results reinforce the paper's central message: implementation choices shape the substantive conclusions drawn from topic models. In the case of selecting the 'optimal' level of topic granularity, for example, the final decision is likely to pivot on the precise nature of

<sup>22</sup>Rare-word filtering may understate highly idiosyncratic disclosures, but our focus is on recognisable risk topics, so the main inferences are unaffected.

<sup>23</sup>An alternative approach to assess the degree of time, industry – and firm-level uniqueness is to perform a variance decomposition using a fixed effects model (Hassan et al. 2019). We favour HLM for our main analysis because results are more parsimonious to report. Findings and conclusions using a variance decomposition analysis are entirely consistent with those we report and are available from the authors on request.

the research objective: aggregate representations may better capture first-order risk categories and industry patterns, while granular models reveal firm-specific nuances that might otherwise appear homogeneous.

## 7. Discussion and conclusions

The popularity of text mining using LDA topic modelling is growing among accounting researchers. The LDA algorithm provides a powerful and relatively accessible tool for describing large financial corpora and categorising individual documents within these corpora. A prevailing view amongst accounting (and finance) researchers is that the algorithmic nature of LDA also promises significant research design benefits in the form of less researcher subjectivity and higher levels of replicability. Indeed, our literature review suggests that the method is frequently portrayed as providing both a tool for mining large volumes of text *and* a solution for minimising the effect of researcher-induced bias.

We review the LDA method and demonstrate numerous implementation aspects that require researchers to apply significant judgements, either implicitly or explicitly. In many ways, the degree of discretion involved in discovering topics with LDA is no different to that required for manual content analysis. The key difference, however, is that researcher choices for LDA are often hidden and their effects overlooked. Our review of LDA applications in the accounting and finance literature confirms that researchers often fail to discuss key design choices when applying the method. We conjecture that this is because the choices are often implicit and not well understood (e.g. relying on default settings in the coding step or on choices made in extant work). Having highlighted the various decisions that researchers need to address we proceed to provide evidence on the effects of (some of) these choices via a series of experiments where we use LDA to discover topics in risk disclosures made by FTSE350 firms during the period 2018-2022. Findings from our experiments confirm that LDA topic model results can be extremely sensitive to implementation choices. Crucially, these decisions extend beyond the LDA algorithm to include the properties of the input text and the labelling strategy for assigning meaning to topics.

Our analysis extends beyond LDA to consider recent advances in embedding-based topic models. We find these newer approaches, such as BERTopic, can generate more coherent and interpretable topics by leveraging semantic embeddings to relax the bag-of-words assumption inherent in LDA. While not universally superior across all evaluation metrics, the evidence suggests that LDA should no longer be viewed as the default topic modelling method for accounting and finance research. Importantly, however, these advanced methods do not eliminate the need for researcher discretion; instead, they introduce a new suite of critical implementation choices related to embedding models, dimensionality reduction, and clustering that must be carefully considered and justified.

There is no recipe book for dealing with these choices. The ‘correct’ choice often depends on the research setting and the research question(s) under investigation. Even then, researchers are likely to face a range of plausible representations of the latent topic space. The good news is that the high degree of ambiguity provides the opportunity for domain expertise to elevate the analysis from a purely mechanical task to a process where accounting researchers can exploit their comparative advantage. Accordingly, we argue that researcher discretion in topic modelling is a feature that accounting researchers should be celebrating rather than downplaying.

We apply the lessons from our experiments to study the properties of FTSE350 firms’ annual report risk disclosures. Two competing models comprising 35 and 100 topics that both display high evaluation scores produce quite different insights. While results using the 35-topic

representation suggest very high levels of stale, generic content, findings using the 100-topic representation indicate that annual report risk disclosures provide the type of dynamic content that investors seek. Precisely which conclusion is correct depends on how one frames the research question, defines a material risk, and ultimately implements the chosen model.

### Acknowledgement

We are grateful to Paulo Alves for support extracting and processing risk disclosures from UK PDF annual reports. We are grateful for comments from Mark Clatworthy and Maria Correia and workshop participants at the University of Bristol, Stockholm School of Economics, Queen's University Belfast, Grudis Global Accounting Research Seminar, and colleagues at the Financial Reporting Council for help interpreting risk disclosures. The authors are entirely responsible for the results and conclusions presented in this paper. Analyses and findings do not represent the views of the Financial Reporting Council or any of the individuals who work for it.

### Disclosure statement

No potential conflict of interest was reported by the author(s).

### Funding

This work was supported by the Financial Reporting Council: [grant number FRC2023-0131].

### Supplemental data

Supplemental data for this article can be accessed online at <https://doi.org/10.1080/00014788.2026.2625716>.

### References

- Abraham, S., and Shrivies, P. J., 2014. Improving the relevance of risk factor disclosure in corporate annual reports. *The British Accounting Review*, 46 (1), 91–107.
- Angelov, D., 2020. Top2Vec: distributed representations of topics, *arXiv*, 2008.09470.
- Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., and Baldwin, B., 2024. LLM stability: a detailed analysis with some surprises, *arXiv*, 2408.04667.
- Ball, C., Hoberg, G., and Maksimovic, V., 2015. Disclosure, business change and earnings quality. Working Paper.
- Beattie, V., McInnes, B., and Fearnley, S., 2004. A methodology for analysing and evaluating narratives in annual reports: a comprehensive descriptive profile and metrics for disclosure quality attributes. *Accounting Forum*, 28 (3), 205–236.
- Bellstam, G., Bhagat, S., and Cookson, J. A., 2021. A text-based analysis of corporate innovation. *Management Science*, 67 (7), 3985–4642.
- Bhalla, V., and Klimcikova, K., 2019. Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications, pp.264–274, Florence, Italy, August 2, 2019. Association for Computational Linguistics (ACL).
- Bianchi, F., Terragni, S., and Hovy, D., 2021. Cross-lingual contextualized topic models with zero-shot learning. *EACL*, 2021.
- Bingler, J. A., Kraus, M., Leippold, M., and Webersinke, N., 2022. Cheap talk and cherry-picking: what ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 102776.
- Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Bozanic, Z., Roulstone, D. T., and Van Buskirk, A., 2018. Management earnings forecasts and other forward-looking statements. *Journal of Accounting and Economics*, 65 (1), 1–20.

- Breuer, M., and DeHaan, E., 2024. Using and interpreting fixed effects models. *Journal of Accounting Research*, 62, 1183-1226.
- Brown, G. W., Gredil, O. R., and Kantak, P., 2022. Finding fortune: how do institutional investors pick asset managers? *Review of Financial Studies*, 36 (8), 3071–3121.
- Brown, N., Crowley, R., and Elliott, W. B., 2020. What are you saying? Using topic to detect financial misreporting. *Journal of Accounting Research*, 58 (1), 237–291.
- Bybee, L., Kelly, B., Manela, A., and Xiu, D., 2024. Business news and business cycles. *The Journal of Finance*, 79 (5), 3105–3147.
- Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H-M., and Steele, L. B., 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19 (1), 396–455.
- Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M., 2009. Reading tea leaves: how humans interpret topic models. *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada: Curran Associates Inc., pp. 288–296.
- Comyns, B., and Figge, F., 2015. Greenhouse gas reporting quality in the oil and gas industry. *Accounting, Auditing & Accountability Journal*, 28 (3), 403–433.
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K., 2019. Bert: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.
- Donovan, J., Jennings, J., Koharki, K., and Lee, J., 2021. Measuring credit risk using qualitative disclosure. *Review of Accounting Studies*, 26, 815–863.
- Dyer, T., Lang, M., and Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64 (2-3), 221–245.
- El-Haj, M., Alves, P., Rayson, P., Walker, M., and Young, S., 2020. Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as PDF files. *Accounting and Business Research*, 50 (1), 6–34.
- El-Haj, M., Rayson, P., Walker, M., Young, S., and Simaki, V., 2019. In search of meaning: lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, 46 (3-4), 265–306.
- Fedyk, A., 2024. Front-Page news: the effect of news positioning on financial markets. *The Journal of Finance*, 79 (1), 5–33.
- Financial Reporting Council, 2014. *Guidance on the Strategic Report*. London, UK: Financial Reporting Council.
- Financial Reporting Council, 2018a. *Guidance on the Strategic Report*. London, UK: Financial Reporting Council.
- Financial Reporting Council, 2018b. *Business Model Reporting; Risk and Viability Reporting - Where are We Now?* London, UK: Financial Reporting Council.
- Financial Reporting Council, 2022. *Review of Corporate Governance Reporting*. London, UK: Financial Reporting Council.
- Financial Reporting Council Lab, 2021. *Reporting on Risks, Uncertainties, Opportunities and Scenarios*. London, UK: Financial Reporting Council.
- Goldsmith-Pinkham, P., Hirtle, B., and Lucca, D., 2016. *Parsing the Content of Bank Supervision*. New York: Federal Reserve Bank of New York.
- Groendorst, M., 2022. BERTopic: neural topic modelling with a class-based TF-IDF procedure. *arXiv*, 2203.05794.
- Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y., 2023. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection, *arXiv*, 2301.07597.
- Hanley, K. W., and Hoberg, G., 2019. Dynamic interpretation of emerging risks in the financial sector. *Review of Financial Studies*, 32 (12), 4543–4603.
- Hassan, T. A., Hollander, S., van Lent, L., and Tahoun, A., 2019. Firm-Level political risk: measurement and effects. *The Quarterly Journal of Economics*, 134 (4), 2135–2202.
- Hoberg, G., and Lewis, C., 2017. Do fraudulent firms produce abnormal disclosure? *Journal of Corporate Finance*, 43, 58–85.
- Hoffman, M., Bach, F., and Blei, D., 2010. Online learning for latent dirichlet allocation. *Advances in neural information processing systems*, 23, 856–864.
- Hofmann, T., 1999. Probabilistic Latent Semantic Analysis. *UAI '99 Proceedings*, 289–296.
- Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A., 2014. Interactive topic modeling. *Machine Learning*, 95 (3), 423–469.

- Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R., 2018. Analyst information discovery and interpretation roles: a topic modeling approach. *Management Science*, 64 (6), 2833–2855.
- Huang, A. H., Wang, H., and Yang, Y., 2023. Finbert: a large language model for extracting information from financial text. *Contemporary Accounting Research*, 40 (2), 806–841.
- Jegadeesh, N., and Wu, D. A., 2017. Deciphering fedpeak: The information content of FOMC meetings. Working paper.
- Jehnen, S., Ordieres-Meré, J., and Villalba-Diez, J., 2025. FinTextSim: enhancing financial text analysis with BERTopic. arXiv, 2504.15683.
- Koelbel, J. F., Leippold, M., Rillaerts, J., and Wang, Q., 2024. Ask BERT: how regulatory disclosure of transition and physical climate risks affects the CDS term structure. *Journal of Financial Econometrics*, 22 (1), 30–69.
- Lee, D. D., and Seung, H. S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.
- Li, H., Gao, H., Wu, C., and Vasarhelyi, M. A., 2025. Extracting financial data from unstructured sources: leveraging large language models. *Journal of Information Systems*, 39 (1), 135–156.
- Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J. Y., and Wen, J. R., 2024. The dawn after the dark: an empirical study on factuality hallucination in large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10879–10899.
- Li, Z., Zhang, X., Long, D., Xie, P., and Zhang, M., 2023. Towards general text embeddings with multi-stage contrastive learning. arXiv, 2308.03281.
- Lowry, M., Michaely, R., and Volkova, E., 2020. Information revealed through the regulatory process: interactions between the SEC and companies ahead of their IPO. *Review of Financial Studies*, 33 (12), 5510–5554.
- McInnes, L., Healy, J., and Melville, J., 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3 (29), 861.
- Menkveld, A. J., Dreber, A., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Neusüss, S., et al., 2024. Non-standard errors. *Journal of Finance*, 79 (3), 2339–2390.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv, 1301.3781.
- Mu, Y., Bai, P., Bontcheva, K., and Song, X., 2024a. Addressing topic granularity and hallucination in large language models for topic modelling. arXiv, 2405.00611.
- Mu, Y., Dong, C., Bontcheva, K., and Song, X., 2024b. Large language models offer an alternative to the traditional approach of topic modelling. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, Torino, Italy: ELRA Language Resources Association and International Committee on Computational Linguistics.
- Raudenbush, S. W. and Bryk, A. S. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, Calif.: Sage.
- Reimers, N., and Gurevych, I., 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv, 1908.10084.
- Röder, M., Both, A., and Hinneburg, A., 2015. Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai, China: Association for Computing Machinery, pp. 399–408.
- Schofield, A., and Mimno, D., 2016. Comparing apples to apple: the effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4, 287–300.
- Siano, F., and Wysocki, P., 2021. Transfer learning and textual analysis of accounting disclosures: applying big data methods to small(er) datasets. *Accounting Horizons*, 35 (3), 217–244.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I., 2017. Attention is all you need. arXiv, 1706.03762.
- Wu, X., Nguyen, T., and Luu, A. T., 2024. A survey on neural topic models: methods, applications, and challenges. arXiv, 2401.15351.