Bridging Topic Modelling Outputs to Bayesian Hierarchical Model Using LLM and WordNet Parameter Labelling

Vadrianey Asas[1], Sarah Samson Juan[2]*, Stephanie Chua[3], Evan Lau[4], Jane Labadin[5]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia[1, 2, 3, 5]
Faculty of Economics and Business, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia[4]

***Abstract*—This study investigates the challenge of generating accurate and interpretable topic labels for integration into Bayesian Hierarchical Models (BHM), a critical step for interpretable probabilistic risk modelling from unstructured textual data. Using a corpus of 35,667 Malaysian business news articles published between 2019 and 2023, four topic modelling approaches, such as Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), Top2Vec, and BERTopic, were evaluated. Among these, NMF produced the most coherent and thematically consistent topics. To address the topic-labelling challenge, this study proposes an NLP-BHM framework that maps topic model outputs into a hierarchical Bayesian structure comprising interpretable topic labels and higher-level risk categories. Two semantic labelling strategies were examined: Large Language Model (LLM) prompting and WordNet-based semantic analysis. The proposed approaches enabled systematic topic interpretation and semantic clustering within the BHM framework. A case study on Malaysian business risks demonstrates that LLM-based labelling produced more coherent and contextually relevant results, while WordNet-based labelling provided a semantically consistent but vocabulary-limited alternative. Comparative results based on Mean Opinion Scores (MOS) highlight the effectiveness of LLM-based semantic labelling in improving interpretability for probabilistic business risk analysis.**

***Keywords*—Bayesian hierarchical model; natural language processing; topic modelling; large language model; WordNet**

I. Introduction

In today's data-driven economy, extracting actionable insights from unstructured text is crucial for effective risk assessment and informed decision-making. Business news articles, financial disclosures, and policy updates often contain critical early signals of systemic risks. However, their narrative form poses challenges for direct integration into quantitative models. To generate interpretable probabilistic risk representations, it is essential to translate this unstructured information into structured, interpretable inputs suitable for Bayesian Hierarchical Models (BHM).

Bayesian Hierarchical Models (BHM) provide a probabilistic framework for modelling uncertainty across multiple levels of abstraction, enabling interpretable relationships between observed variables and higher-level latent structures [1]. The power of Bayes' theorem lies in its adaptability: parameters, hypothesised relationships, and confidence intervals can all be estimated as probabilistic statements derived from the model. Traditionally, parameter labels for BHM are manually annotated, a process that can be highly time-consuming and inconsistent ([2], [3]). One notable work applies manual annotation to obtain the interpretable labels for the factor inputs in their study [4]. Despite these efforts, no significant work has integrated topic modelling outputs into BHM structures, leaving the process of bridging these outputs into BHM largely unexplored.

Prior investigations have identified the best-performing topic model based on coherence scores, which will be utilised for the integration into the BHM structure [5]. Research on topic modelling, such as the foundational work by Blei et al. [6], has established methods for extracting topics from text data, but the challenge remains in linking these outputs with BHM factor inputs. However, the core challenge in this study lies in bridging the topic outputs from the best-performing topic model with the factor inputs to obtain clear, interpretable labels required by the BHM structure.

This study aims to fill this gap by proposing a framework that integrates topic-modelling outputs into BHM structures. We evaluate two parameter labelling strategies: Large Language Model (LLM) prompting and WordNet-based semantic analysis. These strategies will help derive interpretable labels for BHM parameters, facilitating the connection between natural language processing (NLP)-derived topics and Bayesian models. The central research question guiding this work is: How can LLM-based and WordNet-based labelling strategies be leveraged to generate interpretable topic representations and construct a BHM structure for probabilistic risk analysis?

The contribution of the study lies within the framework itself. In contrast to the traditional BHM that relies on defined variables and structured inputs, the proposed approach automatically derives risk factors from unstructured business news, enabling a fully data-driven representation of emerging risk signals. In addition, the framework systematically compares two complementary semantic labelling strategies, namely LLM-based and WordNet-based approaches, to enhance the interpretability and robustness of topic-to-risk mapping. Furthermore, the model enables dynamic estimation of risk probabilities by propagating topic-level distributions across evolving news corpora, thus capturing temporal variations in business risk environments.

The study is organised as follows. In Section II, we briefly discuss related work on the approaches used in our work. We

*Corresponding author.

present our framework for integrating the topic model's output using two labelling approaches in Section III. In Section IV, we discuss the results from the Mean Opinion Scoring (MOS) evaluation and present the BHM structures built using both approaches. Finally, we present our conclusions and future work in Section V.

II. Literature Review

A. Bayesian Hierarchical Model

Bayesian modelling provides a probabilistic framework for reasoning under uncertainty by updating prior beliefs using observed evidence [2]. Bayesian Hierarchical Models (BHM) extend this framework by organising model components across multiple levels, thereby enabling information sharing and uncertainty propagation throughout the hierarchy ([2], [3]).

0 illustrates a Bayesian hierarchical structure with population-level hyperparameters, group-specific parameters, and observed data. Shared priors enable information sharing across groups, while group-level parameters capture variation and govern the observations. Posterior inference updates uncertainty across the hierarchy, supporting partial pooling and more stable estimates when group-level data are limited.

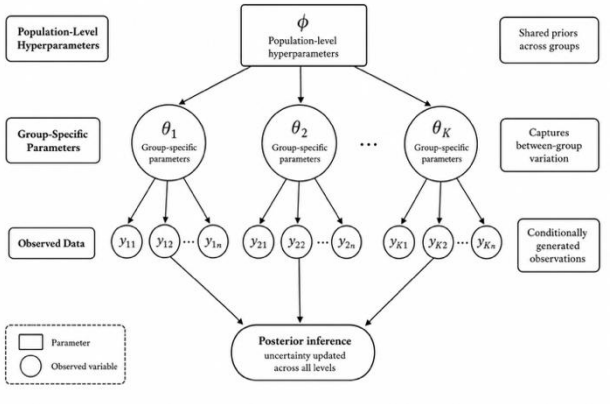

Fig. 1. A general structure of a Bayesian hierarchical model.

Various studies have applied BHM for predictive modelling across different domains. In supply chain management, BHM is used to assess disruption risks with limited data availability by estimating supplier risk probabilities and financial impacts, and incorporates downstream risk assessment measures, such as Value-at-Risk (VaR), to support decision-making [4]. Similarly, in supply chain financing, BHM improves risk evaluation by integrating macro-level credit ratings with micro-level transaction data, resulting in more accurate predictions of payment delays and revealing behavioural patterns such as correlations between delivery and payment delays [7].

Beyond financial applications, BHM has shown strong capability in ecological and environmental modelling. In species distribution studies, it integrates environmental variables with ecological interactions and latent processes, improving predictive performance when combined with complementary models such as food-web approaches [8]. In public health, BHM is applied to estimate disease burden from incomplete surveillance data, providing robust and scalable estimates for long-term monitoring and decision-making during the COVID-19 pandemic [9]. In hydrology, BHM models extreme flood risks under nonstationary conditions by integrating climate indices within a hierarchical framework, improving estimation accuracy and reducing uncertainty through partial pooling [10].

Recent studies in natural language understanding have demonstrated that uncertainty-aware probabilistic representations can better capture semantic ambiguity and subjective variation in textual interpretation, further motivating the integration of probabilistic reasoning within NLP frameworks [11].

These studies demonstrate the robustness of BHM in integrating heterogeneous data sources and modelling complex dependencies across domains. However, most existing approaches rely on predefined model structures and manually specified parameters, limiting flexibility in adapting to evolving data patterns and latent factors. This limitation reduces its effectiveness when dealing with large-scale unstructured data, where important information is often embedded within latent textual representations.

B. Topic Modelling

Topic modelling has long been a popular technique for uncovering hidden themes in large text corpora [6]. A study applies LDA to a large-scale corpus of artificial intelligence literature to extract latent topics from publication abstracts[12]. The topic distributions are then aggregated across time, journals, and countries to identify key AI subfields and analyse research trends and similarities across publication sources. Another study applies a semantically assisted NMF approach (SeNMFk) to extract latent topics and determine the optimal number of topics [13]. By integrating TF-IDF and semantic information through SPPMI matrices, the model improves topic stability and coherence compared to traditional methods.

Top2Vec has been increasingly used across domains due to its ability to combine word embeddings with clustering to produce semantically coherent topics[14]. It was applied to analyse 15,000 mobile app user feedback entries[15], where it outperformed LDA and ETM by automatically selecting the number of topics and producing more interpretable results. Evaluation metrics such as the coherence score and topic diversity demonstrated their effectiveness in capturing issues such as app glitches and performance problems. BERTopic, a transformer-based topic modelling method, has also been widely used to uncover latent themes in text data[16]. One study applied BERTopic to 2,846 PubMed abstracts on depression, anxiety, and burnout in academia[17], identifying 27 topics including pandemic-related anxiety and medical resident burnout.

A comparative study of LDA, NMF, Top2Vec, and BERTopic on COVID-19-related Twitter data highlights challenges in analysing noisy social media text and suggests that BERTopic and NMF are more effective approaches for social science applications[1]. The study also emphasises the importance of qualitative interpretation and domain expertise. However, while prior work compares topic modelling methods for extracting insights from unstructured text, most studies primarily focus on topic extraction quality, coherence, and interpretability. Limited attention has been paid to transforming latent topic representations into structured probabilistic

frameworks, such as Bayesian Hierarchical Models, for downstream risk modelling and uncertainty-aware decision-making.

C. Large Language Models

Recent advances in Large Language Models (LLMs) have significantly improved natural language understanding and semantic reasoning capabilities in NLP tasks [18]. Models such as GPT, PaLM, and LLaMA can generate context-aware responses by learning rich semantic representations from large-scale corpora [19]. Unlike traditional rule-based or ontology-driven approaches, LLMs can capture contextual relationships between words and concepts, enabling more flexible and semantically meaningful interpretation of unstructured text.

This capability has led to increasing use of LLMs in semantic extraction, topic interpretation, summarisation, and decision-support applications. Recent studies have shown that LLMs can generate coherent topic labels and semantic summaries that better align with human interpretation compared to conventional lexical resources and statistical approaches [18]. Their ability to contextualise latent textual representations makes them particularly well-suited to interpreting topic modelling outputs, where extracted keywords often lack semantic clarity when analysed in isolation. Compared to ontology-based resources such as WordNet, LLMs offer stronger contextual reasoning and semantic flexibility, especially in domain-specific environments such as finance and business analytics. Traditional lexical resources rely on predefined vocabularies and hierarchical relationships, which may limit their ability to capture emerging terminology, implicit semantics, and context-dependent meanings. In contrast, LLMs can dynamically infer semantic relationships from textual context, improving interpretability and semantic coherence in topic labelling tasks.

Despite these advantages, several challenges remain in applying LLMs within probabilistic modelling frameworks. LLM-generated outputs may vary with prompt design, sampling configurations, and model randomness, potentially affecting consistency and reproducibility. Furthermore, hallucinated or overly general outputs may reduce reliability when semantic labels are directly integrated into downstream analytical models. Consequently, human validation and structured prompting strategies remain important to ensure semantic accuracy and interpretability.

While recent work has explored the use of LLMs for semantic extraction and financial text analysis [20], limited attention has been given to integrating LLM-generated semantic representations into hierarchical probabilistic frameworks for interpretable business risk modelling. This limitation motivates the present study, which leverages LLM-based semantic labelling to bridge latent topic representations with Bayesian Hierarchical Models for uncertainty-aware risk analysis.

D. WordNet-Based Semantic Analysis

WordNet is one of the most widely used lexical databases in natural language processing for semantic analysis and knowledge representation[21]. It organises words into sets of cognitive synonyms known as synsets, which are connected through semantic relationships such as synonymy, hypernymy, and hyponymy. These structured semantic relationships enable WordNet to support tasks including word sense disambiguation, semantic similarity measurement, text classification, and ontology-based information retrieval.

In topic modelling and semantic analysis, WordNet has been widely used to improve topic coherence and interpretability by identifying semantically related concepts and hierarchical relationships between keywords. Several studies employ WordNet-based semantic similarity measures, such as Wu–Palmer similarity and path-based similarity, to cluster semantically related terms and generate interpretable topic labels [22]. By leveraging lexical hierarchies, WordNet enables semantically related topics to be grouped into broader conceptual categories, supporting more structured knowledge representation. Compared to statistical topic modelling methods, ontology-based approaches offer greater consistency and reproducibility because semantic relationships are derived from predefined lexical structures rather than probabilistic generation. This deterministic behaviour makes WordNet particularly useful in applications requiring stable semantic mappings and explainable semantic relationships. Furthermore, hierarchical lexical structures align naturally with probabilistic hierarchical models, making WordNet-based semantic grouping well-suited to constructing interpretable latent-factor hierarchies.

However, WordNet also presents several limitations, particularly in domain-specific applications such as business and financial risk analysis. Since WordNet relies on manually curated vocabularies, it may fail to capture emerging terminology, contextual nuances, and specialised business concepts commonly found in dynamic textual corpora. In addition, lexical similarity does not necessarily reflect contextual semantic meaning, which may lead to overly generic or semantically weak topic labels when applied to complex business news data.

These limitations motivate the exploration of context-aware semantic approaches that use Large Language Models to dynamically infer semantic relationships from textual context. Nevertheless, WordNet remains a valuable baseline for evaluating semantic consistency and hierarchical topic clustering within interpretable probabilistic modelling frameworks.

E. Challenges in Bridging Text Data and BHM

Integrating unstructured text data into BHM presents several critical challenges that must be addressed to enhance the effectiveness of risk analysis and decision-making processes. Understanding these challenges provides a foundation for the proposed framework outlined in the subsequent section. One of the primary challenges is data preprocessing. Unstructured text often contains significant noise and irrelevant information, which can obscure meaningful insights. Effective preprocessing is crucial for filtering out this noise and involves steps such as tokenisation, stemming, and stop-word removal. Failure to adequately preprocess text data can lead to inaccurate feature representation and difficulties in model training [23].

Another significant hurdle is feature extraction. Converting text into numerical features is a complex task that can lead to high dimensionality, which complicates model training.

Techniques such as TF-IDF and word embeddings are commonly employed. However, as the number of features increases, the model may struggle to generalise effectively, leading to overfitting and reduced performance. This phenomenon, known as the curse of dimensionality, emphasises the need for careful feature selection and dimensionality-reduction strategies to maintain a balance between feature richness and model performance [6]. Furthermore, semantic interpretation also poses difficulties. The contextual meaning of words can vary significantly, making it challenging to derive consistent interpretations. Phenomena such as polysemy (words with multiple meanings) and synonymy (different words with similar meanings) can lead to ambiguity in feature representation. This complexity necessitates advanced methods for semantic analysis to ensure the model captures the text's intended meaning [24], [25].

Establishing appropriate hierarchical structures within the BHM that accurately reflect the relationships derived from text data is another complex endeavour. The challenge lies in defining clear parameters and dependencies that are both interpretable and statistically robust[26]. This lack of clarity can hinder the model's effectiveness at capturing data traces. Moreover, the integration of topic modelling outputs with the inputs required for BHM remains largely unexplored. While prior investigations have identified effective topic models [5], linking these outputs to a BHM structure requires careful consideration of how to represent topics as factors in the model. This gap highlights the need for innovative approaches to align topic modelling outputs with BHM requirements.

Interpretability is another significant challenge. The probabilistic nature of BHMs can make it difficult to extract actionable insights from model outputs derived from unstructured text data. Ensuring that the results are interpretable and meaningful to stakeholders is essential. Thus, developing strategies to enhance interpretability is crucial for the practical use of these models [27].

Overall, existing literature demonstrates significant advances in topic modelling, semantic analysis, and Bayesian hierarchical modelling for analysing unstructured textual data. Prior studies have shown the effectiveness of topic modelling approaches in extracting latent themes from large textual corpora, while BHM provides interpretable probabilistic structures for modelling uncertainty and hierarchical relationships. However, most existing studies focus primarily on topic extraction or semantic interpretation independently, with limited attention given to systematically transforming latent textual themes into interpretable hierarchical probabilistic risk representations. Although recent advances in LLMs enable context-aware semantic interpretation, their integration within Bayesian hierarchical frameworks for business risk analysis remains underexplored.

To address this gap, this study proposes an NLP-BHM framework that integrates topic modelling, semantic labelling, and Bayesian hierarchical modelling to enable interpretable probabilistic business risk analysis from unstructured business news data.

III. Integrating Topic Modelling Outputs into Bayesian Hierarchical Models

This study presents a comprehensive framework that bridges the output from the best topic model into a Bayesian Hierarchical Model (BHM) to support probabilistic business risk modelling from unstructured business news articles. Building on prior research that identifies Non-negative Matrix Factorization (NMF) as the best-performing topic model [4], this study focuses on integrating it into the BHM structure. The framework proposes two approaches for generating interpretable topic labels, clustering related topics into higher-level factors, and modelling probabilistic relationships between these factors and associated risks. The dataset comprises 35,667 Malaysian business news articles written in English published between 2019 and 2023, collected from multiple online portals known for business-related content in Malaysia. TABLE I. provides an overview of the number of articles collected between 2019 and 2023.

TABLE I. The Number of Articles Collected from 2019-2023

Year	Number of Articles
2019	5,979
2020	6,992
2021	8,496
2022	7,824
2023	6,376

A. Framework Overview

The NLP-BHM framework operates in two phases. The first phase involves labelling topics extracted from a topic model. Each topic, represented by its top keywords, is assigned interpretable labels using two complementary strategies: LLM prompting and WordNet-based semantic analysis. The LLM approach uses OpenAI's GPT-3.5-turbo-16k to generate context-aware labels, while the WordNet approach employs semantic similarity measures and hypernym extraction to derive linguistically consistent labels. Each method is carried out independently, which allows for systematic comparison between the two approaches. The primary objective of the proposed NLP-BHM framework is to transform latent textual themes into interpretable probabilistic risk representations through hierarchical semantic structuring and Bayesian reasoning. Fig. 2 presents the proposed framework of our study.

B. Selecting Topic Model Outputs

Four topic modelling methods (LDA, NMF, Top2Vec, and BERTopic) were evaluated using coherence scores across yearly datasets from 2019 to 2023. NMF consistently achieved the highest coherence scores (0.65–0.80), indicating superior semantic consistency and interpretability for the Malaysian business news corpus. Consequently, NMF was selected as the topic-extraction component of the proposed NLP-BHM framework.

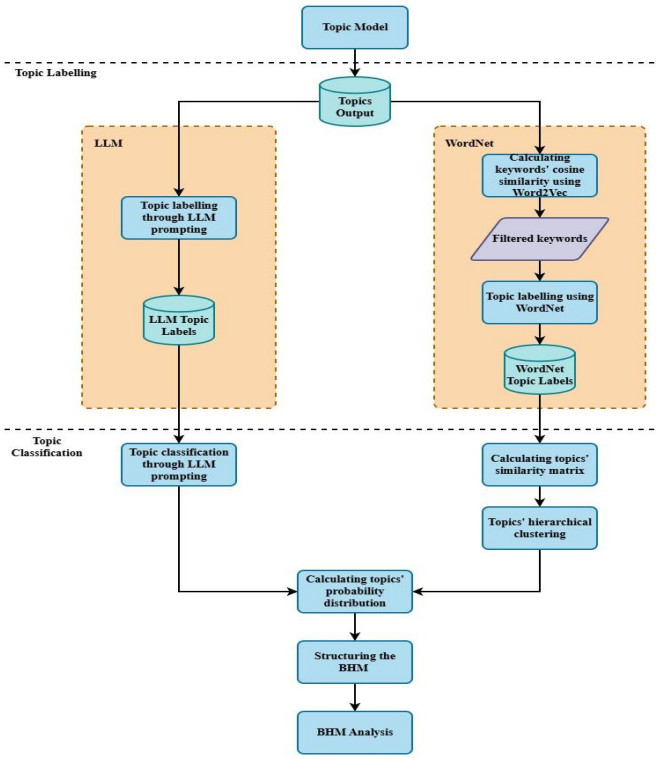

Fig. 2. An NLP-BHM framework for probabilistic business risk modelling.

C. Large Language Model Implementation

To generate interpretable topic labels, the LLM approach was employed with OpenAI's GPT-3.5-turbo-16k. For each topic extracted via NMF, the top keywords were provided to the LLM with a structured prompt instructing the model to suggest a single-word label relevant to business risk. The following prompt was used:

"I have a list of keywords as follows: {keywords}. Based on these keywords, can you suggest a single-word topic label related to risk?"

To enhance reproducibility, the API was configured with a fixed random seed, a temperature of 0.7, and a maximum token limit of 150. A second round of prompting was then applied to cluster the first-round labels into broader risk categories, using the following instruction:

"I have a list of topic labels as follows: {topic labels}. Based on these labels, can you cluster the topics together and suggest a name for each cluster?"

This two-step process helped us to transform latent topics into semantically meaningful factors and risk categories, which were then used to build the two-layer BHM structure.

D. WordNet Implementation

The WordNet-based labelling procedure involved multiple stages to ensure linguistic and semantic consistency. In the first phase, top keywords from each NMF-extracted topic were refined using cosine similarity derived from Word2Vec embeddings to eliminate semantically redundant or weakly related terms. The remaining keywords were then mapped to WordNet synsets, and their hypernyms were extracted to generate the initial topic labels. This process ensured that each label captured the core semantic meaning of its associated topic while maintaining linguistic coherence across topics. A similarity matrix was constructed from the initial topic labels using Wu-Palmer similarity, followed by hierarchical clustering with average linkage to group semantically related topics. Hypernyms were again extracted from each cluster to produce representative cluster-level labels, which were mapped as latent risks in the second layer of the BHM. This hierarchical organisation allowed the model to represent risk categories in a structured and probabilistically interpretable manner.

E. Mean Opinion Scoring Evaluation

Human evaluation remains important for assessing semantic quality and interpretability in NLP tasks, particularly where automatic evaluation metrics may not sufficiently capture contextual meaning and semantic coherence [28]. Therefore, a Mean Opinion Score (MOS) evaluation was conducted to assess the semantic clarity, contextual relevance, and interpretability of the topic labels generated by the LLM- and WordNet-based approaches. The quality of the generated labels was evaluated based on the level of understandability and interpretability. Understandability refers to how easily a label can be read and comprehended by non-specialists, while interpretability reflects how well the label captures the meaning and intent of the topic keywords in a risk-related context. Before the MOS evaluation, 20 respondents were recruited to assess the quality of the generated topic labels. The participants included a balanced representation of genders, ages, and professional statuses. This diverse composition provided a broad perspective on how different backgrounds may influence the perception of label clarity and relevance.

Each was presented with sets of topic keywords and the corresponding labels generated by the LLM- and WordNet-based approaches, while expert-assigned labels were included as a reference benchmark. Respondents were instructed to rate each label on a 5-point Likert scale, where 1 indicated a very poor fit with the topic keywords (unclear or misleading label), and 5 indicated an excellent fit (clear, relevant, and interpretable label). The evaluation was conducted via a Google Form that displayed the topic–keyword pairs alongside the candidate labels in a structured format. Respondents can enter their scores directly, ensuring consistency and ease of submission. The collected ratings were aggregated and averaged to compute the mean opinion score for each labelling method.

F. BHM Formulation

Drawing on the hierarchical architecture proposed by [4] for supply chain risk assessment, this study adopts a two-layer BHM tailored for high-dimensional textual data streams. While the framework in [4] relies on structured, predefined variables such as credit ratings and delivery metrics, our proposed model dynamically parameterises the hierarchy using latent topics (z_k) extracted from business news articles. These topics are subsequently mapped to higher-level business risk categories (r_i) through a context-aware taxonomic aggregation layer facilitated by an LLM- and WordNet-based semantic labelling. This two-layer structure allows the BHM to model uncertainty at both the thematic (topic) and categorical (risk) levels, providing the

flexibility to incorporate emerging risk factors that are typically omitted in traditional, static risk frameworks.

Let z_k denote an observed topic variable extracted from the news corpus using topic modelling, where k =1, …, K. These variables form the first layer of the hierarchy and represent the observable thematic signals derived from the textual corpus. Let r_i denote a business risk category, where i =1, …, I, represents the second layer of the hierarchy. At the topic level, the distribution of topics is represented as:

$$z_k \sim \text{Categorical}(\pi) \quad (1)$$

where, $\boldsymbol{\pi} = (\pi_1, \pi_2 \ldots, \pi_K)$ represents the topic probability distribution estimated from NMF-derived topic weights. At the risk level, each business risk category is conditionally dependent on the observed topic structure:

$$r_i \sim \text{Categorical}(\theta) \quad (2)$$

where, θ represents the probability distribution linking latent topics to higher-level business risk categories. This formulation captures how groups of latent topics contribute to higher-level business risk categories. This hierarchical design separates topic-level signals from risk-level representations, ensuring interpretability by explicitly modelling how individual topics influence higher-level business risk categories.

The probabilistic dependencies defined within the hierarchical structure are parameterised using topic-weight distributions and semantic associations derived from the topic labelling process. Following the Bayesian aggregation strategy proposed in [4], probability estimates are propagated through the hierarchy to quantify the relative influence of latent topics on higher-level business risk categories.

G. Obtaining Prior Probabilities

Initially, the NMF model produced weights for each keyword within a topic, representing the strength of the keyword's contribution to that topic. For each topic, the mean keyword weight was computed and subsequently used as the first-level probability estimate within the Bayesian structure. Following the strategy adopted in [4], these mean topic weights serve as the prior probability estimates for the corresponding latent topics. The prior probability for topic z_k is calculated as:

$$\pi_k = \frac{1}{n_k} \sum_{j=1}^{n_k} W_{kj} \quad (3)$$

where, W_{kj} represents the NMF weight of each keyword j within topic k,

n_k is the total number of keywords within topic k,

and π_k represents the prior probability estimate of topic k.

H. Parameter Estimation and Bayesian Risk Aggregation

The topic probability distribution π was estimated using the mean NMF keyword weights obtained from Eq. (3). Each topic probability reflects the relative prevalence of a latent topic within the business news corpus and serves as the initial probabilistic representation at the factor level of the hierarchy. The relationship between latent topics and business risk categories was established through semantic clustering generated using the LLM- and WordNet-based labelling approaches. Topics assigned to the same semantic cluster were grouped to form higher-level business risk categories.

Following the hierarchical aggregation approach proposed in [4], the probability of a business risk category, r_i was estimated by aggregating the probabilities of its associated topics:

$$P(r_i) = \frac{\sum_{k=1}^{m} \pi_k}{m} \quad (4)$$

where:

- π_k denotes the probability of topic k,
- m denotes the number of topics associated with risk category r_i.

Consistent with [4], the contribution weight associated with topic k was assumed to be equal to 1. This aggregation produces a probability estimate for each business risk category by combining the contributions of semantically related latent topics. Subsequently, the overall Business Risk Impact (BRI) score was computed by aggregating the estimated probabilities of all business risk categories, again assuming the contribution weight assigned to the risk category is equal to 1:

$$BRI = \frac{\sum_{i=1}^{I} P(r_i)}{I} \quad (5)$$

where:

- $P(r_i)$ denotes the probability of business risk categories i,
- I denote the total number of business risk categories.

The resulting BRI score provides a comparative measure of the overall business risk represented within each annual news corpus. To maintain computational efficiency and interpretability, probability estimates were obtained through hierarchical aggregation of topic probabilities and semantic cluster associations, following the Bayesian hierarchical framework in [4]. The resulting probability estimates form the basis of the comparative business risk analysis presented in the Results section.

IV. Results and Discussion

This section presents a comparison of LLM- and WordNet-based labelling integrated into a BHM. The results are organised to demonstrate each method's contribution to the NLP-BHM framework. Due to space constraints, the final BHM structures for the 2020 and 2021 data are illustrated in the figures.

A. Comparison of LLM and WordNet Labelling Effectiveness Through Mean Opinion Scoring

Fig. 3 shows the MOS results for 2019 topics generated by LLM and identified by WordNet.

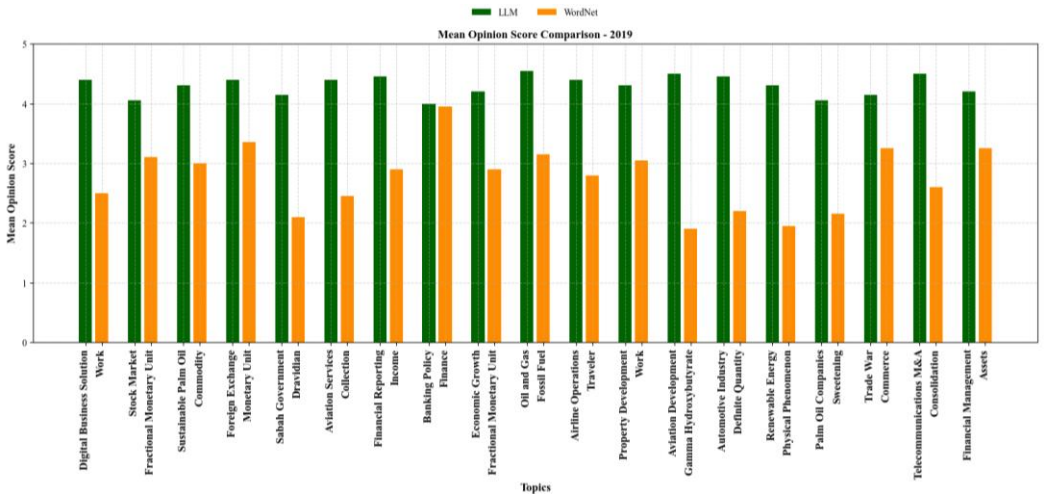

Fig. 3. Mean Opinion Scores from respondents to vote on LLM and WordNet topics from 2019 data.

The results indicated that respondents consistently favoured LLM labels (dark green), which outperformed the WordNet-based labels (orange). The LLM labels were rated as more understandable and better aligned with the business risk context. In contrast, WordNet labels were often perceived as too broad or disconnected from the underlying risk themes. Similar MOS results are observed in the 2020 to 2023 data, consistent with the 2019 data and highlighting patterns favouring LLM labels over WordNet labels. TABLE TABLE II. shows the average MOS results for 2019-2023 topics.

TABLE II. AVERAGE MOS SCORES FOR THE TOPICS GENERATED BY LLM AND IDENTIFIED BY WORDNET

Year	Average MOS (LLM)	Average MOS (WordNet)
2019	4.303	2.766
2020	4.288	2.683
2021	4.358	2.721
2022	4.369	2.634
2023	4.316	2.729

Over the five years, the average MOS for LLM topics increased from 4.2 to 4.4, reflecting a growing appreciation for their clarity and relevance. Conversely, WordNet topics maintained lower scores, ranging from 2.6 to 2.8, as they were often seen as too broad and less connected to specific risks. This trend emphasises the effectiveness of LLMs in enhancing topic interpretation and relevance in business contexts.

B. LLM-Based Labelling and Clustering Results

The factors and risks derived from the LLM topics were initially assigned prior probabilities based on the topics' weights from the NMF topic model. These priors represent the baseline importance of each factor before considering any risk evidence. Using the BHM, these priors were updated into posterior probabilities, reflecting the likelihood of specific risks once related factors are considered. The BHM structures for 2020 and 2021, as depicted in Fig. 4 and Fig. 5, reveal evolving business risk landscapes. In 2020, the focus was on "Transportation and Travel" and "Finance and Economy," with finance-related topics heavily influencing risk perception. By 2021, attention shifted to "Industry Analysis" and "Business and Finance," indicating a heightened emphasis on sectors like aviation and automotive. The business risk impact score increased from 0.408 to 0.444, reflecting heightened concerns, particularly in finance and governance. The reported risk impact scores represent relative probabilistic risk representations derived from posterior topic–risk associations within the BHM framework, rather than direct forecasts of future business outcomes. This evolution highlights shifting priorities and perceived risks in the business environment over the years.

Similar thematic transitions were observed across the 2019, 2022, and 2023 datasets, reflecting evolving business priorities and emerging economic concerns over time. TABLE III. shows the overall business risk impact score obtained from the LLM-based BHM structuring.

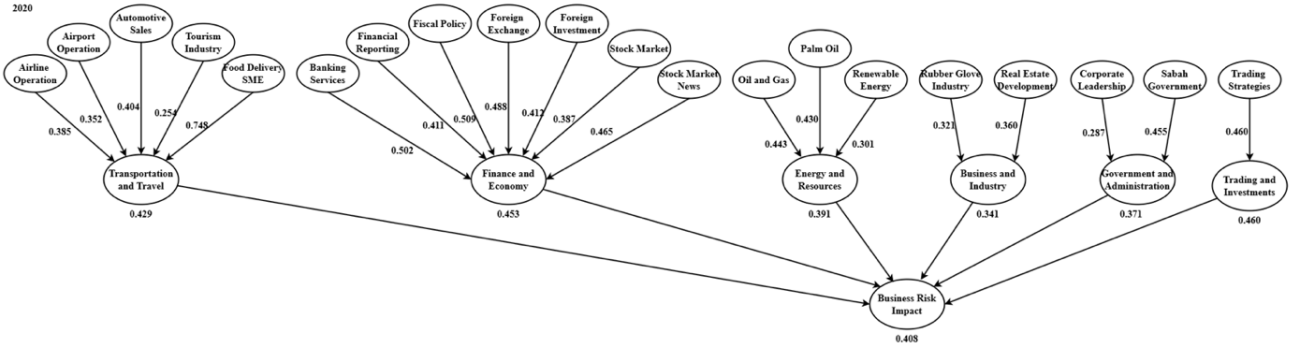

Fig. 4. LLM-based BHM structure for 2020 business news data.

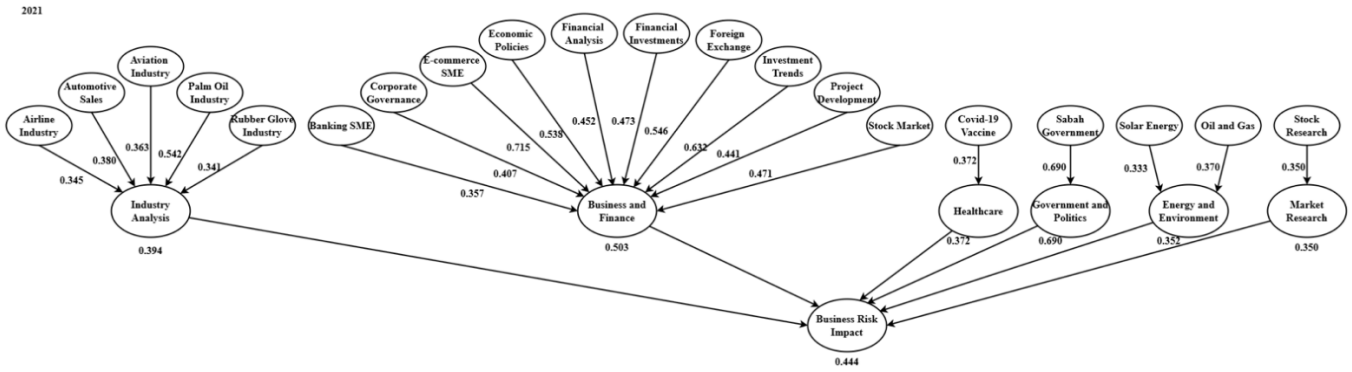

Fig. 5. LLM-based BHM structure for 2021 business news data.

TABLE III. OVERALL BUSINESS RISK IMPACT SCORE FOR LLM-BASED BHM STRUCTURE

Year	Business Risk Impact
2019	0.492
2020	0.408
2021	0.444
2022	0.430
2023	0.458

As shown in Table III, the LLM-based BHM structure produced business risk impact scores ranging from 0.408 to 0.492 across the study period. The highest risk impact score was observed in 2019 (0.492), followed by a decline in 2020 (0.408), before gradually increasing to 0.444 in 2021, 0.430 in 2022, and 0.458 in 2023. These comparisons illustrate evolving business priorities, reflecting changes in the economic landscape and policy environment over the years. These results highlight the NLP-BHM's ability to capture both persistent and emerging risks in Malaysia's business landscape.

C. WordNet-Based Labelling and Clustering Results

The WordNet-based Wu–Palmer similarity scores were used to construct a semantic similarity matrix representing the conceptual relatedness among topic labels. Higher similarity values indicated stronger semantic associations between topics, while lower values reflected weaker conceptual relationships. The similarity matrix was subsequently transformed into a distance matrix to support hierarchical clustering using the average-linkage method.

Fig. 6 depicts the resulting dendrogram for the sample, which highlights the semantic organisation of the topics in 2020. The height of each merge reflects the degree of dissimilarity between clusters, allowing for interpretation of how closely or distantly related the topics are. By applying a predefined threshold of 0.7, distinct semantic groups are identified, enabling clearer interpretation of the underlying conceptual structure derived from WordNet. It is observed that in that particular year, the topics were divided into four groups. After WordNet-labelled labels were processed through a similarity matrix and hierarchical clustering to group the factors into appropriate risk labels, the BHM's priors were estimated to reflect how common each factor was in the data, and then posteriors were calculated for the risks associated with each factor.

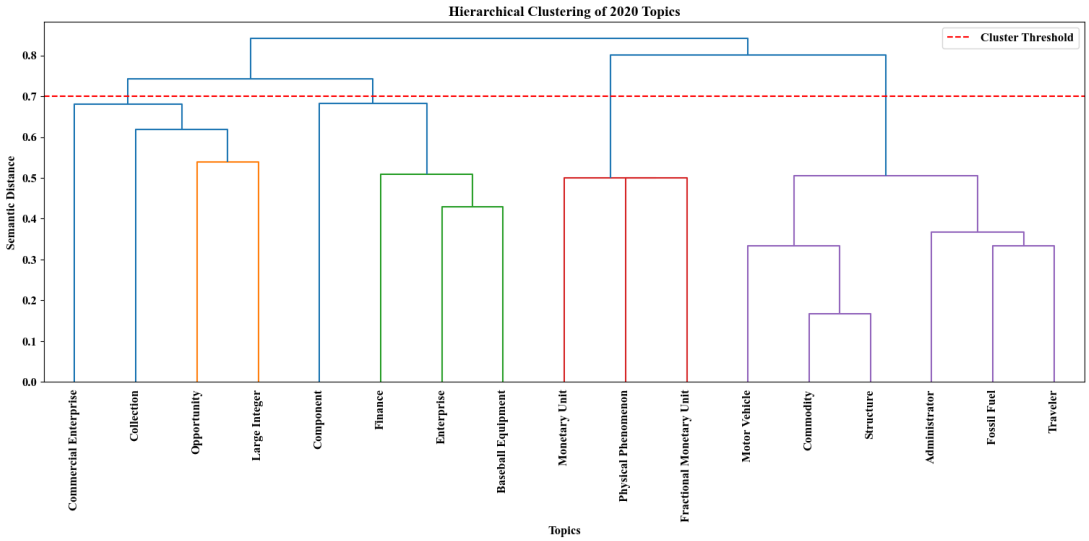

Fig. 6. Result of hierarchical clustering for topics in the 2020 data.

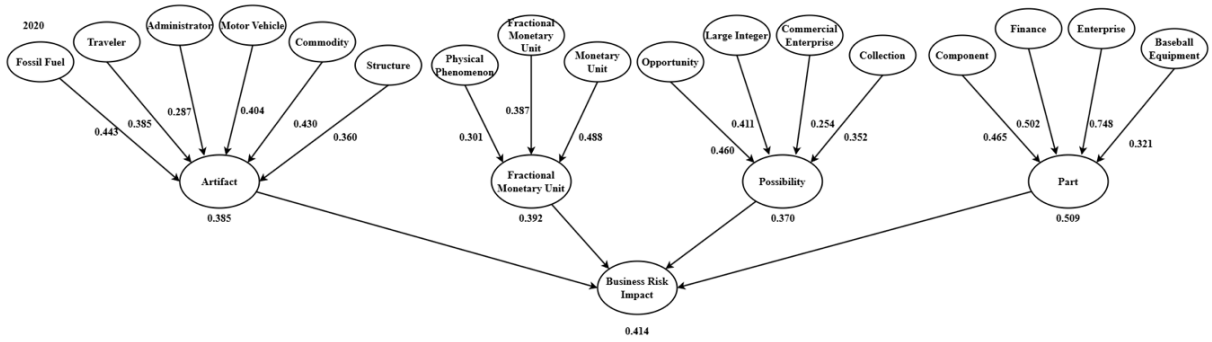

Fig. 7. WordNet-based BHM structure for 2020 business news data.

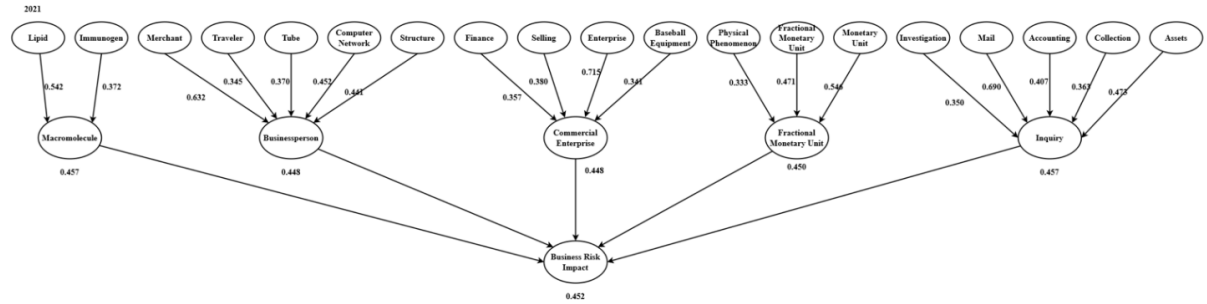

Fig. 8. WordNet-based BHM structure for 2021 business news data.

Fig. 7 and Fig. 8 show the resulting 2020 and 2021 BHM structures, respectively, obtained by leveraging WordNet-based labelling. The WordNet-based structures for 2020 and 2021 reveal distinct thematic focuses on business risk analysis. In 2020, the structure highlights categories such as "Artifact" and "Part," emphasising elements like fossil fuels, commodities, and financial components, leading to a business risk impact of 0.414. By 2021, the focus shifts to "Businessperson" and "Inquiry," with themes involving finance, enterprise, and accounting, reflecting a comprehensive approach to commercial enterprise risks with an impact score of 0.452. These structures illustrate evolving priorities in risk management, with a transition from tangible assets to more abstract business and financial concerns. Similar semantic transitions were observed across the remaining yearly datasets, where the WordNet-based structures progressively shifted from tangible asset-oriented concepts toward broader business and financial categories. TABLE IV. shows the overall business risk impact score obtained from the WordNet-based BHM structuring.

TABLE IV. OVERALL BUSINESS RISK IMPACT SCORE FOR WORDNET-BASED BHM STRUCTURE

Year	**Business Risk Impact**
2019	0.423
2020	0.414
2021	0.452
2022	0.416
2023	0.451

As shown in Table IV, the WordNet-based BHM structure produced business risk impact scores ranging from 0.414 to 0.452 over the five-year study period. The highest risk impact score was recorded in 2021 (0.452), followed closely by 2023 (0.451), while the lowest score was observed in 2020 (0.414). The remaining years, 2019 and 2022, yielded scores of 0.423 and 0.416, respectively. The variation in risk impact scores across the five years is relatively small, with a difference of only 0.038 between the highest and lowest values.

D. Discussion

The proposed NLP-BHM framework represents an integrated approach that combines topic modelling, BHM, and language model techniques to transform unstructured business news into structured, actionable risk insights. By integrating probabilistic reasoning with semantic enrichment, the framework addresses the complexity and uncertainty inherent in business risk analysis, enabling a more structured interpretation of latent risk factors. A key strength of the framework is the dynamic nature of the BHM, which allows risk probabilities to be continuously updated as new information becomes available. This makes the framework suitable for fast-changing business environments and supports near real-time risk assessment and decision-making.

The hierarchical structure of the model further improves interpretability by explicitly modelling relationships between high-level risks and underlying sub-factors, allowing clearer tracing of dependencies within complex risk systems. In addition, the use of LLM-based topic labelling enhances semantic representation by producing contextually meaningful, human-interpretable labels, thereby improving the usability of extracted risk topics. Based on the MOS results, LLM-based labelling is perceived as more interpretable than WordNet-based labelling, indicating its strength in capturing richer semantic context in business news. Importantly, in this study, LLM

outputs are made consistent through a fixed random seed, ensuring reproducibility and addressing concerns regarding variability in language model generation. In contrast, WordNet-based labelling provides deterministic, consistent labels that support reproducibility and standardisation. However, its limited coverage of domain-specific vocabulary, particularly in business and financial contexts, often leads to overly general topic labels that reduce interpretability and granularity. As a result, WordNet is less effective in distinguishing fine-grained risk categories, although it remains useful as a stable baseline for structured comparison.

It should be noted that the resulting business risk impact scores represent relative risk indicators derived from textual topic distributions rather than direct measures of real-world business risk severity. The framework quantifies the prominence of risk-related themes within the news corpus and enables comparative analysis across different periods. As such, higher scores indicate a greater concentration of risk-related topics within the analysed news data, rather than a verified increase in actual business risk outcomes. The current framework terminates at the estimation of business risk impact. However, these probabilistic indicators could be further served as inputs to more advanced risk quantification and decision-support models. Similar to the approach adopted in [4], where disruption probabilities were further utilised to estimate supplier revenue impact and Value at Risk (VaR), the proposed business risk impact scores could potentially be integrated with financial, operational, or market-based indicators to quantify the potential consequences of emerging business risks. Consequently, the business risk impact score should be viewed not only as a comparative textual risk indicator but also as a foundation for future quantitative risk assessment frameworks that link news-derived risk signals with real-world business outcomes.

V. Conclusion and Future Work

This study proposed an NLP-BHM framework for transforming topic modelling outputs into interpretable hierarchical probabilistic representations for business risk analysis. Using a corpus of Malaysian business news articles published between 2019 and 2023, four topic modelling approaches were evaluated, where NMF consistently produced the most coherent and semantically interpretable topic structures. The framework subsequently integrated semantic topic labelling and clustering using both LLM-based prompting and WordNet-based semantic analysis within a Bayesian hierarchical modelling structure.

The study demonstrates that topic modelling outputs can be effectively transformed into interpretable hierarchical probabilistic risk representations through semantic labelling and Bayesian hierarchical structuring, thereby addressing the challenge of integrating unstructured textual themes into uncertainty-aware business risk analysis. Comparative findings further show that LLM-based semantic labelling generated more contextually relevant and interpretable topic representations than the WordNet-based approach, as reflected by consistently higher MOS scores. While WordNet-based clustering provided semantically structured and reproducible groupings, its limited contextual flexibility reduced its effectiveness in capturing domain-specific business terminology and evolving semantic relationships.

Overall, the proposed NLP-BHM framework provides a structured approach to bridging latent textual themes and higher-level probabilistic risk representations through hierarchical semantic structuring and Bayesian reasoning. By integrating topic modelling, semantic interpretation, and posterior risk representation within a unified framework, the study demonstrates the potential of NLP-driven probabilistic analysis for interpretable business intelligence and decision support.

Future work may explore integrating dynamic temporal Bayesian structures, automated probabilistic parameter learning, and multiple LLMs to evaluate the generalisation ability, robustness, and consistency of semantic topic labelling across different model architectures. In addition, further research should investigate the validation of the generated business risk indicators against external economic and financial benchmarks, such as market volatility indices, business confidence indicators, or observed business risk events, to provide stronger empirical evidence for interpreting the resulting risk impact scores as meaningful measures of real-world business risk severity and to enhance the practical applicability of the proposed NLP-BHM framework.

Acknowledgment

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2023/ICT06/UNIMAS/02/2). The authors thank the MOS participants. The survey was approved by the UNIMAS Human Research Ethics Committee (HREC(NM)/2023(2)/57).

Declaration on Generative AI

This work utilises OpenAI's ChatGPT to enhance data analysis and interpretation. All outputs generated through the technology have been critically reviewed and validated by the authors to ensure accuracy and integrity.

References

[1] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Frontiers in Sociology, vol. 7, May 2022, doi: 10.3389/fsoc.2022.886498.

[2] G. M. Allenby, P. E. Rossi, and R. E. Mcculloch, "Hierarchical Bayes Models: A Practitioners Guide Hierarchical Bayes Models," 2005.

[3] M. Veenman, A. M. Stefan, and J. M. Haaf, "Bayesian hierarchical modeling: an introduction and reassessment," Behav. Res. Methods, May 2023, doi: 10.3758/s13428-023-02204-3.

[4] S. M. Ali, A. B. M. M. Bari, A. A. M. Rifat, M. Alharbi, S. Choudhary, and S. Luthra, "Modelling supply chain disruption analytics under insufficient data: A decision support system based on Bayesian hierarchical approach," International Journal of Information Management Data Insights, vol. 2, no. 2, Nov. 2022, doi: 10.1016/j.jjimei.2022.100121.

[5] V. Asas, S. S. Juan, V. C. Kerbun, S. Chua, J. Labadin, and E. Lau, "A Comparative Analysis of Topic Modelling Techniques for Malaysian Business News Data: LDA, NMF, Top2Vec, and BERTopic," in 2025 14th International Conference on Information Technology in Asia (CITA), IEEE, Aug. 2025, pp. 42–47. doi: 10.1109/CITA66455.2025.11198734.

[6] D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Michael I. Jordan," 2003.

[7] L. Kong and A. Brintrup, "A hierarchical Bayesian model for payment delay prediction in supply chain financing," Int. J. Prod. Res., pp. 1–24, Oct. 2025, doi: 10.1080/00207543.2025.2546029.

[8] M. Coll, M. Grazia Pennino, J. Steenbeek, J. Sole, and J. M. Bellido, "Predicting marine species distributions: Complementarity of food-web and Bayesian hierarchical modelling approaches," Ecol. Modell., vol. 405, pp. 86–101, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.05.005.

[9] A. Couture et al., "Estimating COVID-19 Hospitalizations in the United States with Surveillance Data Using a Bayesian Hierarchical Model: Modeling Study," JMIR Public Health Surveill., vol. 8, no. 6, 2022, doi: 10.2196/34296.

[10] H. Zeng et al., "Climate-informed clustering based nonstationary regional extreme flood events spatio-temporal evolution using hierarchical Bayesian modeling," J. Hydrol. Reg. Stud., vol. 56, Dec. 2024, doi: 10.1016/j.ejrh.2024.102066.

[11] A. Talman, H. Celikkanat, S. Virpioja, M. Heinonen, and J. Tiedemann, "Uncertainty-Aware Natural Language Inference with Stochastic Weight Averaging," in Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), T. Alumäe and M. Fishel, Eds., Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 358–365. [Online]. Available: https://aclanthology.org/2023.nodalida-1.37/

[12] D. Yu and B. Xiang, "Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling," Expert Syst. Appl., vol. 225, Sep. 2023, doi: 10.1016/j.eswa.2023.120114.

[13] R. Vangara et al., "Finding the Number of Latent Topics With Semantic Non-Negative Matrix Factorization," IEEE Access, vol. 9, pp. 117217–117231, 2021, doi: 10.1109/ACCESS.2021.3106879.

[14] D. Angelov, "Top2Vec: Distributed Representations of Topics," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.09470

[15] S. S. Pangastuti, E. N. Rohmatullayaly, and N. Najmi, "TOPIC MODELING FOR USER FEEDBACK DATASET," Communications in Mathematical Biology and Neuroscience, vol. 2025, 2025, doi: 10.28919/cmbn/8932.

[16] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.05794

[17] O. Lezhnina, "Depression, anxiety, and burnout in academia: topic modeling of PubMed abstracts," Front. Res. Metr. Anal., vol. 8, 2023, doi: 10.3389/frma.2023.1271385.

[18] I. Widiastuti and H. S. Yong, "TR-GPT-CF: A Topic Refinement Method Using GPT and Coherence Filtering," Applied Sciences (Switzerland), vol. 15, no. 4, Feb. 2025, doi: 10.3390/app15041962.

[19] S. Minaee et al., "Large Language Models: A Survey," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2402.06196

[20] Y. Li, S. Wang, H. Ding, and H. Chen, "Large Language Models in Finance: A Survey," in ICAIF 2023 - 4th ACM International Conference on AI in Finance, Association for Computing Machinery, Inc, Nov. 2023, pp. 374–382. doi: 10.1145/3604237.3626869.

[21] G. A. Miller, "WordNet: a lexical database for English," Commun. ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995, doi: 10.1145/219717.219748.

[22] Z. Wu and M. Palmer, "Verb Semantics and Lexical Selection," in 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, USA: Association for Computational Linguistics, Jun. 1994, pp. 133–138. doi: 10.3115/981732.981751.

[23] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[24] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models, 3rd ed. 2026. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[25] D. Jurafsky and J. H. Martin, "Speech and Language Processing (2008)," vol. 1, 2008, doi: 10.1162/089120100750105975.

[26] R. Navigli, "Is Word Sense Disambiguation Dead in the LLM Era?," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 40, no. 46, pp. 39753–39762, Mar. 2026, doi: 10.1609/aaai.v40i46.41331.

[27] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," arXiv: Machine Learning, 2017, [Online]. Available: https://api.semanticscholar.org/CorpusID:11319376

[28] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "BLEU is Not Suitable for the Evaluation of Text Simplification," in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 936–941