Towards the development of an explainable e-commerce fake review index: An attribute analytics approach

Abstract:

Instruments of corporate risk and reputation assessment tools are quintessentially developed on structured quantitative data linked to financial ratios and macroeconomics. An emerging stream of studies has challenged this norm by demonstrating improved risk assessment and model prediction capabilities through unstructured textual corporate data. Fake online consumer reviews pose serious threats to a business' competitiveness and sales performance, directly impacting revenue, market share, brand reputation and even survivability. Research has shown that as little as three negative reviews can lead to a potential loss of 59.2% of customers. Amazon, as the largest e-commerce retail platform, hosts over 85,000 small-to-medium-size (SME) retailers (UK), selling over fifty percent of Amazon products worldwide. Despite Amazon's best efforts, fake reviews are a growing problem causing financial and reputational damage at a scale never seen before. While large corporations are better equipped to handle these problems more efficiently, SMEs become the biggest victims of these scam tactics. Following the principles of attribute (AA) and responsible (RA) analytics, we present a novel hybrid method for indexing enterprise risk that we call the Fake Review Index (R_{FRI}). The proposed modular approach benefits from a combination of structured review metadata and semantic topic index derived from unstructured product reviews. We further apply LIME to develop a Confidence Score, demonstrates the importance of explainability and openness in contemporary analytics within the OR domain. Transparency, explainability and simplicity of our roadmap to a hybrid modular approach offers an attractive entry platform for practitioners and managers from the industry.

Keywords: Fake Reviews, Amazon, Risk Analysis, AI Explainability, BERT, Topic Model Indexing, LIME Confidence Score

1.0 Introduction

A global e-retail giant like Amazon, with a turnover value of $576 billion in 2023, has claimed to increase their year-on-year sales by an average 14% in the post-pandemic world (Statista, 2024). The World Economic Forum (2021) reported that with the recent growth of e-commerce, online reviews accounted for $3.8 trillion worth of global e-retail expenditure in 2021. Buyers place immense trust in fellow buyer reviews while making critical purchase decisions. 93% of B2C buyers have admitted to being influenced by online reviews, while 92% of B2B buyers rely on trusted reviews before making purchase decisions (Kaemingk, 2020). An average UK household was found to spend £900 per year directly influenced by online reviews (Statista, 2022). Growing digital transformations have led to the emergence of new business risks like fake reviews. Yet, Business Failure Prediction (BFP) and risk assessment literature are heavily focused on macro factors like the economy, and micro factors like cost sensitivity and credit risk for modern-day risk assessment (Xu et al., 2019).

In current operations management research framework, fake reviews were described as reviews that exhibit embellishment qualities, purposefully exaggerating or defaming an organisation's image and operational legitimacy (Banerjee & Chua, 2023). Fake reviews are often strategically deployed to achieve unfair competitive advantages by directly uplifting corporate reputation, or distorting competitor(s') ranking in third-party e-commerce platforms like Amazon (Which, 2019). Large corporations, often equipped with substantial financial resources and sophisticated Online Reputation Management (ORM) strategies, are better positioned to mitigate the risks associated with fake reviews. In contrast, small and medium-sized enterprises (SMEs) are particularly vulnerable to these scams. In contrast, small to medium-sized businesses (SMEs) have become the biggest victims of these scams. An online e-commerce platform like Amazon hosts more than 85,000 SMEs, just in the UK. Together, these enterprises sold 950 million products worldwide, reaching hundreds of millions of customers who, in turn, contribute millions of product reviews (Amazon, 2022). Despite Amazon's recent efforts, fake reviews remain a growing problem, directly affecting (positively or negatively) ranking, visibility and sales volume of strategically targeted products (BBC, 2023).

The recent advent of generative AI (GAI), has further complicated the issue of consumer trust and the integrity of online reviews. As GAI's ability to generate realistic and convincing fake reviews increases, concerns about their proliferation have risen amongst businesses and consumers, leading to hesitancy and eroding trust in online marketplaces. Empirical studies have corroborated the close association between the abundance of GAI fake reviews and declining trust in organisations (Dwivedi et al., 2023). Distinguishing between company-generated reviews and those dishonestly generated by competitors or individuals is crucial for BFP literature and modern-day enterprise risk assessment. The former undermines overall market trust, while the latter can diminish the prospect of a promising business.

The increasing number of fake reviews is a significant threat to an enterprise (Jabeur et al., 2023). Hence, e-commerce operations managers and risk assessors are increasingly interested in online reviews and their impact on organisational risk assessment (Luo et al., 2023). Early research by Mayzlin et al. (2012) highlights the inherent corporate incentive to generate favourable fake reviews, blurring the lines between legitimate information management and deceptive online review manipulation, which ultimately erodes consumer trust. In contrast, competitors may use fake negative reviews to manipulate search results and gain an unfair competitive advantage in terms of reputation building. Examining both positive and negative fake reviews on a travel e-commerce platform, Zervas et al. (2021) pointed out how these deceptive practices have a scientifically demonstrable negative impact, including reputational damage (decreased user trust), lost sales (lower booking rates), and potential legal issues (regulatory

fines). Buyers are willing to spend 31% more on a business with excellent reviews, and, falsely increasing one review star rating is shown to increase 5%-9% organisational revenue (Luca, 2016). On the other hand, false reviews can severely diminish product, brand and corporate reputation. For small to medium-sized businesses (SMEs), operating on e-commerce platforms like Amazon, consequences are detrimental (World Economic Forum, 2021).

A recent stream of research has emphasised the importance of tackling online review manipulation for e-commerce platform sustainability and consumer right protection (Marsh 2023). Given the enterprise risk impact, fake reviews in the form of unstructured data have become an important field of research within traditional Online Reputation Management (ORM) frameworks, but the topic appears to be neglected within Business Failure Prediction (BFP) and corporate risk assessment literature – a point underscored by Stevenson, Mues, & Bravo (2021) and further supported by Borchert et al. (2023). Unstructured textual data is becoming increasingly important for corporate risk, reputation, and business failure modelling, given 80% of organisational data is now identified to be unstructured (Kriebel & Stitz, 2022). However, integrating NLP-based research with consumer voice, UGC, and quantitative corporate performance metrics is complex; and primarily disregarded in operations management research. Only a few studies have applied this concept using advanced NLP principles (Lu et al., 2023), yet none of them comprehensively considered consumer reviews (or fake reviews) as an important attribute metric.

Research on fake reviews have seen a relatively slow growth in the past ten years, with only 165 selected articles identified across 122 journals until 2019, from various discipline using systematic literature review (Wu et al., 2020). Since then, there has been a recent surge of research into fake reviews, and this stream of research has shown promising growth within the following clusters: (i) fake review intervention (detection, response strategy) (Barbado et al., 2019), (ii) fake review consequences (consumer decision) (He, Hollenbeck, & Proserpio, 2022), (iii) fake review antecedents (reviewer motivation, product life-cycle) (Jiang et al., 2020). Despite growing multidisciplinary interests, we identified five key research gaps within the stream: (i) Research context: most fake review research appears to be focused on the hospitality and restaurant industry due to the easy availability of data from third-party review websites like Yelp and TripAdvisor. For the e-commerce industry, especially for an organisation like Amazon, there is clear lack of standard labelled dataset; (ii) Lack of authentic labelled dataset: most studies create their own labelled datasets which leads to inconsistent evaluation and performance benchmarking (Salminen et al., 2022); (iii) Unstructured data, model interpretation, and tailored performance metrics: there is a noticeable over-reliance on structured data-driven fake review detection (model accuracy comparison) and consumer distrust modelling in current research (Bathla et al., 2022). Fake reviews provide a rich source of unstructured textual data that can improve prediction accuracy through advanced vector space modelling technique(s); (iv) Behavioural linguistic patterns: there is insufficient understanding of behavioural and psychological cues in fake reviews. While linguistic analysis alone may be inadequate for discerning subtle human behavioural patterns; (v) Organisational risk assessment: research on fake e-commerce reviews mostly appears to reside on consumer decision-making, distrust, and online reputation management domain rather than being linked to organisational vulnerability assessments (Presti & Maggiore, 2021).

Traditionally, fake review detection research has treated the phenomenon as a single, undifferentiated entity, neglecting the diverse motivations and actors behind its creation (Luo et al., 2023). This one-size-fits-all approach overlooks crucial distinctions, as recent studies emphasise (Hajek, Hikkerova, & Sahut, 2023). Recognising the difference between positive fakes commissioned by companies to inflate ratings and negative fakes employed by competitors to damage reputation is essential for understanding

the varied risks and developing effective detection strategies. Existing literature heavily focuses on fake positive reviews, exemplified by He, Hollenbeck, & Proserpio (2022). However, this emphasis neglects the distinct corporate risks posed by fake negative reviews, particularly for external stakeholders. Therefore, acknowledging and analysing the different types of fake reviews is crucial for comprehensively understanding corporate risk. Current research in this domain has also paid limited attention to the responsible application of analytics to support small and medium sized enterprises (SMEs). Large corporations have resources to mitigate fake review risks, but SMEs remain highly vulnerable without clear solutions for scalable and dedicated explainable NLP techniques.

To address this gap, this study proposes developing an explainable **Fake Review Index (**R_{FRI}**)** that can be incorporated into overall organisational risk assessments and crisis response strategies. The proposed FRI can detect fake reviews considering a series of structured metrics and unstructured text using hybrid machine learning and deep learning approaches that combine semantic analysis of textual data and other normative review features. Differentiating between positive and negative fake reviews is crucial from a corporate risk perspective. Accurately classifying fake reviews based on their likely intent equips organisations with vital context for comprehensive risk assessments and strategic decision-making.

In order to make the FRI interpretable for operations managers, model explanations are generated using a three-stage study describing how individual text fragments and structured attributes influence the fake review predictions. Furthermore, new metrics were devised to evaluate the confidence level of explanation quality using Confidence Score (CS). The proposed FRI (R_{FRI}) further benefits from credibly sourced enterprise-level labelled fake review training data, tested on original Amazon reviews gathered over a period of approximately 12.5 years. This approach helped us overcome the limitations of self-generated labelled data related to inconsistent benchmarking, as noted within previous studies. We address the following research questions (RQ) towards developing the fake review index.

RQ1: What metadata and contextual cues beyond textual content exhibit strong predictive capabilities for detecting fake reviews on Amazon (Study 1)?

RQ2: Can semantic indexing of linguistic patterns and cues, along with human interpretation, help to improve the predictive performance of an LSTM-based Fake Review (R_{FRI}) indexing model (Study 2 and 3)?

RQ3: How can important identified variables (from Study 1 and 2) be used to optimise a neural network architecture for improved fake review detection? How such LSTM model outputs can be explained and integrated into overall enterprise risk assessment (Study 3)?

These research questions combined aims to uncover hidden patterns of deception in fake reviews beyond textual content. Investigating *RQ1*, we seek to identify non-textual structured metadata based variables that are highly significant in describing fake review characteristics. Our study further aims to uncover new insights into the non-textual patterns of deception in online reviews by identifying non-textual cues within Amazon customer review data. This enhances the theoretical and empirical understanding of fake review writing behaviour beyond textual content. *RQ2* helps us explore cross-category variations in fake review language and semantic level thematisation of fake review linguistics. Outcome of the study promises to provide variations in fake review linguistic themes helping generate niche topic insights into each product category. This approach enables more nuanced risk assessments tailored to the dynamic fake review landscape, empowering better corporate decision-making. Finally, following *RQ3*, we propose to design and optimise a hybrid fake review risk indexing model; to advance

the protocol for leveraging representation learning, addressing the advancement of attribute and responsible analytics. Using an explainable attribute analytics approach, we aim to establish new benchmarks for fake review detection, enterprise risk assessment and managerial decision-making.

2.0 Related Literature

The proliferation of fake online reviews presents multifaceted risks that parallel established organisational risk taxonomies like financial, operational, competitive and regulatory parameters. However, academic literature has yet to substantially explore the dimensions of corporate risk posed by fake reviews across enterprises operating on e-commerce platforms. This paper proposes, a step-by-step explainable attribute analytics approach towards developing a fake review-oriented risk-indexing framework that compiles heuristic criteria innate to these types of reviews. Our approach to developing such a hybrid risk assessment and supportive decision-making tool addresses a common research gap that cuts across three important research streams in operations research: Explainable AI (XAI), Crisis Management and Business Failure Prediction (BFP), and Online Reputations Management.

Our analysis addresses two key gaps identified by De bock et al. (2023) in explainable AI and operations research (OR):

1. Lack of explainability in applying Analytics to OR problems: prior scholarship has not adequately highlighted transparent internal logics and mechanisms that shape AI and ML applications to the modern-day operations research business problems. This limitation obfuscates further application and theoretical development.

2. Limited focus on practical applications: research grounded outside operations research; primarily focuses on methodological advancements through 'sub-dimensions of explainability', neglecting the need to develop pragmatic solutions that can directly enhance organisational decision-making.

Study Positioning: we position the study by addressing De Bock et al.'s (2023) calls to advance research-led decision-making and managerial problem-solving in operations research by providing explainable insights into the ML/DL mechanisms by which predictions based decisions are made. In particular, we position this study within the Attributable Analytics (AA) and Responsible Analytics (RA) domains. In terms of AA, we endorse De Bock et al.'s (2023) criticism of recent research that has an obsession with maximising model performance metrics (AUC, ROC etc.); instead, we align this study to their appeal for the development of tailored performance metrics that focus on dimensions of organisational risk. Regarding RA, we acknowledge the - 'under-investigated' - extant literature involving operations research and explainable AI. In our endeavour, we propose to develop a novel e-commerce Fake Review Index (FRI); by combining a series of explainable and interpretable ML/DL-based analytics techniques that not only benefit from combined structured and unstructured information metrices, but the index can also be integrated with other weighted risk factors to assess organisational stability and Business Failure Prediction (BFP) in real-time relative to other risk assessments. Our proposed FRI contributes to responsible social analytics by offering competition fairness to small to medium-sized enterprises (SMEs), who often lack the knowledge or resources to combat fake reviews and assess risk impact on their business. By illuminating the inner workings of models, explainability in modern-day analytics is a vital tool for evaluating model validity, reducing uncertainty, and ensuring model comprehensibility, leading to improved ML/DL model adoption within organisations (De Bock et al., 2023).

Research on leveraging unstructured data for risk assessment is a developing avenue. Previous research in BFP primarily relied on structured data like financial ratios, while recent studies have recognised the hidden potentials of unstructured textual data. There is also a scarcity of research focusing on this particular topic for risk assessment within smaller businesses (Borchert et al., 2023). Few recent studies have acknowledged the importance of developing DL models for BFP, incorporating textual data, for better explainability and prediction power (Borchert et al., 2023; Mai et al., 2019). Although recent BFP research acknowledges the benefits of incorporating unstructured textual data, only a few explored the option to optimise model prediction capacity, using both, structured and unstructured data (Borchert et al., 2023; Mai et al., 2019).

Explainable Analytics and Fake Review Research: the prevalence of black-box models in operations research has increased the demand for model explanations among practitioners and decision-makers (Hassija et al., 2023). Explainable models allow operations research professionals to interpret model decisions, validate models, reduce uncertainty, and increase the adoption of complex machine learning methods across business processes. Model explainability enables practitioners and decision-makers to understand, trust, and productively leverage black-box model outputs. Previous research on fake reviews has attempted to apply various modelling techniques to uncover review credibility and falseness in understanding broad characteristics of review manipulation (Salminen et al., 2022). Relevant to our study, the potential of their ML models were developed through a series of planned computational analytics, that aided the resolution of intricate issues in assessing online retail review quality (Joung & Kim, 2023). However, this line of study suffers from large drawbacks that not only limit the quality of their output, but also encourages the reproduction of more and more comparative model accuracy validation work that can barely be translated into managerial decision making. Below we highlight some of the potential drawbacks of the current fake review research:

- Synthetic training data: overreliance on machine-generated training data potentially limits generalisability to real-world human-written fake reviews. Consider incorporating real-world data or data augmentation techniques.
- Focus beyond accuracy: while benchmarks are valuable, overemphasising accuracy without considering interpretability, fairness, and real-world applicability hinders progress.
- Limited feature usage: Explore non-textual metadata (e.g., reviewer length, review rating) to enhance detection capabilities.
- Underutilized NLP innovations: leverage recent advancements like self-supervised, contrastive, and semi-supervised learning for efficient feature extraction and representation.
- Comprehensive hyperparameter exploration: systematically analyse different model architectures, activation functions, and regularisers to optimise performance.
- Robust evaluation: move beyond simple 80-20 split and adopt rigorous validation methods like k-fold cross-validation.
- Deepen interpretability: utilise advanced techniques like LIME and SHAP to provide actionable insights into model explainability.
- Continuous learning: Explore methods for adaptive models that learn and update continuously with new data to maintain effectiveness in a dynamic environment.

Several aspects related to model architecture, hyperparameter tuning, training schemes, interpretability analysis and evaluation metrics could be explored further in fake review detection from a more explainable and attribute analytics perspective. De Bock et al. (2023) proposed that XAI should be considered a significant area of focus within the future of operations research, particularly with regard

to corporate risk assessment. The authors emphasised the importance of algorithm interpretability in assessing and mitigating emerging organisational risks while meeting the demands of both internal and external stakeholders. Darwish (2022) further argued that XAI is essential for ensuring that AI systems are used responsibly and ethically to mitigate corporate risks. In this context, this is essential to recall earlier work by Colley, Väänänen, & Häkkilä (2022) who introduced the concept of tangible explainable AI (TangXAI), which explores the use of physical artefacts and tangible user interfaces for explaining AI systems. The authors argue that TangXAI can provide a more effective and engaging way to explain AI systems, particularly to users unfamiliar with AI or who have difficulty understanding traditional text-based explanations. Their work was built on Arrieta et al. (2020), helping to lay the foundation for the XAI field alongside proposing a set of guidelines for evaluating the explainability of AI systems. Despite these foundational advancements, limited research has gone into developing an applied and explainable analytics solution to tackle fake reviews as a disinformation-tackling mechanism at an enterprise level. Adoption of an explainable artificial intelligence (XAI) philosophy in future research will make automated detection systems more understandable, interpretable, and applicable by managers and decision-makers. Therefore, the impetus for XAI is essential to increase trust and accountability of automated fake review detection in business practice.

A limited number of emerging studies have tested the effects of XAI interventions to deal with misinformation/disinformation problems. For example, Janssens et al. (2023) discussed the value of explainability in social media rumour detection. To accurately identify social media rumours, the authors constructed hybrid machine learning and deep learning models that leverage both unstructured textual data and structured contextual information. The textual data consists of the raw tweet content, while the structured data includes metadata. By combining these two data types, their custom models aim to enhance predictive performance compared to using single-dimensional data points. The authors applied LIME to generate interpretable explanations from these complex hybrid models. Such adoption of a model-agnostic method meaning the authors could explain underlying mechanism within the black-box of their machine learning and deep learning model. Crucially for this study, LIME enabled analysis of the impact of individual words and structured features on the predicted rumour probability. This approach provided fine-grained and human-understandable explanations about why a particular tweet was flagged as a rumour or not. Without LIME, the hybrid models would act as impenetrable black boxes, making it impossible to diagnose their behaviour. The model-agnostic nature of LIME allows for a fair comparison of the explanation quality across different machine learning and deep learning architectures (Stevenson, Mues, & Bravo, 2020). The authors' focus on interpretability and rigorous explanation evaluation promotes AI transparency alongside reliable and trustworthy AI systems that can enable better policy and better decision-making. While research like Janssens et al. (2023) demonstrates the efficacy of XAI-powered approaches for mitigating misinformation, a significant knowledge gap remains regarding modelling the intent behind fake reviews and their impact on their positive/negative distribution.

Unstructured Data-Driven Enterprise Risk Assessment: Janssens et al.'s (2023) work on XAI inspired us to explore its potential in bridging the gap between fake reviews and unstructured data-driven business risk assessments. XAI can offer valuable insights into distinguishing the types and motivations behind fake reviews, leading to more nuanced and effective risk assessments. Stevenson et al. (2021) supported this approach by highlighting the limitations of vector-based methods and advocating for deeper learning techniques like CNNs and RNNs. Building on their work with a large dataset of structured and unstructured data, they demonstrated the effectiveness of BERT and LIME in understanding the difference in remarks and decision-making factors used in assessing customer-led

business risks. This study provides pioneering insights into the importance of language model fine-tuning by extensively applying "Gradual unfreezing" and "Discriminative [learning rate] fine-tuning" (Howard & Ruder, 2018).

In contrast, Borchert et al. (2023) leveraged latent semantic indexing using a paragraph vector model. Borchert et al. (2023) used a semi-explainable approach in deploying pre-trained transformer models like BERT. In their study, BERT was used as an attention mechanism to create context-specific word embeddings, i.e., the same word gets different vector representations based on its context. Their study further employed Latent Semantic Indexing (LSI), a semantic indexing technique, to extract a compact yet informative representation of textual data. LSI was applied on the initial high-dimensional TF-IDF matrix obtained from the company websites. By using singular value decomposition, LSI values approximated the TF-IDF matrix in a lower dimensional space that preserved semantic relationships between terms. Reducing the number of unnecessary textual features this way provides two key benefits: first, it improves computational and storage efficiency for modelling compared to the sparse high-dimensional TF-IDF matrix; second, LSI's lower dimensional representation discards noise and focuses on the underlying semantic concepts. This approach resulted in a more meaningful representation of textual similarities compared to just matching surface term frequencies. Overall, the use of LSI semantically enriched lower-dimensional textual features improving the predictive modelling accuracy while reducing model complexity. This technique also helped the authors to capture conceptual relationships between topics rather than just word statistics. Overall, the study empirically demonstrated LSI's effectiveness by showing improved performance over raw TF-IDF features for business failure prediction. It further showed that for prediction tasks, semantic representations can unlock latent thematic patterns and non-linear relationships from text that word frequencies cannot capture well. It also empirically proves the superiority of semantic BERT features over TF-IDF for business failure prediction, validating the power and importance of semantic indexing.

On this note it is important to recall the limitations of Moon & Kamakura’s (2017) work, as this was one of the first studies in the domain that attempted to analyse expert wine reviews to linguistically understand better product positioning and product development strategies in marketing and operations research. Their approach is rather simplistic from a methodological perspective, as the authors followed a specialised dictionary-based BoW and part-of-speech (POS) tagging method. Although, the novelty of their work derived from unique topic cluster modelling where individual product characteristics (Sweet, Tannin, Fruity, Woody) were weighted and represented in three-dimensional vectorised space based on a unique topic and sentiment taxonomy designed by them. Moon & Kamakura's (2017) topic modelling approach assigned weights and vectors to product attributes based on the custom taxonomy, but it is rather unclear from their study how specific topic values were derived for topic mapping. More details on the weighting criteria and 3D vector representation of product attributes would improve model explainability and transparency.

Similarly, Chakraborty, Kim, & Sudhir (2022) developed a word-frequency-based deep learning convolutional long short-term memory (LSTM) hybrid model that was capable of converting unstructured open ended textual reviews into attribute level ratings using Yelp reviews. Although their CNN model converted reviews to ratings, the logic behind such a feature was deeply obscured within the ‘black-box’ of the model. Both Moon & Kamakura (2017) and Chakraborty et al. (2022) highlight the potential of customer review analysis for extracting valuable business insights. However, both studies share limitations regarding explainability, interpretability and generalisability due to the undescribed nature of their LSTM black-box. Our study address such limitations in detail and proposes solutions for enhancing explainability and generalisability in the context of fake review risk indexing.

Table 1:
Studies that applied unstructured textual data towards organisational risk assessment through advanced NLP

Study	Summary	Industry	Data Source	Explanability	Applied NLP		
					Vector Space	Neural Network	Semantic Analysis
Text based Operational Risk Assessment							
Stevenson, Mues, Bravo (2021)	Value of textual data towards small business default prediction	Specialised Micro and SME lenders	Loan dataset from an m-SME specialised lender	No		✓ BERT	✓ LIME
Borchert et al. (2023)	Business failure prediction integrating unstructured textual website content.	Finance, Accounting and Services	Corporate website content	Yes Call for better model explainability	✓ Doc2Vec	✓ CNN BERT	✓ Latent Semantic Indexing
Review based Corporate Risk Assessment							
Chakraborty, Kim, & Sudhir (2022)	LSTM based transformation of online review text into attribute summary rating.	Hospitality, Service, Restaurants	Yelp restaurant reviews	No		✓ Hybrid LSTM	
Moon and Kamakura (2017)	Translating product reviews into vector space product positioning map	Wines, Hotels	Expert wine reviews and meta reviews	No	✓ POS BoW Word taxonomy based vector space mapping		
This Study	Developing an E-Commerce Fake Review Index (FRI)	E-commerce (e-retail)	Original Amazon customer reviews (75.26 million)	Yes In depth explainability for transparency and managerial application	✓ Top2Vec	✓ Modified Bi-LSTM (designed on BERT based transform er output)	✓ Top2Vec HDBSCAN UMAP (LIME) Semantic Indexing & Thematic Analysis

Note: Table 1 demonstrates the potential for advanced natural language processing techniques to extract valuable insights from unstructured textual data in the business risk prediction domain. However, there remain some limitations around model interpretability that present opportunities for further research, especially within the fake review stream.

While complex transformer architectures like BERT achieve state-of-the-art performance on language tasks, their 'black-box' nature hinders transparency and openness. Techniques like attention mapping provide some insight into model behaviour but fall short of full explainability capacity. As businesses increasingly rely on NLP models to automate decisions, trust and transparency become crucial. Hybrid approaches that combine neural networks with linear models or rule-based systems could balance accuracy and explainability. Additionally, domain-specific taxonomies and ontologies designed by experts, as used in some of the listed emerging studies, have been shown to enhance model semantics and explainability. Therefore, while emerging interest is in applying NLP for business text analytics, opportunities remain to improve model explainability, reduce bias, and develop hybrid methods combining neural networks with structured knowledge. Advances in these areas will be vital to increasing the adoption of text analytics for critical business decision-making in the future. This article breaks new ground by analysing the categorisation and intent of fake reviews, shedding light on their impact on corporate risk assessment and offering valuable insights for businesses. For the reasons stated above, incorporating the believability and intentions behind online fake reviews into corporate risk assessment becomes crucial, moving beyond simply understanding their impact on BFP.

3.0 Methodology

In developing the proposed explainable Fake Review Index (R_{FRI}) framework, we used a combination of three complementary experimental setups (Study 1-3) that helped us identify a series of important structured and unstructured data-driven metrics in developing a unique deep learning bi-LSTM architecture. Sequentially, we designed **Study 1** to fine-tune BERT (FTFR-BERT) and identify normative characteristics of Fake Reviews (FK) compared to Original Reviews (OR) as identified by our trained FTFR-BERT classification model. **Study 2** was designed to perform advanced semantic-level topic model indexing using Top2Vec. Normative and semantic information gathered from Study 1 and Study 2 was used to train and define the bi-LSTM model architecture (**Study 3**). This process has helped us develop a novel deep learning binary and multiclass classifier that provides unprecedented future opportunities for reiterative improvement in identifying, classifying and indexing e-commerce fake reviews as part of the organisational risk index (see Fig. 1).

The process started by fine-tuning BERT, and then using the trained BERT model (FTFR-BERT) to identify fake reviews from 75.26 million original customer review data. Textual data from the classified (fake vs original) customer reviews were then processed with Top2vec to develop semantic indices with interpretable dimensionality. The resulting embeddings retained interpretable semantic meanings with less complex dimensionality. Semantic meanings within identified topics were further interpreted by a team of researchers using a three-stage interpretative coding method (Vanover, Mihas, & Saldana, 2021); identified topics were then transformed into a semantic modified BoW format. Thereafter, a series of statical validation methods were applied to identify important and significant structured and unstructured data features that became the building blocks of our bi-LSTM model. By controlling the model parameters and processing steps, we improved interpretability and trust in the final model. We developed four different versions of the bi-LSTM model and identified the best-suited architecture using a combination of validation techniques. This increased confidence in understanding the model's inner workings necessary for developing fair, responsible, and interpretable AI systems. The final validated LSTM model offers LIME interpretability and softmax output that were further used to develop the Fake Review risk Index alongside calculating a confidence value devised from explainable LIME results.

This study utilises the strengths of complex contextual embeddings by leveraging their rich information while simplifying them in a controlled manner. This approach aims to achieve a balance between model performance and interpretability, ultimately contributing to the development of an explainable Fake Review Index (R_{FRI}) (Section 4.4). In order to maintain methodological rigour, we applied three important statistical measures in validating our data and process: (i) nearest neighbour propensity score matching (Study 1) – used to identify highly significant structured meta-data, (ii) SMOTE (Study 3) – to balance any discrepancy arising from dataset imbalance, (iii) Mann-Whitney U test (Study 3) – model performance variation measure. Although our study exploits the advantages of transformer-based model architecture and representation power offered by BERT, we add further value to the fine-tuned BERT outcome in developing our bi-LSTM, which is much more scalable and explainable to the managers and decision-makers by design. By creating the bi-LSTM, we provide more control of neural network architecture to the future user base; for example, future development of the model can consider additional covariance, such as review time, review readability score, etc., towards designing a more robust but scalable model.

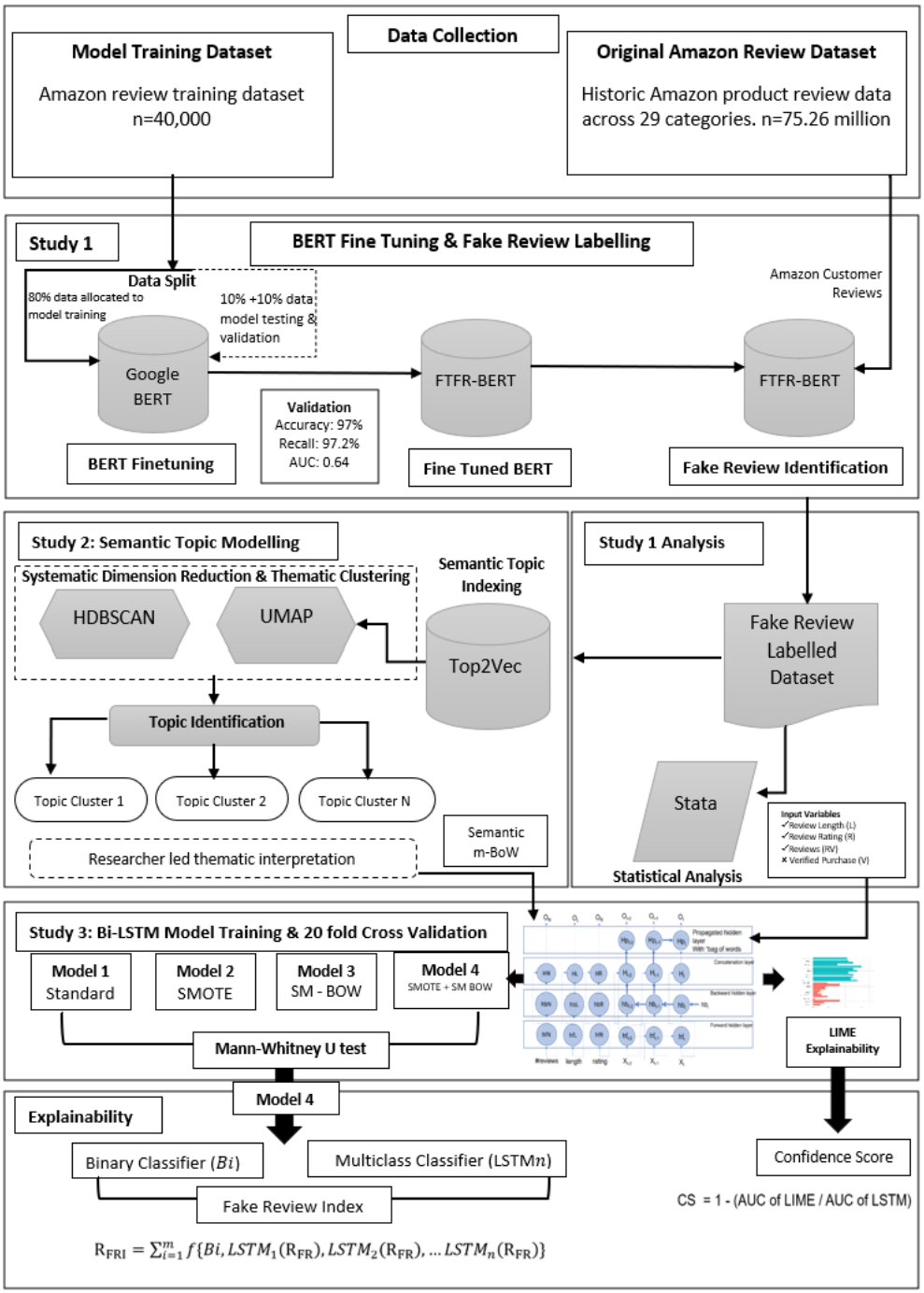


Fig. 1. Schematic representation of experiment Design

3.1 Study design and experimental setup

Study 1 *– BERT fine-tuning & Fake review detection & labelling:* during the first stage of analysis, we aimed to identify the linguistic nature and characteristics of Amazon fake reviews using customer reviews from all 29 selected categories of Amazon products. Using the Transfer Learning concept we developed a fine-tuned version of Google's BERT (Bidirectional Encoder Representations from Transformers) language engine (Devlin et al., 2018).

Fine-tuning of the BERT was conducted using a labelled training dataset. A common practice for BERT fine-tuning involves an 80/10/10 split of the data for training, validation, and testing, respectively.

BERT weights were unfreezed and updated with a small learning rate of $2e-5$. This value is pretty much standard for a slow and steady good fine-tuning. 3-4 epochs were trained and monitored for validation performance. Although this can vary, previous studies suggested it as an efficient number, and our performance validation monitoring complied with it. During the cross-validation experiment, our fine-tuned BERT model (FTFK-BERT) produced 97% fake review detection accuracy when compared against pre-labelled values (1 = Original ('OR') or 0 = Fake ('FK')). In addition, Precision = 97.3, Recall = 97.2, F1=97.2, and AUC = 0.64, meaning our model achieved linguistic detection capability equivalent to RoBERTa (Briskilal, & Subalalitha, 2022).

The fine-tuned version of BERT (FTFK-BERT) was then used on the 75.26 million Amazon customer review dataset, representing 29 different product categories, to identify and label fake and original reviews (Kaliyar et al., 2021). The dataset comprises a total of 233.1 million Amazon product reviews collected over a period of 12 years and 5 months. To manage the scalability of the project and considering available computational power, we selected a K-core subset of this dataset with 75.26 million product reviews (Ni, Li, & McAuley, 2019). Twenty-nine varied Amazon product categories were identified within the dataset, including Books, Automotive, Arts and Crafts, Amazon Fashion, Prime Pantry, Electronics etc.

Due to the volume and veracity of our dataset, deep learning experimentation and advanced semantic topic modelling tasks were performed using a High Performance Computer (HPC) hosted at the Hartree Centre, University of Oxford. JADE, or the Joint Academic Data Science Endeavour, is a tier 2 high-performance computing facility with 63 MAXQ Deep Learning Systems comprising 8 Tesla V100 GPUs and 4TB of SSD, designed to handle complex machine learning tasks. In total we used 1 Tesla for around 473 hours to perform our experiments.

The fake review identification and labelling process took approximately seven days (non-stop) to complete using one Tesla V100 GPU. Following the training dataset analogy, customer reviews were analysed and labelled either as '1=OR' or '0=FK'. Once the entire dataset was labelled, associated metadata (rating, verified etc.) and numeric values were used for statistical validation (propensity score nearest neighbour matching) (**Appendix 1**), while the textual data was used for large semantic topic modelling indexing (Study 2). Applying fine-tuned BERT offered this study multidimensional advantages, i.e., (i) vocabulary modification, (ii) context-dependent vector representation and (iii) self-attention towards capturing feature representations. Using BERT helped us investigate semantic sentence-level linguistic characteristics through modified hyperparameters (Huang et al., 2022). Output from the machine learning experiment was further statistically analysed and validated using Stata version 17.

Study 2 – *Semantic Topic Modelling:* this study was designed to perform advanced semantic-level topic model indexing using Top2Vec. This part of the study was inspired by an emerging stream of operational research that has shown the added value of unstructured textual data in improving operational performance prediction (Stevenson, Mues, & Bravo, 2021; Borchert et al., 2023). These studies claim that there is a real lack of investigation into how unorganised textual data can be systematically modelled and used aside structural data to improve business performance prediction.

To date, fake review research is highly concentrated on detection parameters that are derived from Large Language Models (LLMs) (Salminen et al., 2022). No specific work could be identified that benefits from additional topic model indexing as recommended by the studies above. Borchert et al. (2023) argue that unstructured textual data provides a complementary perspective and, when processed with

other structured information, can enhance business performance prediction. The authors tuned TF-IDF textual embeddings to develop a Latent Semantic Index (LSI) that helped them to extract better semantic textual features that combined with structured data, were used to improve business failure prediction compared to using only structured data. The textual features identified also provided additional insights. Inspired by these studies, we developed and applied our own Semantic Indexing technique that not only offers a new dimension to fake review research but also provides technical advancements to the study highlighted above.

Top2Vec is an unsupervised learning algorithm that jointly embeds words, documents, and topics into a common vector space. It uses document co-occurrence patterns to learn semantic relationships between words/documents. We used Top2Vec, which does not require any explicit preprocessing like bag-of-words or tokenisation. The algorithm is capable of directly ingesting raw textual data and output meaningful vector representations capturing semantics and similarity (Egger & Yu, 2022). The document vectors can then be used as features. Labelled fake review textual data identified in Study 1 was fed into Top2Vec. It analysed co-occurrence patterns to capture semantic relationships between words and documents, similar to Borcher et al.'s (2023) LSI, but with greater efficacy. Output embedding vectors from Top2Vec for words and topics produced encoded semantic similarities based on co-occurrences. Due to the volume of data, the output from Top2Vec was too large and needed dimensionality reduction. Dimensions of document embedding vectors were reduced using UMAP, and then topic clusters were created and visualised using HDBSCAN. This approach enriched document representations by capturing semantic relationships analogous to LSI's document vectors. The output document embeddings and topics from Top2Vec summarise semantic information more effectively than LSI. The semantically enriched, reduced dimension document vectors from Top2Vec was further analysed by two researchers, and it was used as input to the bi-LSTM model along with other features model in the form of semantic modified BoW (SM-BoW or M-BoW). The document vectors injected semantic knowledge that complements the bi-LSTM model's predictive performance.

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was used to identify density based topic clustering. Due to the large nature of our database, topics within high dimensional semantic embedding space often became widely dispersed, making it difficult to identify prominent clusters. To eliminate this problem, Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm was used to reduce dimension space where a sufficient amount of densely topic clusters could be identified through HDBSCAN. After a number of trial and error, we identified that a cluster size of 25 words were best representative of a topic theme, and a total of 100 themes per product category presented a good linguistic insight into fake review content.

Topic consistency values were also calculated to understand the coherent or dispersed nature of fake reviews themes within each product category. Finally, macro-level topics were analysed and decoded by human researchers using axial coding principles. Following Angelov (2020), we use the following equation to describe how topic-level information was generated.

Topic information gain: Topic information gain calculation was used to measure the informative nature of individual cluster outputs. Total information (I) gain from documents (D) with words (W) can be represented as

$$I(D,W) = \sum_{d \in D} \sum_{w \in W} P(d,w) \log \left(\frac{P(d,w)}{P(d)P(w)} \right)$$

Here, the contribution of each document (d) with word (w) can be represented as probability-weighted information (PWI).

$$PW\ I(d,w) = P(d,w)\log\left(\frac{P(d,w)}{P(d)P(w)}\right)$$

To reduce unnecessary dimensions and to increase interpretation efficiency, *n* number of words were selected within a topic (t). Hence, probability-weighted information across all the topics can be represented as:

$$PW\ I(T) = \sum_{t\in T}\sum_{d\in D_t}\sum_{w\in W_t} P(d,w)\log\left(\frac{P(d,w)}{P(d)P(w)}\right)$$

$$= \sum_{t\in T}\sum_{d\in D_t}\sum_{w\in W_t} P(d|w)\,P'(\mathrm{w})\log\left(\frac{P(d,w)}{P(d)P(w)}\right)$$

Here, P(w) represents the marginal probability of an identified word (w) across all the documents (D), leading to the calculation of 'pointwise mutual information'. $P'(\mathrm{w})$ represents the probability of selecting a word (w) within a topic, leading to the calculation of 'expected mutual information' (Li et al., 2022). At the beginning of the process $P'(\mathrm{w}) = 1$, modifying the equation as:

$$PW\ I(T) = \sum_{t\in T}\sum_{d\in D_t}\sum_{w\in W_t} P(d,w)\log\left(\frac{P(d,w)}{P(d)P(w)}\right)$$

During the later stage of the process, we assumed that each document represented multiple topics, replacing $P'(\mathrm{w})$ with $P'(\mathrm{t})$:

$$PW\ I(T) = \sum_{d\in D}\sum_{t\in T}\sum_{w\in W_t} P(d|w)\,P'(\mathrm{t})\log\left(\frac{P(d,w)}{P(d)P(w)}\right)$$

Therefore, Top2Vec helped us generate semantically meaningful document vectors (for each Amazon product category) that were used as an input to LSTM models, providing advantages over LSI for dimensionality reduction and semantic enrichment of textual data.

Study 3 – *Bi-LSTM Neural Network Model Development & Validation:* results from Study 1 not only provided us with an extensive database of categorically labelled fake and original reviews, but statistically it also helped us identify other important variables to be considered towards designing a bi-LSTM neural network model. Based on the statistical analysis results from Study 1 (see **Appendix 1**), input variables for our bi-LSTM model consisted of Review Content (N), Review Length (L), and associated Review Rating (R). Although 'Verified Purchase' (V) co-variance appeared to be an important factor associated with fake Amazon reviews, importance of this factor could not be statistically validated, hence it was disregarded. Semantic topic model information related to fake reviews in each category (Study 2) was also introduced into the bi-LSTM architecture as an added layer in the form of semantic document vectors.

For the training and validation, we used 20-fold cross-validation method (Wong & Yeh, 2019). Primarily, we used 60 million Amazon review data for training purposes. The cross-validation process is then repeated with 10 million (500K x 20 folds) review data, with each of the 20 (19+1) subsamples of 500K review data used exactly once as the validation data and the remaining 19 subsets as training data. This process allows all 10 million reviews to be used for both training and validation across the 20 iterations. Using each fold only once for validation avoids overlap and bias. After cross-validation,

a separate 500K review set was used for final testing to avoid further bias. This method is beneficial in maximising both the training and testing data, allowing a robust assessment of the model's performance.

Input data was uniquely assigned and fed into three 'stacked' layers of LSTM cells. 'Unrolled' loops were used to keep persisting information hidden within recurring networks, while outputs from these networks were further fed into other time-distributed layers as recurring inputs for model training. Finally, a softmax function was applied to transform raw output into vectors of probabilities (Zhu et al., 2020). In total, we created four different versions of the bi-LSTM model **(Model 1: Standard; Model 2: SMOTE Data; Model 3: SM-BoW; Model 4: SMOTE + SM-BoW).** Following Borchert et al. (2023), we used F1-score, AUC and MAP (mean average precision) as accuracy metrics to identify the best performing model. This process has helped us develop a novel deep learning binary and multiclass classifier that provides unprecedented future opportunities for reiterative improvement in identifying, classifying and indexing e-commerce fake reviews as part of the organisational risk index.

Our study design exploits the advantages of transformer-based model architecture and representation power offered by BERT (in Study 1), and adds further values of the fine-tuned BERT outcome in developing a bi-LSTM architecture, that is much more scalable and explainable to the managers and decision-makers by design. By creating the bi-LSTM, we provide more control of neural network architecture to the future user base, for example, future development of the model can consider additional covariance such as review time, review readability score etc. towards designing a more robust but scalable model. We improve interpretability and trust in the final model by controlling the model parameters and processing steps. This increased confidence in understanding the inner workings of the model is necessary for developing fair, responsible, and interpretable AI systems. Overall, we leverage the power of complex contextual embeddings and simplify them in a controlled manner to create performant and interpretable models for developing a Fake Review Index (R_{FRI}).

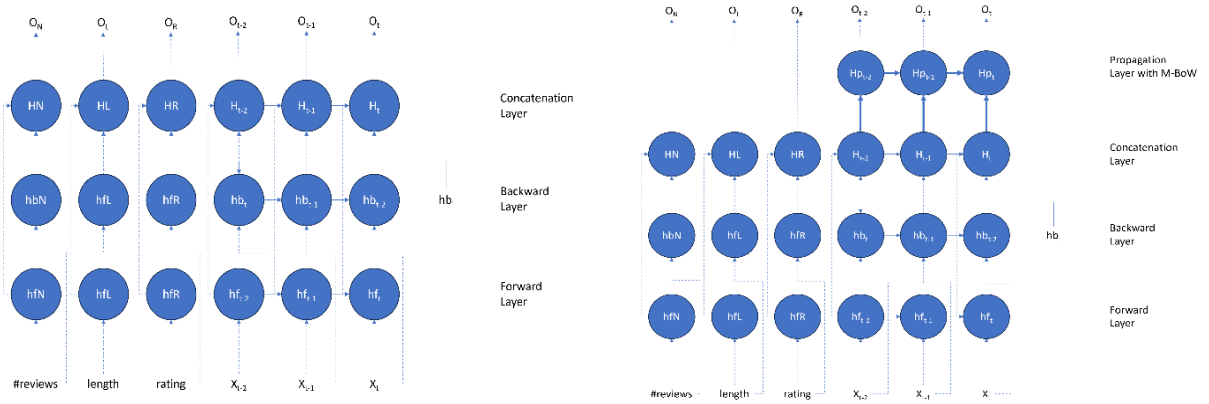
In addition, the LSTM model offers LIME interpretability and softmax output that are further used to develop the Fake Review risk Index and further validate the model with a calculated confidence value devised from explainable LIME results.

3.2 Bi-Directional LSTM model architecture

Our proposed Fake Review Index (R_{FRI}) was developed on a bi-directional LSTM architecture output. The primary LSTM model ran on dataset provided input sequence, while the secondary LSTM model followed a reverse-order input sequence. Such an architecture offers benefit by using past and present (both) temporal directions at the same time, and outperforms single LSTM outputs (Jang et al., 2019).

We further believe that using a Bi-LSTM allowed us to create two hidden states within a given time frame, preserving information from forward and backward state cells. This approach helped the model to learn and encode information and context with better accuracy, given the complex nature of textual data. One of the key operational challenges of our Bi-LSTM architecture was to merge the two layers to determine the final output, while the models were functioning independently of each other. To overcome this problem, we created a Concentration layer and used a softmax activation function to merge and prepare the values to be propagated to the output layer.

The novelty of our Bi-LSTM architecture design was further boosted by three additional neurons (Fig. 2a) in each layer: Reviews (N), Length (L), and Rating (R). These neurons were trained and weighted like other neurons within their respective layers, and all the information was congregated within the Concentration layer in separation. Finally, results from these selected neurons were carried forward to the output layers, for cross-validation, without passing through the propagation layer (see Fig. 2b).



(a) **(b)**

Fig. 2. (a) dash line: independent connection to O_t, thick blueline: new connections within Bi-LSTM. *t* represents position of a words within a sentence. (b) Bi-LSTM with Semantic Modified Bag of Words (M-BoW). Dash lines: independent connection to O_t, thick blue line: new connections in Bi-LSTM. *t* represents position of a words within a sentence.

A single-cell LSTM architecture was the key building block of our neural network (**Appendix 2**).

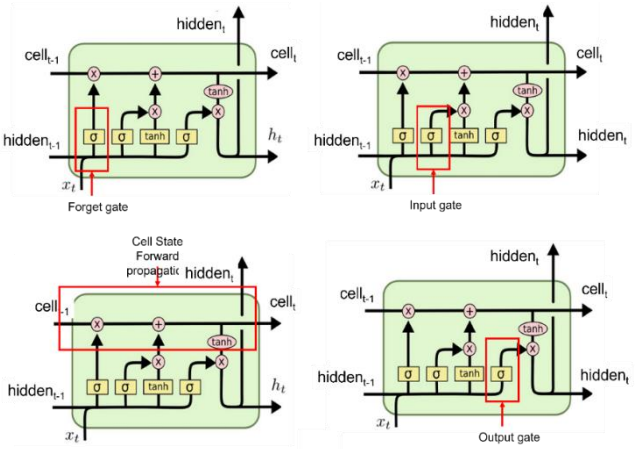


Fig. 3. Four key components of an LSTM cell architecture. Red squared areas: (a) Forget gate sigmoid function, (b) Input gate sigmoid function, (c) Cell state forward propagation, (d) Output gate sigmoid function.

Semantic Modified BoW: to further experiment and compare accuracy and validity we created another two different versions (Model 3 & 4) of our bi-LSTM model by adding a propagated hidden layer with a semantic modified bag of words (SM-BoW) identified from Study 2. Using vector space we designed robust n-gram bag of words (BoW) by observing and analysing complex semantic topic models resulting from Study 2. We believe that adding the SM-BoW layer could act as an important layer for learning and identification of fake reviews with higher accuracy.

Together this neural network architecture defines output for: (i) identifying the probability of a review to fall into 'Fake' or 'Original' category (binary classification), (ii) Identify review rating associated with a fake review within specific product categories (multiclass classification).

The two different classification modes add an extra dimension to the fake review risk assessment beyond just the binary classification (fake or not). Review rating scores help quantify the degree of distress across multiple ordered categories, rather than just a binary outcome. This provides a more nuanced risk spectrum whether fake reviews inflate or deflate the future of certain products and organisations (considering ranking, visibility, credibility, compliance etc.). Second, predicting the auxiliary review score gives a useful signal for interpreting the black-box Bi-LSTM model. Since

review scores correlate with the risk label, being able to accurately predict them helps validate that the Bi-LSTM is capturing meaningful risk patterns that independently feed into the indexing system. Applying LIME across review rating behaviour also helped to explain the correlation between two different spectrums of fake review writing and rating behaviour.

Jointly modelling the review score with topic interpretation strengthened the risk indexing mechanism by adding an intermediate level of quantitative assessment.

3.3 Bi-LSTM output (binary and multiclass classification):

A softmax activation function was used to produce the following parameters for the FRI (R_{FRI}).

Binary Classifier: for binary fake review detection, the softmax takes the logit score z for a given review and converts it into a probability p that the review is fake. Specifically:

p(fake) = exp(z) / (exp(z) + exp(0))

Since only two classes (fake and real) exist, the non-fake probability would be 1 - p(fake). The higher p(fake) is, the more likely the model predicts the review is fake. We set a median threshold of 0.5, where p(fake) >= 0.5 indicates the model predicts the review is fake. This binary prediction is directly derived from the softmax probability.

Multiclass Classifier: the softmax formula for multiclass classification deserves a more in-depth explanation and concrete examples to demonstrate how it assigns predictions. Let me provide some additional details: For a 5-class rating analysis case, the softmax would produce a probability distribution over the five classes. For example, if the logit scores are [2.1, 1.0, -0.5, 0.2, 3.2], the softmax probabilities could be: [0.2, 0.1, 0.02, 0.05, 0.63].

Here, the highest probability is class 5, so the model predicts that selected fake review has 5-star rating. The probabilities provide a complete distribution for all classes. Here is an example of how the softmax formula would work for 5-star rating classification: Let's say we have a review with logit outputs from the model as:

z1 = 1.5 (logit score:1-star); z2 = 3.0 (logit score: 2-stars); z3 = 2.8 (logit score: 3-stars); z4 = -0.2 (logit score: 4-stars); z5 = 1.2 (logit score: 5-stars)

Applying the softmax formula as:

p(1-star) = exp(1.5) / (exp(1.5) + exp(3) + exp(2.8) + exp(-0.2) + exp(1.2))

p(2-stars) = exp(3) / (exp(1.5) + exp(3) + exp(2.8) + exp(-0.2) + exp(1.2))

same for 3, 4 and 5 stars.

This normalises the logits into a probability distribution: [0.1, 0.5, 0.3, 0.02, 0.08]. Since 2-stars has the highest probability of 0.5, the model would predict this review is a 2-star rating.

4.0 Results

4.1 RQ1: Exploring fake review characteristics beyond textual information

From the 75.25 million Amazon review dataset spread across all 29 product categories, the fine-tuned BERT model identified 8.27 million fake reviews. Following the Pareto principle, Fig. 4 depicts a distributed view of identified fake reviews across all categories. The first eight categories, ranging from 'Books' to 'Toys and Games,' accounted for roughly 80% of the detected fake reviews. The remaining

20% was detected within the twenty-one remaining categories. Books had the most fake reviews (n=2.27 million, or 27.49%), followed by Clothing, Shoes, & Jewellery (n=1.52 million, or 18.41%), and Electronics (n=772.28 thousand) (9.33%). Amazon Fashion n=509 (0.01%), Magazine Subscription n=429 (0.01%), and Appliances n=61 (0.001%) had the fewest identified fake reviews.

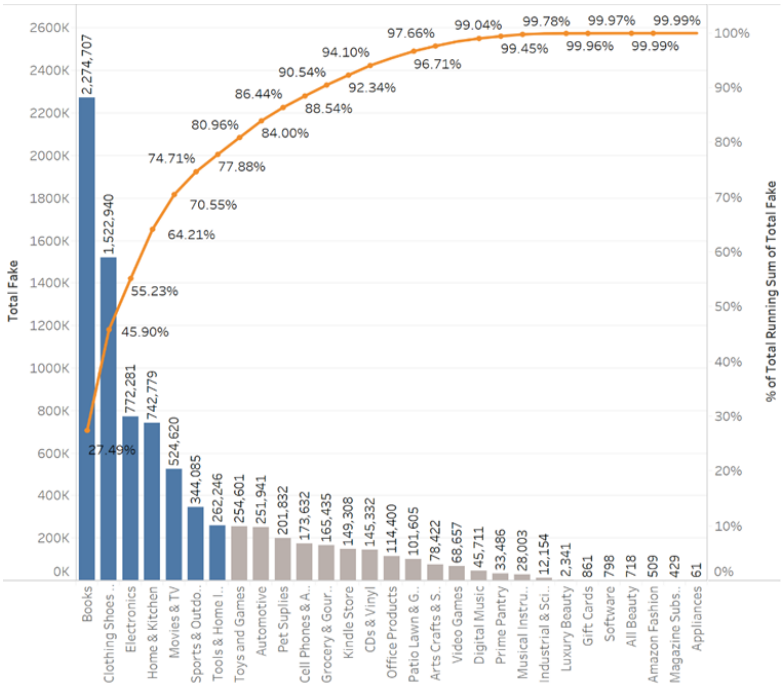


Fig. 4. Distribution of fake reviews Pareto chart

Following their initial identification, reviews were evaluated further using three key variables: review rating (R), length of review (L), and purchase verification (V). Table 2 provides a comparative overview of these review characteristics for genuine and shill reviews.

Table 2:
Real vs Fake Review Characteristics

	Number of Reviews (N)		Mean ($\bar{x}$)		Minimum		Maximum		Standard Deviation (σ)		Variance (σ^2)	
	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake	Real	Fake
Review Rating	66.98*	8.27*	4.32	4.68	1		5		1.05	0.76	1.10	0.58
Length of review	66.98*	8.27*	69.97	9.31	1		1039	692	81.99	24.18	6722.36	584.67
Verified purchase	52.77*	7.62*	—	—	—		—		—	—	—	—

Non-Verified Purchase	14.21*	0.65*	—	—	—	—	—	—	—	—

*values in a million (000,000)

Review Rating (Fig 5 & 6)*:* the majority of the fake reviews 76% (n=6.29 million) are 5-star. In comparison, 15.1% (n=1.25 million) of the ratings were 4-stars, while 4.8% (n=0.39 million) were 3-stars. In comparison to fake reviews, only 63% (n=42.18 million) of original reviews received 5-star ratings, 19.3% 4-star ratings (n=12.9 million), and 8.8% received 3-star ratings (n=5.9 million). Fake reviews have a greater right skewness towards 5-star rating with a mean value $\bar{x}$ = 4.68. In contrast, original reviews are more gravitate towards 4-star rating with a mean value $\bar{x}$ = 4.32. Genuine ratings are also more evenly distributed than fake evaluations, with standard deviations and variances σ=1.05 and σ^2=1.10. Such findings indicate that the majority of fake reviews are written to boost the overall rating of specific product lines, strengthen algorithmic search ranking and influence critical decision-making and purchase behaviour.

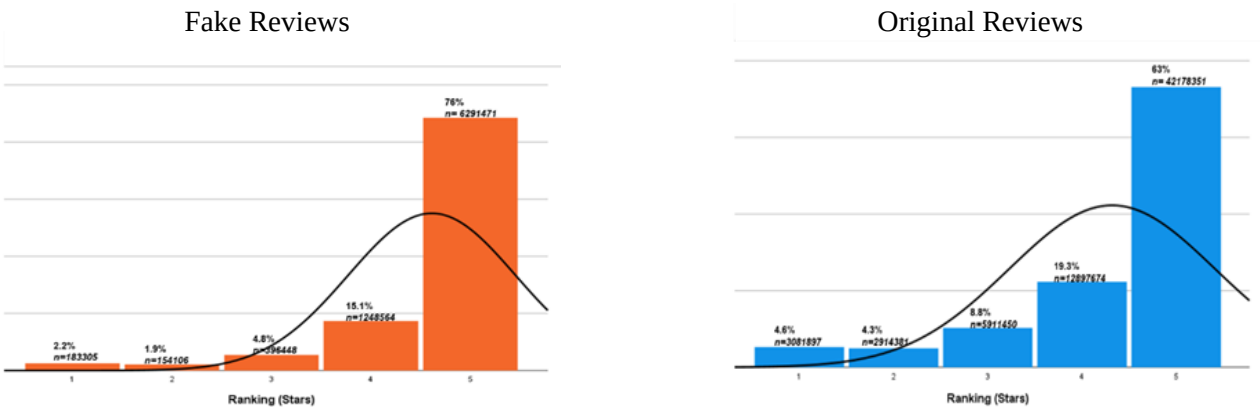


Fig. 5. Distribution of review rating

Fig. 6 depicts more granular review rating behaviour in various categories. The original review rating behaviour appears more realistic and distributed based on the heatmaps, except for the Kindle Store (66%) and Software (67%) categories, where the majority of fake reviews (70%) are 5-star.

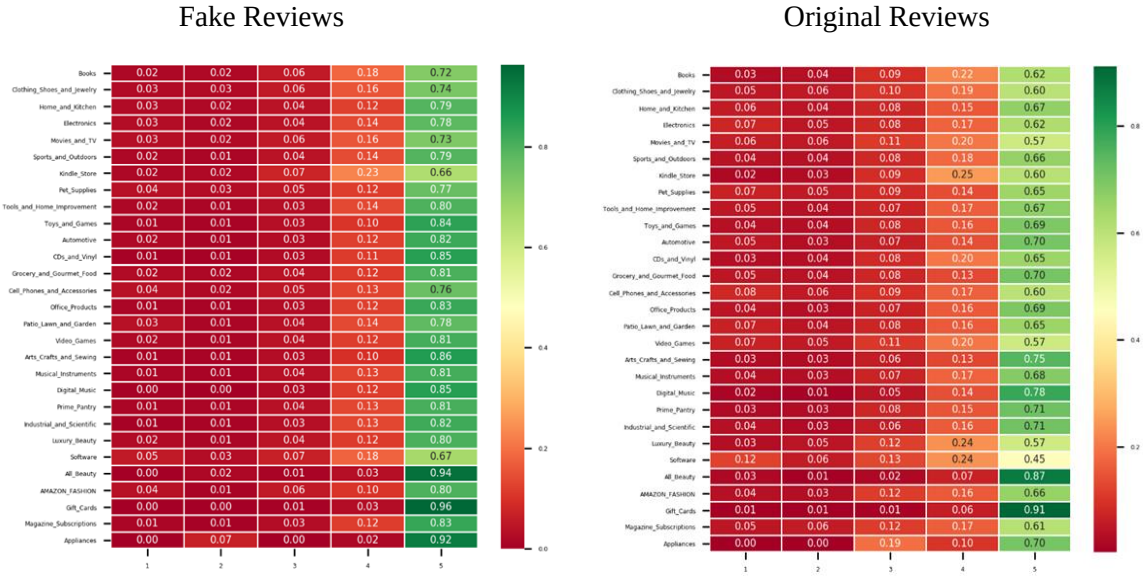


Fig. 6. Review rating heatmap

Review Length (Fig. 7)*:* aside from review ranking, our findings show that fake reviews are typically short, with the average number of words per fake review being less than ten words ($\bar{x}$ = 9.31) compared to approximately 70 words ($\bar{x}$ = 69.97 words) for real reviews. Fake reviews are written in lengths of 100 or more words on infrequent occasions. The average length of a fake review was 692 words, compared to 1039 for original reviews. Fig. 7 shows that original reviews in specific categories, such as Appliances, may have a high (47%) number of comprehensive reviews (250 words or more). Original review standard deviation and variance (σ = 81.99, σ^2 = 6722.36) are significantly higher than fake review standard deviation and variance (σ = 24.18, σ^2 = 584.67). These comparative characteristics show that genuine reviews are more varied than fake reviews.

Fig. 7. Length of review (words) heatmap

Nearest Neighbour Propensity Score Matching - due to the large volume of the Amazon review dataset, along with dispersed distribution of reviews across different categories, it was essential to avoid any endogeneity issue arising from sample variation. Possible bias was avoided through a propensity score matching test comparing fake reviews and original reviews within and across each product category. Using propensity scores (Rosenbaum & Rubin, 1983), we reduced the likelihood of statistical bias (see **Appendix 1**). Random samples were gathered from fake and real reviews in each category, and then matching experiments were carried out involving 1, 2 and 3 nearest neighbours, with an aim to check result sensitivity. Difference coefficients (real vs fake) were found to be highly significant (1%, 5%, 10% significance level) across all the tested variables, except 'Verified Purchase'. In these day and age, fake reviewers buy products from Amazon to further authenticate the validity of their reviews in the form of 'Verified Purchase'. This practice appears to be so significant that this factor cannot be used as a statistically distinguishing factor. Therefore we decided we decided to drop this variable from our bi-LSTM model.

4.2 RQ 2: Semantic topic modelling - linguistic detail and thematic characteristics

In order to generate more linguistic nuance and thematic insights into fake reviews, fake reviews from all 29 product categories were processed through Topic2Vec, creating 2900 thematic topics (100 topics per category) with up to 25 keywords per topic (a total of up to 72,500 keywords). This exhaustive list of modelled topics and words were studied and deciphered by two researchers from the group. To create broad thematic descriptions, we used axial coding principles combining researcher interpretation, inter-topic distance map, and topic coherence (TC) values. These broad keyword-based themes are referred to as "clusters". Topic Coherence values across different product categories ranged from 0.302 (Magazine) to 0.775 (All Beauty), indicating strong coherence and similarity in the majority of categories. Due to the volume of results and level of linguistic detail, **Appendix 3** provides a further insight into one of the largest reviewed categories - books.

Books: The Books category (Fig. 8) represents interesting and distinct elements of fake reviews with a selective purpose of influencing prospective buyers' decisions. Four distinct clusters and associated thematic descriptions were identified as part of our analysis, with a moderate Topic Coherence value (TC = 0.574), indicating a dispersed level of comments and remarks that characteristically exist within fake Amazon book reviews. Fig. 8. highlights how keywords and thematic topics in this category are largely dispersed, although their underlying themes are intertwined in many aspects.

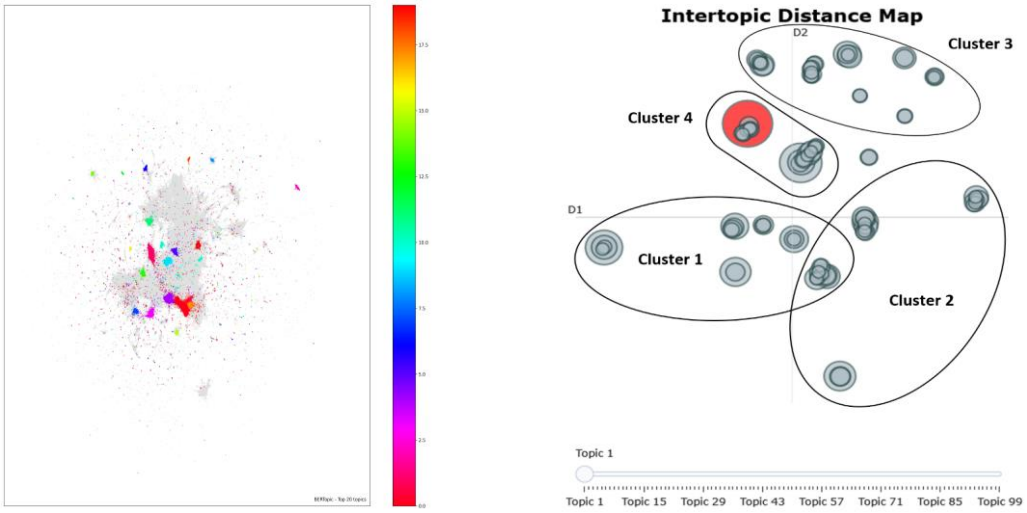


Fig. 8. Topic cluster mapping (Category: Books).

Note: the image on the left shows high dimensional distribution of topic clusters in a three dimensional vector space. The image on the right shows reduced dimensional topic distribution. Clusters were developed through researcher-led topic interpretation.

4.2.1 Believability, Misguiding Intensions and Trust:

Cluster 1: the topic of fake reviews within Cluster 1 was developed on the characteristics of a book and its storyline. A significant proportion of reviewers intended to promote 'suspense', 'thriller' and the 'intriguing' nature of selected storylines to misguide review readers and prospective buyers by creating a sense of anticipation. The documented features of these reviews, with their heavy emphasis on evoking anticipation for specific genre-based storylines, reveal an ulterior motive beyond genuine experience sharing. They aim to generate hype and intrigue, ultimately manipulating and misleading consumers by creating the illusion of authenticity. This calculated approach makes them more believable than reviews based on genuine experiences.

These reviewing activities were highly concentrated on a specific genre of books based on dystopia, ‘paranormal romance’ and crime series. Such a narrow concentration deviates from broader topic distribution, selectively highlighting intentions of hyping specific emotional aspects rather than well-rounded assessments. The overall goal here is to influence purchase intentions. These rhetorical patterns exhibit specific characteristics that have not been highlighted in prior false narrative research (Hajek, Hikkerova, & Sahut, 2023), with a market positioning intention where subjective emotional appeals are used over factual criticism to gain strategic advantage.

Cluster 2: fake reviews within this cluster demonstrated unique characteristics, distinct from the hype-generating ones in Cluster 1. This cluster focused on science fiction and erotic novels, exhibiting a more discursive tone and balanced, dichotomous nature. Interestingly, these reviews displayed a clear split between positive and negative sentiments, suggesting a polarisation in positioning and misguiding goal of manipulating a specific group of reviewers. This bifurcated sentiment aligns with "agree-disagree" formulations (Hajek, Hikkerova, & Sahut, 2023), potentially aiming to manufacture mixed reactions, manipulate, and misguide audience perception. The disconnect between positive tone and expressed dissatisfaction ("pleasant" and "enjoyable" alongside "abrupt" and "surprising") hints at external influence rather than genuine experience. For example, reviewers might endorse books as "enjoyable" but explicitly lament the "abrupt" or "unsatisfying" ending. This unnatural polarity resembles patterns seen in Markov chain opinion modelling (Zhang et al., 2023). Since genuine human reviews are unlikely to follow such precise mathematical patterns, this trend suggests the work of algorithmically generated reviews lacking genuine experience and intention to manufacture mixed reactions. Our research unveils the alarming sophistication and diversity of fake review tactics, posing a significant challenge to online trust and consumer decision-making. Traditional methods relying solely on sentiment analysis fall short in detecting these increasingly nuanced manipulations. Furthermore, believability in fake reviews often deviates from actual accuracy. This is evident in the phenomenon of "fake review navigation," where consumers overestimate their ability to identify fake content despite lacking effective strategies. This highlights a critical gap between perceived and actual media literacy, posing a significant challenge for consumers and platforms.

Cluster 3: this cluster was focused on fake reviews for cookbooks and recipes, particularly those related to healthy eating and vegan cuisine. These reviews aimed to combat negative perceptions associated with strict diets by frequently using positive phrases like "tasty," "tested," and "wonderful." Notably, the emphasis on personal health improvement significantly exceeded what was found in genuine reviews, highlighting a specific feature designed to appeal to consumers seeking weight loss goals.

The explicit link between personal goals and self-disclosure differed from the natural reflection observed in genuine reviews. Cluster 3 reviews instead constructed a narrative of setting and achieving goals like weight loss on a public platform like Amazon, mirroring the objectives often marketed by the very cookbooks they promoted. This suggests a strategic alignment with commercial objectives rather than genuine personal experiences, deviating from the gradual self-discovery principles advocated by nutrition and dietetic research. This misrepresents the product’s effectiveness and misguides consumers by creating the illusion of widespread positive sentiment. Cluster 3 reviews exclusively focused on promoting veganism and vegan food recipes, employing similar linguistic patterns. This opportunistic framing suggests an attempt to capitalise on current consumer trends and fads by creating the illusion of widespread positive sentiment towards veganism and specific products.

These findings highlight the sophistication and potential harm of fake reviews to exploit specific consumer interests. By masquerading as genuine testimonials and aligning with popular trends, these

reviews can mislead consumers about the effectiveness of products and manipulate their buying decisions. This necessitates the development of more robust detection methods that can go beyond simple sentiment analysis and identify subtle linguistic cues like those observed in Cluster 3.

Cluster 4: revealed distinct fake reviews targeting historical fiction and romance novels. These reviews displayed distinct negativity, aiming to deter readers and potential buyers. The excessive criticism ("predictable," "lack of dimensions," "flat," "boring") and focus on negativity ("dark," "darkness," "darkest") deviate from impartial characteristics expected in genuine reviews. Interestingly, many targeted World War II stories, focusing solely on the "violent and evil side" while dismissing their potential educational value. War stories are generally educational, and people read them to gain knowledge of history, politics, leadership, society and economy. By promoting the violent and evil side of war, fake reviews appear to deter consumers from the educational aspects of the literature.
This suggests an attempt to downplay the historical and societal insights often explored in such narratives. These fake reviews risk misleading readers and unfairly discrediting entire genres by promoting such a one-sided view. As manipulative tactics in fake reviews increasingly rely on specific terms and themes, research on robust detection methods (beyond sentiment analysis) is crucial to combat their detrimental impact on consumers and online platforms. This includes uncovering the underlying intentions behind reviews, as Román et al. (2023) highlighted in their work on the role of credibility and the reviewer's emotions in shaping trust.

These findings open doors to further insights beyond previous fake review research findings, enabling us to delve deeper into review believability, identify manipulative tactics, and personalised stories of fake user experiences based on review trustworthiness. Our cluster-based interpretation shows that despite diverse forms, fake reviews share a common goal of deception: exploiting users' cognitive biases and emotional vulnerabilities to mislead them and build unwarranted trust/mistrust (Wang, Fong & Law, 2022). Deception tactics like inauthenticity, misleading information, and emotional appeals can be particularly effective in this context as they exploit consumer reliance on central and peripheral cues (Mousavizadeh et al., 2022).

4.3 RQ3: Bi-LSTM model validation and explainable FRI development

Following the results of Study 1 & 2, we developed the bi-LSTM architecture with the aim of developing a Fake Review Index (R_{FRI}). We trained, validated and tested four different models involving two primary factors (i) fake review identification (binary classification) (ii) review rating classification (multiclass classification).

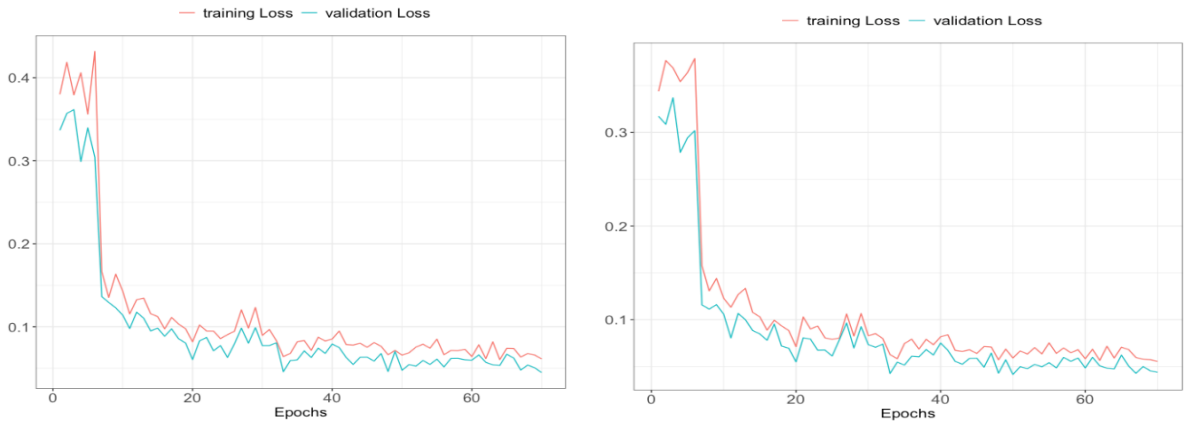
4.3.1 Model performance assessment: binary and multiclass classification

The Bi-LSTM model discussed in Section 3.3 was used to classify the fake and original reviews. Considering the imbalanced nature of the original data distribution across various categories, we further applied SMOTE method to transform the dataset and reduce minority representation. Table 3 presents a comparison of F1, AUC, and MAP for all four models. Observations from our study show that applying SMOTE method generally improves model performance and accuracy. Adding SM-BoW layer to SMOTE dataset makes little apparent noticeable difference. This apparent result can mean that Study 2 semantic modelling adds little value to improve model performance. In order to further verify this claim we apply the Mann-Whitney U test in the next section, which helps us identify the best model for FRI development.

Table 3:
F1 score, MAP and AUC comparison across four models

		Model 1 Original	**Model 2** SMOTE	**Model 3** SM-BoW	**Model 4** SMOTE + SM-BoW
F1 score	Training	0.91	0.96	0.90	0.96
	Validation	0.90	0.95	0.89	0.95
	Testing	0.88	0.94	0.87	0.89
	baseline	0.17	0.50	0.17	0.50
MAP	Training	0.77	0.84	0.79	0.84
	Validation	0.76	0.83	0.76	0.84
	Testing	0.76	0.82	0.76	0.82
AUC	Training	0.81	0.83	0.79	0.81
	Validation	0.79	0.82	0.76	0.79
	Testing	0.78	0.81	0.76	0.79

Fig. 9 demonstrates loss curves for the training and validation phases for Model 2 & 4. Characteristically, these graphs show that training losses (red) fluctuate over a small range of epochs before stabilising; validation loss (green) also shows similar patterns with lower ranges compared to training loss. Evidence and patterns from these graphs show no noticeable difference between training and validation loss, confirming the elimination of potential bias arising from data overfitting. For further information, see **Appendix 4**.



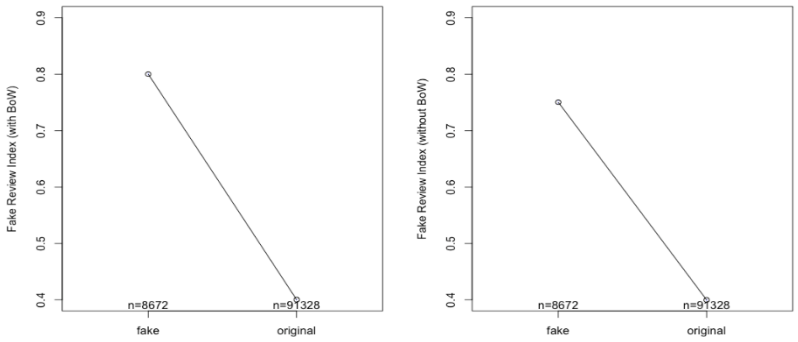
(a) Model 2 (b) Model 4

Fig. 9. Loss curves for the binary classification with (a) SMOTE and (b) SMOTE + SM-BoW

4.3.2 Validity and applicability of semantic topic model index

Results observed from the binary classification models makes it difficult to identify the best classification model due to the close proximity of F1 score, AUC and MAP between Model 2 & 4. Observations from the Mann-Whitney U test show that fake review identification capacity is significantly higher when SM-BoW (Model 4) was added as an additional hidden layer. A Mann-Whitney U test value of 27928 ($p<.00001$) shows that statistically, Model 4 outperforms Model 2 in identifying fake reviews. Similar results were identified when applied the same test to multiclass

classification. Following this statistical comparison, we concluded that adding semantic topic modelling (Study 2) adds greater accuracy to model output when classifying fake reviews.



(a) **(b)**

Fig. 10. Mann-Whitney U test demonstrating validity Fake vs Original with (a) and without (b) SM-BoW model

The multiclass classifier produces the following results. We present findings from Model 4 because of its superiority, and findings from other Models are available in **Appendix 4**. Higher accuracy towards Rating 5 clearly demonstrates that there is simply a lack of enough training data in other categories. This finding indeed confirms our findings from Study 1 (see Figure 6) that reviews are highly biased towards 5-star ratings.

Table 4:
Model 4 (SMOTE + BoW) multiclass classification performance

		Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
F1 score	Training	0.67	0.66	0.72	0.75	0.97
	Validation	0.66	0.65	0.70	0.73	0.94
	Testing	0.63	0.63	0.69	0.71	0.94
	baseline	0.20	0.20	0.20	0.20	0.20
MAP	Training	0.65	0.64	0.70	0.71	0.89
	Validation	0.63	0.62	0.70	0.70	0.89
	Testing	0.61	0.61	0.69	0.69	0.88
AUC	Training	0.67	0.66	0.72	0.75	0.87
	Validation	0.66	0.65	0.70	0.73	0.86
	Testing	0.63	0.63	0.69	0.71	0.85

The loss curves obtained from studies show similar patterns like binary classification, eliminating possibilities of bias arising from data overfitting. For further details see **Appendix 4 & 5**.

4.4 Managerial implications: explainable fake review risk index

We propose a Fake Risk Index (R_{FRI}) that systematically combines a selected number of output from the binary and multiclass classifier through weighting and aggregation to quantify overall enterprise risk. The proposed indexing system can be adapted to individual enterprise contexts by appropriating functions and weights for each category. Further benchmarks and backtesting can help validate the index at an individual enterprise level.

Explainable Fake Review Index

Overall Enterprise Risk Index,

$\mathrm{R(t)} = \sum_{i=0}^{x} w_i * f_i(t)$

$\mathrm{R(t)} = w_1 \cdot f_1(t) + w_2 \cdot f_2(t) + \cdots + w_n \cdot f_n(t)$

Where,

R(t) = overall organisational risk index over time t

w_i = combined risk factor weight based on multiple operational factors

f_i = overall risk factor function considering multiple operational factors

$f_1 .. f_n$ = function of individual risk indices based on multiple operational factors

$f_1(t) .. f_n(t)$ = functions representing changes in specific risk factor over time t

The function f for each sub-index (fake review risk, operational risk, compliance risk etc.) could be based on a model, aggregating key risk indicators. Individual risk factor based weights ($w_{1\ldots}w_n$) could be determined through expert judgment, statistical methods, or optimisation techniques. Our proposed enterprise risk index offers the provision to incorporate covariance/correlations between risk categories.

Fake Review Index:

Considering the above equation, R_{FR} denotes risk associated with fake reviews.

Therefore…

$$P(FR) = f\{\mathrm{Bi}, \mathrm{LSTM}_1(\mathrm{R_{FR}}), \mathrm{LSTM}_2(\mathrm{R_{FR}}), \ldots \mathrm{LSTM_n}(\mathrm{R_{FR}})\}$$

Where:

P(FR) = probability of overall risk posed by individual fake review ($\mathrm{R_{FR}}$)

$\mathrm{R_{FR}}$ = individual fake review associated risk

Bi = output of the binary classifier

$\mathrm{LSTM}_1(\mathrm{R_{FR}})$ = output of LSTM classification 1 for risk $\mathrm{R_{FR}}$

$\mathrm{LSTM_n}(\mathrm{R_{FR}})$ = output of LSTM classification *n* for risk $\mathrm{R_{FR}}$

m= total number of reviews

Fake Review Index (FRI):

$$\mathrm{R_{FRI}} = \sum_{i=1}^{m} f\{Bi, LSTM_1(\mathrm{R_{FR}}), LSTM_2(\mathrm{R_{FR}}), \ldots LSTM_n(\mathrm{R_{FR}})\}$$

Topic weighted accuracy increment: more in depth machine and human led topic mapping from Study 2 can help to develop more advanced multiclass classifiers to determine one or more critical risk factor(s) related to individual topic, e.g., product criticism, price criticism, delivery complaint.

$$P(\,R_{FR} \mid d, y\,) = P(dy)P'(z)\; log\left(\frac{P(d,y)}{P(d)P(y)}\right)$$

$P(dy)$ = probability of a review (d) containing identified keyword (y). It captures the frequency of co-occurrence of selected words within a review.

$P'(z)$ = probability associated with a particular topic (z). Helps to identify topics that are more prevalent.

$log\left(\frac{P(d,y)}{P(d)P(y)}\right) =$ calculates the mutual prevalence of targeted keywords (y) within a review, measuring how much misinformation selected keywords contribute to a review topic.

If the LSTM outputs a probability distribution over risk categories, P represents the expected value of these probabilities, weighted by the severity of the risk each class represents. The equation would thus provide a dynamic risk index that changes over time based on the outputs of the binary classifier and the multiclass LSTM model output.

Continuous risk assessment and associated cost:

$$R_{CRA} \int_{b}^{a} p\{\mathrm{R(t)}\} * c\{\mathrm{R(t)}\} dx$$

$P\{\mathrm{R(t)}\}$ and $C\{\mathrm{R(t)}\}$ are continuous functions defining the probability density and cost as a function of risk $\mathrm{R(t)}$ over the interval from a to b.

4.4.1 LIME for Performance vs Explainability:

Previous studies (Stevenson, Mues, & Bravo, 2021) have largely discredited the use and application of LIME in measuring model performance and other wider explainability benefits. LIME has been criticised for isolating localised words without context, failure to increase or decrease predictability, failure to capture complex interactions, and trading off model performance to add explainability.

In this section, we provide a vital managerial insight by highlighting how strategic application of LIME can help to optimise model performance vs model explainability. We further highlight how the application of LIME can contribute toward better model selection, in addition to calculating the Confidence Score for the FRI. Application of LIME offers three specific benefits to our approach – (i) measure Performance vs Explainability, (ii) FRI Confidence Score calculation, and (ii) complementary local explanation

LSTM Model Performance vs Explainability – Figure 11 results demonstrate the insights gained by supplementing LSTM evaluation with LIME explanation analysis. Following the principles of Attribute Analysis (AA), we describe how managers can benefit from such applications. Using this technique managers, can identify models that are both accurate and interpretable, while also finding opportunities to improve alignment between a model's predictions and explanations. Evaluating explainability alongside accuracy provides a more complete view of real-world model performance.

Figure 11 shows that higher model performance leads to high interpretability, but this relationship is not linear. For enterprise managers, this guides the selection of the optimal model configuration for a given case based on whether interpretability or accuracy is more important. Differences in LIME AUC for the same LSTM architecture could indicate variability in how explanations align with model reasoning across different data samples.

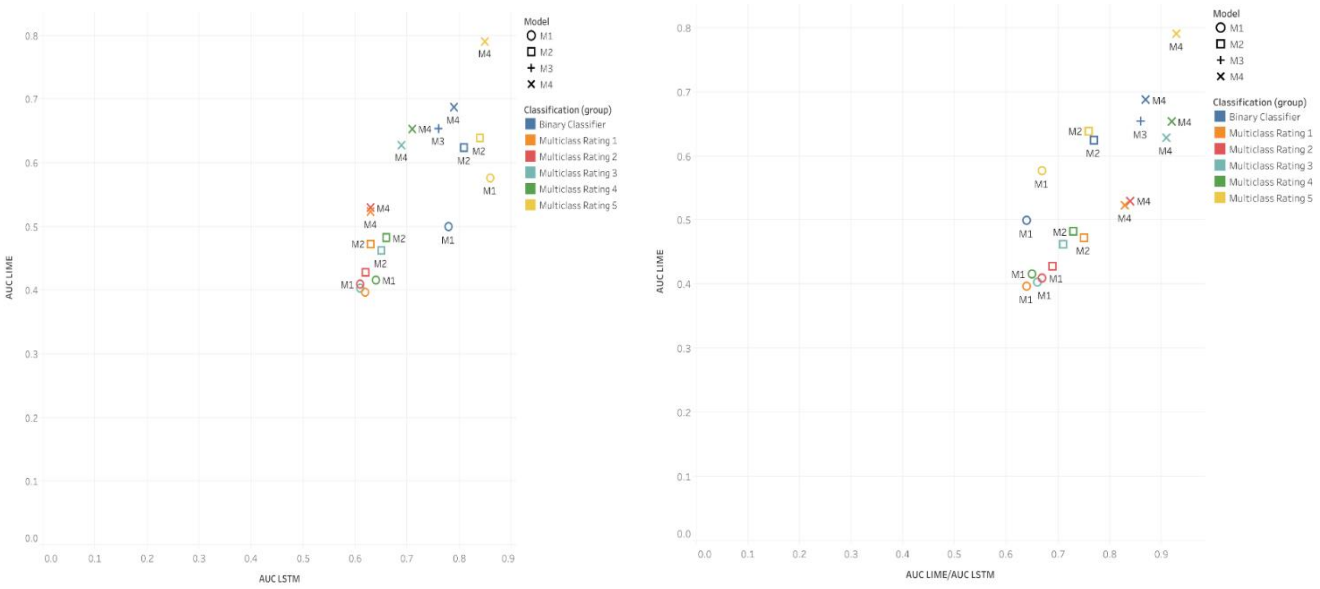


Figure 11: Model Performance vs Explainability

Note: models with a higher AUC LIME/AUC LSTM ratio are more interpretable relative to their LSTM performance. Models with a higher AUC LIME value have a better standalone interpretability.

In this scenario, optimal models appear to be M4, located in the top right quadrant, indicating a good balance between high performance (measured by AUC LSTM) and high interpretability (measured by AUC LIME). Figure 11 indicates that Model 4 (M4) outperformed the interpretation task compared to other models in similar classification task. These findings conclusively consolidate the importance of Semantic Topic Modelling (Study 2) in raising the interpretation capability of LSTM.

Lower LIME AUC signals regions where the model's logic is less sound. Managers could examine if certain model architectures tend to produce more stable LIME explanations across different data samples. This would identify models with more robust and reliable reasoning. For models with similar LSTM AUC, practitioners could select the one with higher LIME AUC for deployment to maximise interpretability. However, they would need to verify that the explanations are valid and aligned with previous reasoning.

Clustering samples by LIME AUC could also help identify subsets of data where model reasoning differs significantly from the overall trends. Statistical analysis of the relationship between LSTM and LIME AUC could quantify this correlation more precisely to guide architecture optimisation.

Confidence Score - by providing numeric confidence scores, LIME enables us to assess the reliability of each individual prediction made by the LSTM model. We can calculate these confidence scores alongside FRI to alert predictions with higher uncertainty. Overall, the local surrogate models created by LIME are crucial for the approximation of the LSTM model's behaviour and for converting differences into calibrated certainty estimates. The confidence scores powered by LIME improve the bi-LSTM model's robustness and utility by quantifying uncertainty.

Confidence Score (CS)

AUC LIME validates the quality of explanations. AUC LSTM represents overall classification performance. Together these two metrices provide complementary information about model interpretability and model accuracy, validating the reliability of the Fake Review Index (R_{FRI})at a given time.

CS = 1 - (AUC of LIME / AUC of LSTM)

So if the LSTM has an AUC of 0.95 and the LIME model has an AUC of 0.85, the Confidence Score would be: CS = 1 - (0.85 / 0.95) = 0.105

AUC LIME = AUC LSTM, indicates optimum performance.

AUC LIME < AUC LSTM, means that the ratio is < 1, so the difference is positive. Therefore, the LSTM has superior performance compared to LIME, highlighting model explainability trade-off.

AUC LIME > AUC LSTM, means that the ratio is > 1, so the difference is negative. This indicates LIME explanations correlate better to the original model.

This value represents the performance we 'lose' by using the interpretable LIME model. It quantifies the trade-off between model performance and interpretability when explaining a black box model with LIME.

Local Explanation - LIME is not used for topic modelling directly in our study, but it rather provides complementary local explanations that, combined with the semantic topics from Top2Vec, give a richer understanding of the model's fake review detection logic. Specifically, LIME is used to analyse the importance of local features for the model's fake review predictions. These local explanations help characterise the linguistic properties of fake reviews identified by the model.

By inspecting the words and phrases that the model weights highly in flagged fake reviews, LIME provides additional linguistic insights into what makes a review inauthentic. These LIME-based linguistic insights bolster and expand upon the topics and semantic meaning extracted through Top2Vec topic modelling.

Together, the semantic topics extracted by Top2Vec and the local lexical explanations from LIME allow us to better understand the textual cues used by the model to detect fake reviews (**Appendix 6**).

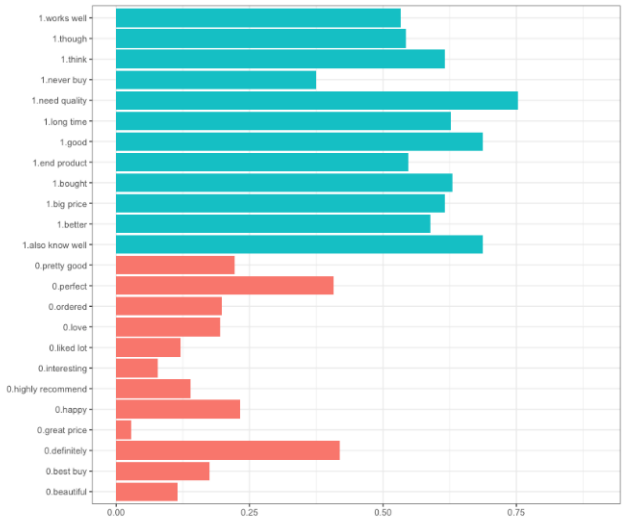


Figure 12: LIME results from the binary classification of fake versus original reviews. The green bars show strongest associations with the original reviews and the red bars show strongest associations with the fake reviews.

Table 5:
Key phrases and associated themes identified through black-box interpretation

Review led black-box interpretation:	
Fake Reviews: • Expressive phrase: ‘liked lot’ • Elevated phrase: ‘best buy’ • Endorsing phrase: ‘pretty good’ • Action oriented phrase: ‘ordered’ • Rewarding phrase: ‘beautiful’	**Original Reviews:** • Cautionary phrase: ‘think’ • Derogatory phrase: ‘never buy’ • Endorsing phrase: 'works well’ • Action oriented phrase: ‘bought’ • Exclamatory phrase: 'big price’
Rating led black-box interpretation:	
Low-rated fake reviews: • Deterrent: ‘problem’ • Derogatory: ‘bad’, ‘old’ • Uninviting: ‘didn’t find’	**Highly-rated fake reviews:** • Endorsing: ‘highly recommended’ • Encouraging: ‘good price’ • Supportive: ‘must buy’ • Anticipation: ‘long need’ • Positive reinforcement: ‘beautiful’, ‘love’ • Alternative purchase: ‘gift’

Note: themes were researcher generated based on the keywords and their context identified using LIME.

On further researcher interpretation it appears that fake reviews mostly comprise highly positive and endorsement related comments. Selective fake review keywords identified are assertive, elevated, and expressive in nature. In contrast, original reviews are comprised of more balanced phrases. There is a balanced amount of key phrases that are more inquisitive, cautionary and exclamatory in nature. Topic clustering showed fake reviews centralised around generic satisfaction words rather than specifics about product attributes and performance. Real reviews covered a wider range of meaningful product-related topics. Fake reviews disproportionately focus on positive emotions and experiences without providing evidence or rationale. Real reviews had a more balanced assessment, mixing positives and negatives backed by details.

5. Conclusion and future research

In this study, we present a novel pathway towards developing a fake review risk index for enterprises. Such initiative is aimed at helping small to medium-sized enterprises to develop and index business failure risks posed by fake reviews across product categories. We use a combination of novel ML and DL approach that benefits from the semantic textual knowledge embeddings of BERT, but reproduces the complexity using a much more scalable and controllable bi-LSTM architecture. We demonstrate that using an unsupervised semantic topic indexing technique (i.e. Top2Vec), supported by human interpretation, leads to the development of semantically modified topics representing fake review characteristics within individual product categories. Adding this layer of information to the proposed bi-LSTM architecture increases model performance and explainability alike. Our proposed model benefits from a 20-fold cross-validation technique and is capable of producing two classification metrics towards the development of the Fake Review Index. The modular aspect of our study offers transparency and explainability accustomed to industry practitioners. In addition, explainability dimensions of our semantic topic models offer thematised fake review linguistic insights to practitioners across different product categories (Books, Movies, Digital Music etc.). Finally, we demonstrate that although explainability features associated with LIME suffer from poor reputation due to restricted interpretation capability, if applied strategically, it can be used to measure and optimise LSTM model

performance against LSTM model explainability. This can be a vital resource for managers trying to settle for either the best performing or the best interpretable model, while trying to evaluate multiple neural network architectures. In our case, LIME output can also be quantitively analysed to develop a Confidence Score (CS), representing FRI reliability at any given time. From, a qualitative perspective, LIME-based linguistic insights also serve as a benchmark for previously extracted topics and semantic meanings. This study is exclusively focused on the negative impact of fake reviews; this can be perceived as the limitations of this research. However, we encourage the next generation of researchers to consider how manipulated positive reviews can impact an organisation's competitiveness when competitors gain an unfair advantage by employing third parties. The study is also exclusively focused on Amazon review data, future studies must incorporate data from other e-commerce platforms.

Inspired by Borchert et al. (2023), our study affirms that unstructured textual data is going to play a pivotal role in predicting business performance in the near future. Academic researchers and practitioners need a conceptual paradigm shift in developing more unstructured data-driven KPIs. Future research needs to demonstrate more openness and adaptability towards introducing textual data from various aspects of business and quantifying them to develop better performance metrices. On this note, we further complement De Bock et al.'s (2023) initiative towards galvanising more explainable attribute analytics (AA) within operational research. As opposed to achieving iterative model performance accuracy and obfuscating model complexity, academic research must also lend a hand to industry practitioners in developing novel attribute analytics methods. Our modular study design provides a clear roadmap for the partitioners to modify and develop Fake Review Indices that are better suited to their organisational risk monitoring system. The FRI offers a quantitative metric, updated in real-time, that gives better visibility into fake review risks. This enables organisations to deploy both proactive and reactive strategies to mitigate fake review risks. Following the ethos of responsible analytics (RA), the FRI can also help identify fake review based unfair practices, ensuring competitive fairness.

Our systematic and scientific approach addresses a critical problem at a critical time when businesses are suffering immense amounts of financial and reputation losses due to fake reviews. Future research in this area has immense potentials, and the stream can grow by investigating and improving the FRI using additional metadata such as product price, reviewer profile, reviewer historic activities, review sentiment, review timing, review readability etc. Other ensemble approaches can also be developed on our proposed study prototype, and the explainability features can be more robustly evaluated by separating human-generated fake reviews from machine-generated ones. Improved fine-tuning and application of more sophisticated large language engines, such as GPT-5, with advanced semantic modelling capabilities, can help improve the index's accuracy and fake review detection capability. Future research can also expand laterally as our proposed principles can also be applied in other e-commerce platforms beyond Amazon to develop a more universal model of e-commerce FRI.

References

Amazon (2022). Amazon continues to invest in the growth of European small and medium enterprises. *Amazon News*, November, https://www.aboutamazon.eu/news/empowering-small-business/amazon-continues-to-invest-in-the-growth-of-european-small-and-medium-enterprises

Angelov, D. (2020). Top2Vec: Distributed representations of topics. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2008.09470.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Banerjee, S., & Chua, A. Y. K. (2023). Understanding online fake review production strategies. *Journal of Business Research*, *156*, 113534. https://doi.org/10.1016/j.jbusres.2022.113534

Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, *56*(4), 1234-1244. https://doi.org/10.1016/j.ipm.2019.03.002

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bathla, G., Singh, P., Singh, R. K., Cambria, E., & Tiwari, R. (2022). Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. *Neural Computing and Applications*, *34*(22), 20213-20229. https://doi.org/10.1007/s00521-022-07531-8#=

BBC (2023). Amazon unveils Alexa-powered home robot. https://www.bbc.com/news/technology-58727057

Belkina, A. C., Ciccolella, C. O., Anno, R., Halpert, R., Spidlen, J., & Snyder-Cappione, J. E. (2019). Automated optimized optimised parameters for T-distributed stochastic neighbor embedding improve visualization visualisation and analysis of large datasets. *Nature communications*, *10*(1), 1-12.

Borchert, P., Coussement, K., De Caigny, A., & De Weerdt, J. (2023). Extending business failure prediction models with textual website content using deep learning. *European Journal of Operational Research*, *306*(1), 348-357. https://doi.org/10.1016/j.ejor.2022.06.060

Briskilal, J., & Subalalitha, C. N. (2022). An ensemble model for classifying idioms and literal texts using BERT and RoBERTa. *Information Processing & Management*, *59*(1), 102756.

Chakraborty, I., Kim, M., & Sudhir, K. (2022). Attribute sentiment scoring with online text reviews: Accounting for language structure and missing attributes. *Journal of Marketing Research*, 59(3), 600–622. https://doi.org/10.1177/00222437211052500

Colley, A., Väänänen, K., & Häkkilä, J. (2022, November). Tangible Explainable AI-an Initial Conceptual Framework. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia* (pp. 22-27).

Darwish, A. (2022). Explainable Artificial Intelligence: A New Era of Artificial Intelligence. *Digital Technologies Research and Applications*, *1*(1), 1. https://doi.org/10.54963/dtra.v1i1.29

De Bock, K. W., Coussement, K., Caigny, A. D., Słowiński, R., Baesens, B., Boute, R. N., Choi, T.-M., Delen, D., Kraus, M., Lessmann, S., Maldonado, S., Martens, D., Óskarsdóttir, M., Vairetti, C., Verbeke, W., & Weber, R. (2023). Explainable A.I. for Operational Research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2023.09.026

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational A.I. for research, practice and policy. *International Journal of Information Management, 71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Egger, R., & Yu, J. (2022). A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, *7*, 886498. https://doi.org/10.3389/fsoc.2022.886498

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*. https://doi.org/10.1007/s12559-023-10179-8

Hajek, P., Hikkerova, L., & Sahut, J. M. (2023). Fake review detection in e-commerce platforms using aspect-based sentiment analysis. *Journal of Business Research*, 167, 114143.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science*. https://doi.org/10.1287/mksc.2022.1353

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Huang, X., Khetan, A., Bidart, R., & Karnin, Z. (2022). Pyramid-BERT: Reducing complexity via successive core-set based token selection. *arXiv preprint arXiv:2203.14380*.

Jabeur, S. B., Ballouk, H., Arfi, W. B., & Sahut, J. M. (2023). Artificial intelligence applications in fake review detection: Bibliometric analysis and future avenues for research. *Journal of Business Research*, *158*, 113631. https://doi.org/10.1016/j.jbusres.2022.113631

Jang, B., Kim, M., Harerimana, G., Kang, S. U., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. *Applied Sciences*, *10*(17), 5841.

Janssens, B., Schetgen, L., Bogaert, M., Meire, M., & Van den Poel, D. (2023). 360 Degrees rumor detection: When explanations got some explaining to do. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2023.06.024

Jiang, G., Shang, J., Liu, W., Feng, X., & Lei, J. (2020). Modeling the dynamics of online review life cycle: Role of social and economic moderations. *European Journal of Operational Research*, *285*(1), 360-379. https://doi.org/10.1016/j.ejor.2020.01.054

Joung, J., & Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management*, *70*, 102641. https://doi.org/10.1016/j.ijinfomgt.2023.102641

Kaemingk, D. (2020). Customer Experience: Online reviews statistics to know in 2022. *XM Blog*, October. https://www.qualtrics.com/blog/online-review-stats/

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, 80(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, *302*(1), 309-323.

Lu, L., Xu, P., Wang, Y.-Y., & Wang, Y. (2023). Measuring service quality with text analytics: Considering both importance and performance of consumer opinions on social and non-social online platforms. *Journal of Business Research, 169*, 114298. https://doi.org/10.1016/j.jbusres.2023.114298

Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016).

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743-758. https://doi.org/10.1016/j.ejor.2018.10.024

Moon, S., & Kamakura, W. A. (2017). A picture is worth a thousand words: Translating product reviews into a product positioning map. *International Journal of Research in Marketing*, *34*(1), 265-285.

Mousavizadeh, S. M., Maghsoodi, S., & Ibrahim, O. (2022). *Machine learning for e-commerce*. Springer.

Ni, J., Li, J., & McAuley, J. (2019). *Empirical Methods in Natural Language Processing (EMNLP)*. Amazon review data [Data set]. https://nijianmo.github.io/amazon/

Plotkina, D., Munzel, A., & Pallud, J. (2020). Illusions of truth—Experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research*, *109*, 511-523.

Román, S., Riquelme, I. P., & Iacobucci, D. (2023). Fake or credible? Antecedents and consequences of perceived credibility in exaggerated online reviews. *Journal of Business Research*, 156, 113466.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services*, 64, 102771. https://doi.org/10.1016/j.jretconser.2021.102771

Statista (2024). Annual net revenue of Amazon.com from 2004 to 2023. Retrieved from https://www.statista.com/statistics/1264135/amazon-net-sales/

Statista (2022), Net revenue of Amazon from 1st quarter 2007 to 3rd quarter 2022, https://www.statista.com/statistics/273963/quarterly-revenue-of-amazoncom/

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, *295*(2), 758-771

Tufail, H., Ashraf, M. U., Alsubhi, K., & Aljahdali, H. M. (2022). The effect of fake reviews on e-commerce during and after Covid-19 pandemic: SKL-based fake reviews detection. *Ieee Access*, *10*, 25555-25564.

Vanover, C., Mihas, P., & Saldaña, J. (Eds.). (2021). *Analyzing and interpreting qualitative research: After the interview*. Sage Publications.

Wang, E. Y., Fong, L. H. N., & Law, R. (2022). Detecting fake hospitality reviews through the interplay of emotional cues, cognitive cues and review valence. *International Journal of Contemporary Hospitality Management*, *34*(1), 184-200.

Which. (2019). Exposed: the tricks sellers use to post fake reviews on Amazon. Retrieved from https://www.which.co.uk/news/article/exposed-the-tricks-sellers-use-to-post-fake-reviews-on-amazon-a9kJS7X2kMB0

Wong, T. T., & Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586-1594. 10.1109/TKDE.2019.2912815

World Economic Forum (2021). Fake online reviews cost $152 billion a year. Here's how e-commerce sites can stop them. https://www.weforum.org/agenda/2021/08/fake-online-reviews-are-a-152-billion-problem-heres-how-to-silence-them/

Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems*, *132*, 113280. https://doi.org/10.1016/j.dss.2020.113280

Xu, Y. Z., Zhang, J. L., Hua, Y., & Wang, L. Y. (2019). Dynamic credit risk evaluation method for e-commerce sellers based on a hybrid artificial intelligence model. *Sustainability*, *11*(19), 5521. https://doi.org/10.3390/su11195521

Zhang, D., Li, W., Niu, B., & Wu, C. (2023). A deep learning approach for detecting fake reviewers: Exploiting reviewing behavior and textual information. *Decision Support Systems*, 166, 113911.

Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews. *Journal of Management Information Systems*, 33(2), 456–481. https://doi.org/10.1080/07421222.2016.1205907

Zheng, X., Zhang, C., & Woodland, P. C. (2021, December). Adapting GPT, GPT-2 and BERT language models for speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 162-168). IEEE.

Zhu, D., Lu, S., Wang, M., Lin, J., & Wang, Z. (2020). Efficient precision-adjustable architecture for softmax function in deep learning. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *67*(12), 3382-3386.

Zhuang, M., Cui, G., & Peng, L. (2018). Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research*, 87, 24–35. https://doi.org/10.1016/j.jbusres.2018.02.016