Towards a Review-Analytics-as-a-Service (RAaaS) Framework for SMEs: A Case Study on Review Fraud Detection and Understanding

Xuan Truong Du Chau[1], Thanh Toan Nguyen[2], Vinh Khiem Tran[3], (iD)
Sara Quach[4], (iD) Park Thaichon[4], (iD) Jun Jo[1], Bay Vo[4,5],
Quang Dieu Tran[5] and Quoc Viet Hung Nguyen[1]

Abstract

With the advancement of internet technology, customers increasingly rely on online reviews as a valuable source of information. The study aims to develop a marketing data analytics framework to manage online reviews, especially fake reviews, which have become a significant issue undermining the creditability of online review systems. As small and medium-sized enterprises often lack the capabilities to automatically derive customer insights from online reviews, this study proposes a cost-effective, extensible Review-Analytics-as-a-Service (RAaaS) framework that can be operated by non-data specialists to facilitate online review data analytics. We demonstrate the framework's application by using two datasets with more than 400,000 online reviews from Yelp to simulate live platforms and demonstrate an analytic flow of review fraud detection and understanding. The findings reveal insights into the influence of fake reviews on product ranking and exposure rate. Moreover, it was found that there was a higher concentration of sadness and anger in fake reviews (vs. organic reviews). In addition, fake reviews tend to be shorter, more extreme (with the use of strong adverbs), and have different patterns of topic distribution. This study has important implications for different stakeholder groups including, but not limited to, SMEs, review platforms and customers.

Introduction

Big data are high velocity, large volume and multiple source data resulting from digitising daily activities, including interactions, online transactions and communications (De Luca et al., 2021; McAfee et al., 2012). User-generated content is an important source of big data that can be converted into useful information for modern businesses (Kauffmann et al., 2020). Online review is one of the most crucial generated content that significantly affects brand engagement because it directly reflects the customer experience with products or services (de Vries et al., 2012). Moreover, with the advances in internet technologies, customers mainly rely on reviews as a primary source to explore product information (Robertson et al., 2021). A recent study (Ha et al., 2015) reveals that 52% of customers explore product information in advance via the internet, and 24% of them decide to buy products based on these online insights. Therefore, product reviews play a crucial role in the success of a business as they directly influence customers' purchase decisions (Choi et al., 2017). Recent research has highlighted the shift to outside-in marketing, which focuses on sensing the market and creating connections with customers (Quach et al., 2022). Therefore, understanding what customers say about firms and their products in a timely manner constitutes a critical factor in succeeding in the era of big data.

On the other hand, fraudulent reviews have become a significant issue undermining the creditability of online review systems (Luca & Zervas, 2016). Fake reviews give authentic firms and customers the wrong signals, which can manipulate the ranking and sales, thereby negatively impacting genuine businesses (He et al., 2022).

Nonetheless, deriving customer insights from online reviews and detecting fake reviews involved a complex process that requires big data capability (Kauffmann et al., 2020). Reviews of the same product may come from multiple data sources, ranging from online selling platforms (e.g. Amazon, eBay) to product review providers (e.g. Yelp, TripAdvisor). Therefore, conventional review analytic frameworks, which are designed and optimised for a specific domain

[1]Griffith University, Brisbane, QLD, Australia
[2]HUTECH University, Ho Chi Minh City, Vietnam
[3]University of Information Technology, Ho Chi Minh City, Vietnam
[4]Vietnam National University, Ho Chi Minh City, Vietnam
[5]Ho Chi Minh National Academy of Politics, Hà Nội, Vietnam

Corresponding author:
Thanh Toan Nguyen, Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Viet Nam.
Email: nt.toan@hutech.edu.vn

of applications such as Yelp (Lee et al., 2022; Luca & Zervas, 2016), app store reviews (Kauffmann et al., 2020) and Amazon reviews (Kauffmann et al., 2020), are unable to generalise to multiple and diverse data sources, thereby suffering a limitation of extensibility. In addition, these approaches are not cost-effective as the company needs continuous infrastructure investment and maintenance, which hinders small and medium-sized enterprises (SMEs) from leveraging customer insights for business planning and development. Although SMEs could save their cost by leveraging emerging cloud services that enable the storage, analysis and visualisation of data, and paying per usage, they require specialised knowledge of analytics algorithms to conduct a review analysis pipeline. To sum up, designing a cost-effective analytics platform for marketing stakeholders who are not data analysis experts to conduct data analytic tasks is still an open problem.

Therefore, we aim to address the gap in the extant literature pertinent to fake reviews by developing the framework that allows marketers from SMEs to adopt fraudulent review detection and understanding. It is worth noting that marketing stakeholders are not required to be data experts as the proposed framework offers algorithm templates that are manageable to apply. The main research question (RQ) of our study is how to exploit the core values of big data and provide a proactive platform for non-data specialists to trigger enormous tasks on review analytics. While recent research (Kauffmann et al., 2020; Salminen et al., 2022) analysed fake reviews and proposed a detection method, and it does not necessarily help to understand fake reviews with multi-faceted insights, including textual patterns, topics and emotions. Therefore, this study aims to provide an in-depth understanding of fake reviews and focus on the additional sub-research questions as follows: RQ1: Can review fraud be automatically detected?; RQ2: How do review frauds affect the product ranking system?; and RQ3: What are the differences between fake and organic reviews?

The study contributes to the current body of knowledge in the following ways. First, overcoming the limitation of conventional framework (Kauffmann et al., 2020), we are among the first that propose a state of the art, cost-effective, extensible cloud-based platform consisting of data acquisition and data analytic planes, that can be operated by non-data specialists to facilitate online review data analytics in marketing. Although the main focus of the study is to support SMEs in implementing review data analytics without expert guidance, their potential adoption is open to businesses of all sizes that want to harness customer data with minimal investment cost. Second, via the use of a case study, we demonstrate the employment of multiple algorithms to detect fake reviews and provide insights into how contemporary detection methods' accuracy can help detect and filter fake reviews. Finally, we evaluate the impact of fake reviews on business activities and determine multifaceted insights and characteristics that distinguish fake reviews from organic reviews, extending previous research such as Kauffmann et al. (2020) and Luca and Zervas (2016). The structure of our paper is as follows. The next section reviews the literature related to the study followed by our proposed RAaaS framework which details the necessary phases for the data acquisition plane and details analytics tasks available in the cloud plane. We then present a case study to demonstrate the use of the framework together with the analysis and intensive discussion on insights. Finally, we conclude our paper with some limitations and future research directions.

Background

In this section, we first review related work on big data analytics in marketing (section 2.2). As our case study is based on fake review detection and understanding, and we then discuss related work on fake review detection (section 2.3).

Challenges for SMEs in big data adoption

Although it would be incorrect to assume that all SMEs cannot afford a private-owned big data analytics framework, the cost typically plays a crucial role. Therefore, it is a possible barrier for SMEs to adopt big data analytics to facilitate their business activities. The reason could be that there is a natural cost-functionality trade-off, as the investment cost outweighs the potential benefits (Coleman et al., 2016), or it is simply that the cost is beyond the capability of an SME. Additionally, SMEs are concerned about the main challenge associated with adopting big data analytics, that is, the skills required to manage it. Not only because these expertises are challenging to locate, but they are also expensive to acquire (Del Vecchio et al., 2018). Moreover, SMEs frequently lack in-house capabilities for selecting, configuring and maintaining sophisticated IT systems; such considerations are another challenge for SMEs to apply big data analytics (Schmidt & Möhring, 2013). With the rise of Software as a Service (SaaS; Kasemsap, 2021) and cloud computing, SaaS could simplify access to elaborate data stores and computer networks. Prior studies suggested that many SMEs will adopt cloud computing due to its scalable nature and relatively low upfront costs (Sultan, 2011). In order to make big data analytics more accessible to SMEs, we propose the idea of big data analytics being given as a cloud-based platform, which is focused on online review analytics for customer behaviour understanding.

Big data analytics framework and online reviews

Big data analytics can transform data from multiple sources into actionable insights, which can be used to maximise the effectiveness of the marketing campaigns and, optimise the enterprises' return on investment (Quach et al., 2022). Online reviews are essential in the era of big data revolution (Erevelles et al., 2016) as these data provide customers' behavioural insights that directly affect their buying decisions. In addition, customer reviews allow firms to sense the market and to build lasting relationships with their customers (Munzel, 2016). Despite the importance of online reviews which often come from various sources such as business websites, social networking sites and e-commerce platforms, most existing works in literature to extract customer insights from online reviews are domain-specific. It means that they are designed to work with a specific data domain. For example, Barbado et al. (2019) proposes a framework for review analytics with applications on fake review detection, and the authors evaluate the framework's performance using reviews on an electronic domain. Jimenez-Marquez et al. (2019) proposes a two-stage framework for analysing social network content and presents a case study to analyse online reviews for tourism-related businesses. Martens and Maalej (2019) proposes an analytic process for understanding online user feedback for supporting decisions for mobile app launching. As being optimised to a data domain, such approaches are hard to generalise and extend to multiple data sources, while reviews of the same product may come from multiple domains. The recent work (Kauffmann et al., 2020) proposes a generic framework that can be applied to multiple domains of online review. However, it requires continuous infrastructure investment and maintenance, which may not be cost-effective for the majority of SMEs, which often have limited technological and financial resources.

To resolve the limitations of existing approaches, we propose Review-Analytics-as-a-Service (RAaaS) – a platform to facilitate

online review data analytics in the cloud that supports marketing-specialised data analytic algorithm templates and enables those not experts in analytic algorithms to discover the customers' insights. Cloud technology offers ready-to-use resources, including computing power, applications and services in the cloud environment in which users can minimise their cost investment without worrying about resources' availability (Quach et al., 2022). Cloud technology has been widely utilised in various fields such as manufacturing (Zulkernine et al., 2013), smart cities analytics (Khan et al., 2015), automation in construction (Chen et al., 2016), health-care (Bastwadkar et al., 2020), design innovation (Liu et al., 2020). This study is among the first to apply this framework in the marketing discipline.

Fake review detection

Internet technology has changed the way customers buy products, and fake reviews have gradually manipulated these purchasing decisions. Fake reviews might be written to promote or demote a business and to intentionally misinform consumers in their decision-making (Choi et al., 2017). Therefore, fraudulent reviews are detrimental because they introduce disturbance and false signals to systems that are meant to reduce information asymmetry, resulting in consumers' purchase of unreliable products, disadvantaging genuine businesses by manipulating ranking and promoting unhealthy competition, and undermining customer trust in the platforms (He et al., 2022). Whereas a major research stream has focused on consumers' perspectives, such as their ability to detect fake reviews (Malbon, 2013; Munzel, 2016) and intention to write fraudulent reviews (Choi et al., 2017), there has been increasing interest in how firms manage online reviews and detect fraudulent attempts (Kauffmann et al., 2020; Salminen et al., 2022). Specifically, online service firms are facing the alarming situation of purchasing fake reviews, which has been highlighted in a recent study (He et al., 2022). Fake review detection is the task of classifying online reviews into a class of fake and organic reviews. Fake review detection methods can be divided into two categories: supervised and unsupervised methods. Supervised methods (Saumya & Singh, 2018) learn a classifying model from labelled information (i.e. fake and organic reviews). Unsupervised methods apply clustering techniques (Liu & Pang, 2018) and graph-based analysis (Ye & Akoglu, 2015) for fake review detection without requiring labelled data. A graph partitioning approach (Manaskasemsak et al., 2023) is proposed to prevent the deceiving of untruthful reviews on product quality and fair commercial benefits. However, this approach focuses on distinguishing fake reviewers from benign ones rather than fake review detection. Similarly, more approaches are raised to combat the detrimental effects of fake reviews (Z. Wang et al., 2020; F. Zhang et al., 2020). Nevertheless, they focus on the group of reviewer detection by considering both time effects and network effects. More recent studies (Birim et al., 2022; W. Zhang et al., 2022) explore the feasibility of employing topic modelling as a trait for fraudulent review detection. Although these methods yield moderate performance, it is unclear how topic distributions affect fraudulent review detection, and the generalisation to various domains is inadequate (X. Wang et al., 2018). In this study, replying to the call of recent research in using machine learning for fake review detection (Salminen et al., 2022), we integrated state-of-the-art supervised and unsupervised methods and made them available in our framework using an algorithm template. A summary of the features used and characteristics of these methods shall be presented in section 5.3, where we reveal a variety of data analytic tasks to facilitate fake review detection and understanding.

The proposed framework

Towards a cost-effective and extendable review data analytics, we propose Review-Analytics-as-a-Service (RAaaS) framework. Before giving an overview of the framework, we summarise its underlying design principles.

Design principles. Reflecting on emerging applications of extracting insights from online reviews for SMEs (Jimenez-Marquez et al., 2019), the limitations of existing work in literature, and on the analytic results reported for state-of-the-art approaches (Kauffmann et al., 2020), we derive the following requirements for a practical framework:

- (R1) *Cost-effective*. The framework shall allow a low-resource setting where high-resource required tasks will be performed within the cloud, and the cloud cost shall be shared between enterprises following a *pay-as-you-go* paradigm.
- (R2) *Non-expert required*. The framework shall enable those not specialising in data science to ultilise data analytic techniques to derive customer insights to facilitate marketing intents.
- (R3) *Extensibility*. The framework shall allow enterprises to extend to multiple data sources as reviews of the same product may come from multiple sources ranging from online selling platforms to specialised online review providers.

Framework overview. The overview of the proposed Review-Analytics-as-a-Service (RAaaS) framework is presented in Figure 1. The framework consists of two planes: the *aquisition* plane and the *analytic* plane.

- *Data acquisition plane*: The function of this plane is to retrieve and structure data generated from various data sources. The enterprise can extend more data sources over time as long as they follow the data structure specified by the data transmission component of the plane. This way, we achieve extensibility (R3) as the enterprises shall be proactive in how many data sources they want to perform the analysis. Details of this component shall be given in section 4.
- *Data analytics plane*: After receiving the high-quality data transmitted from the data acquisition plane, in this plane, we provide various services to perform analysis tasks. The enterprises only pay per their usage of analytic resources and thus, being cost-effective (R1). In addition, the analytical tasks are independent of the analytic experts as each analytic task is built in the form of an algorithm template. Quickly selecting the algorithm template and selecting minimum variable inputs are now open to a non-expert (R2) through the concepts of algorithm template. Details of this component shall be given in section 5.

Data acquisition plane

The first plane of RAaaS proposes sequential phases that represent the pipeline of data processing performed locally at the enterprise site, including data entry (section 4.1), data retrieval (section 4.2), data structuring (section 4.3) and data transmission (section 4.4) with buffering component. The purpose of this plane is to ensure the data availability and quality rather than the complex analysis and,
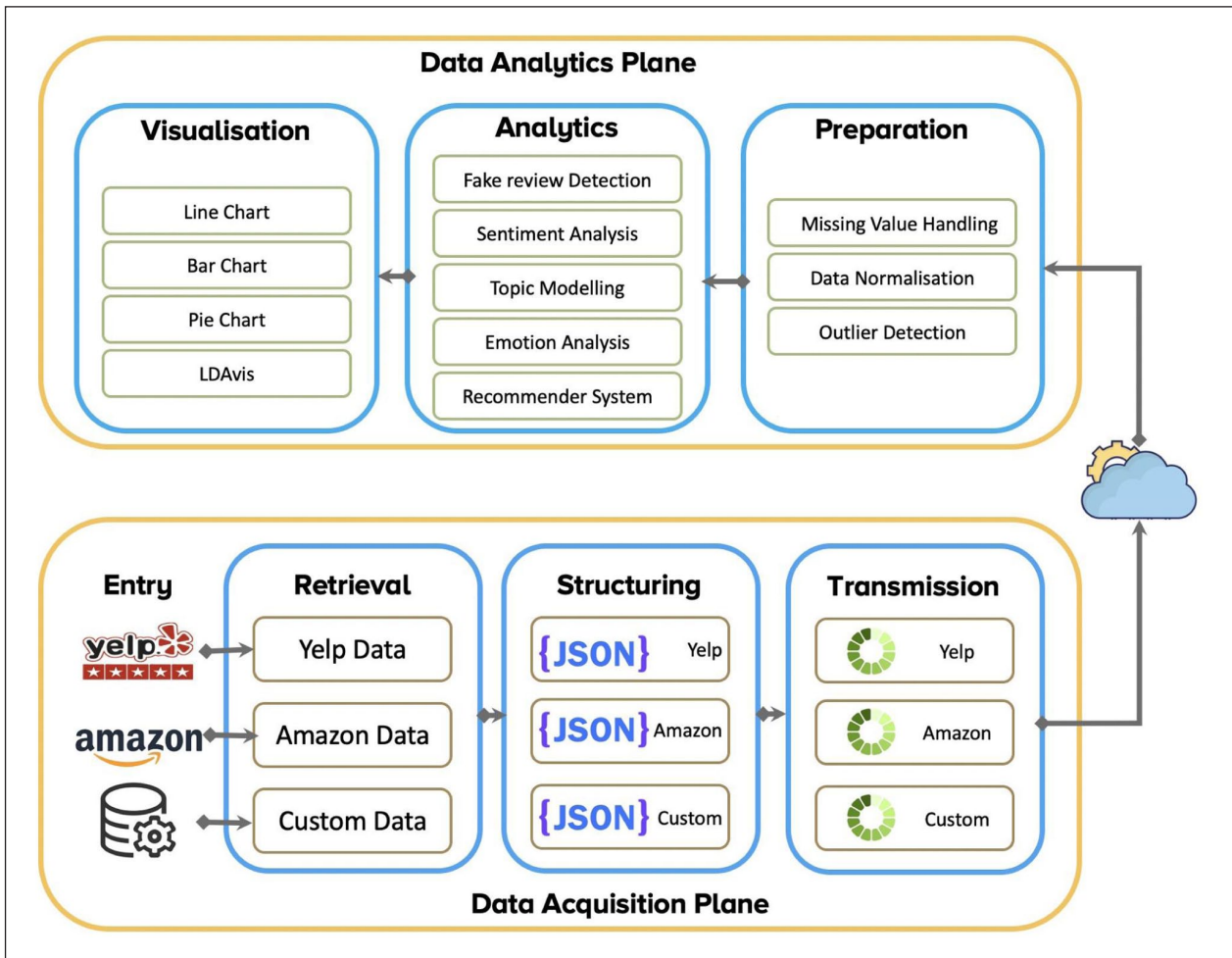
Figure 1. The big data analytic framework for fraudulent reviews.

therefore not required for an analytic expert which is satisfied with the (R2) of design principles.

Data entry

This work aims to propose a cloud-based framework that focuses predominantly, but not exclusively, on a collection of online reviews. The data input to the framework is mainly coming from the textual domain. These textual data can come from online selling platforms (e.g. Amazon, eBay) to product review providers (e.g. Yelp, TripAdvisor). As a demonstrated study, tourism data is chosen to perform data analytics, especially reviews on restaurants and hotels; however, the framework is well adapted to any data source such as questionnaires, surveys, or data collected from offline marketing campaigns.

Data retrieval

To analyse product reviews, it is essential to first accurately identify the reviews and collect them in a timely manner. Therefore, the initial phase of this plane is data collection, which involves adopting the necessary data collection techniques to obtain relevant data. Data collection techniques allow collecting diverse information using web scrappers (Kauffmann et al., 2020) or third-party services (Zulkernine et al., 2013). One of the most

popular means for data collection is the Application Programming Interfaces (APIs) (Olmedilla et al., 2016). Therefore, the APIs for the most popular selling platforms (e.g. Amazon, eBay) and product review providers (e.g. Yelp, TripAdvisor) are made available in this phase. Enterprises are flexible and free to choose based on their capacity as long as it is sufficiently optimal to avoid irrelevant data.

Data structuring

Data analytics systems acquire data from multiple sources with diverse formats. This phase aims to obtain a consistent and standardised structure for a high volume of information that comes from multiple data sources. Although the reviews may come from different data sources, they share some similarities. For example, a review associated with a score value ranging from 1 to 5 indicates a user's preference level for a specific product. A review text represents the assessment and experience of a customer about the product, and this textual review is widely known as a new type of online Word-of-Mouth (WoM) in marketing research (Li & Zhan, 2011). Together with other metadata of online reviews, the datasets of online reviews are structured and normalised using the well-known JSON format (https://www.json.org/json-en.html), where a fundamental online review is defined as follows:

```
1  {
2  "authorID ":" o EFJ29 z AQa CNn Qzeb HQvpg ",
3  "productID ":" lDVAVEPa VSq Hz J2 a 43 _IMA "
4  "authorName ":" A. Adam Martin ",
5  "review Text ": " A solid choice on the expensive side for downtown Chicago . It is a
       great location to walk or hop a short taxi ride to the many sights in the city
   .",
6  "rating Score ": 4.0 ,
7  "summary ": "A great Chicago hotel",
8  "review Time ": "03 25 , 2019"
9  }
```

Data buffering and transmission

This phase aims to transmit data from the *data acquisition* plane to the *data analytic plane* in the cloud in real-time. Due to high velocity, the data availability could be affected by network outages from the enterprise site or the bottleneck at the cloud host site. To eliminate such data loss, a data buffering component is employed to temporarily store data when outages occur and continue sending the data once the cloud host is accessible. Additionally, to avoid the risk of unexpected errors or data loss, we maintain a table structure that matches the data transmitted and data received in the cloud plane. The data buffering and transmission mechanism enables high speed, and real-time processing facilitates mass adoption in the era of big data analysis.

Data analytics plane

Different from *data acquisition* plane, which is located locally at the enterprise site to facilitate extensibility (R3), *review analytic* plane is deployed in cloud computing and enables a pay-as-you-go payment schema. In this way, the framework achieves cost-effectiveness (R1). Before revealing the algorithms supported in the framework, we present an essential concept of marketing-specialised algorithm template.

Concept of marketing-specialised algorithm template

The primary goal of the marketing-specialised algorithm template is to design a template in which the analyst can quickly define the analysis types and the expected outcomes. The template must also consider which algorithms shall be involved in the analysis flow and how these algorithms are connected together. In addition, the template must be simple enough for non-experts with little knowledge about data analysis to utilise available algorithms to conduct their analysis. To facilitate the template formulation, we divided the analytics plane into three stages: data preparation, data analysis and data visualisation. Each stage supports various types of algorithms that can be ultilised by simply specifying the algorithm template. The logical definition of an algorithm template can be formulated as follows.

$$MSAT = Metadata + P + D + V \qquad (1)$$

where each component of the equation is detailed below.

- *MSAT* is Marketing-Specialised Algorithm Template,
- $P = \{Pi|i = the\ cardinality\ of\ the\ set\ of\ preparation\ tasks\}$,
- $A = \{Aj|j = the\ cardinality\ of\ the\ set\ analytic\ tasks\}$,
- $V = \{Ak|k = the\ cardinality\ of\ visualisation\ set\}$.

Metadata is the essential data required for specifying the analysis type and the data input. We use extensible markup language file (XML; http://www.xml.org/) to store all configurations of the algorithm template. Details of XML example shall be provide in the case study design

and implementation in section 6.2. In the subsequent sections, we present the details of algorithms supported by the framework including *data preparation* algorithms (section 5.2), *data analytic* algorithms (section 5.3), and *visualisation* algorithms (section 5.4).

Data preparation

Data preparation is a dominantly important step in a data analytics process. It allows for maintaining data quality, which is a crucial factor in the successful outcome of a data-driven framework. In this framework, the following operations are made available for the data preparation phase:

Missing values handling. Although the data is uniformly structured before transferring into the analytic plane, a considerable amount of missing value may occur, which degrades the quality of the hidden knowledge patterns. In this framework, a variety of missing value handlings are integrated and ready for adopting the analysis pipeline, including data elimination, sketching, and imputation-based methods (Roiger, 2017).

Data normalisation. The purpose of data normalisation is to transform data into the same distributions. It helps to improve the analytic results significantly as it brings equal importance and contribution to every variable during the analysis. The following normalisation methods are available in the framework:

- Rescaling: This normalisation ("min-max normalisation") transforms the mean and the standard deviation of the data to 0 and 1, respectively. More precisely, the transformation is formulated as follows.

$$z = \frac{x - \min(x)}{max(x) - min(x)} \qquad (2)$$

- Mean normalisation: This normalisation computes and subtracts the average value of every feature of the data input. The formulation is as follows.

$$z = \frac{x - \mu}{max(x) - min(x)} \qquad (3)$$

where μ is the mean of all data items in the dataset.

- Z-score normalisation: This normalisation is also known as standardisation. This method computed as the same process of mean normalisation and the standard deviation is adopted as the denominator:

$$z = \frac{x - \mu}{\sigma} \qquad (4)$$

where σ is the standard deviation of the dataset.

Table 1. Description of Fake Review Detection Methods.

Category	Method	Feature
Supervised	FRD-1 (Ott et al., 2011)	Linguitics n-gram
	FRD-2 (Jeevanandam & Kumaraswamy, 2013)	Inverse document frequency
	FRD-3 (Saumya & Singh, 2018)	Sentiments, content and rating deviation
Unsupervised	FRD-4 (Ye & Akoglu, 2015)	Group behaviours
	FRD-5 (Liu & Pang, 2018)	Sentiments and rating deviation
	FRD-6 (Kauffmann et al., 2020)	Cosine similarity measure

Outlier detection. Another important activity in data preparation is detecting and filtering outliers – abnormal values concerning the normal distribution of the data. In review data, we mainly work with numerical and textual data, and thereby, we integrate outlier detecting operations as follows:

- Numerical outlier detection: Numerical values are frequently validated to be in a specific range of values that is suitable for the study (Roiger, 2017). Therefore, to avoid a significant deviation from the expected analytic results, values far exceed or below the typical indicators are identified and filtered out by the numerical outlier detector.
- Textual outlier detection: We only consider reviews in English under the conditions of the present study. As a result, textual values that are null or correspond to different languages shall be detected and filtered out by the textual outlier detector.

The output of the data preparation phase after the outlier detection operations is 'clean data' with a homogeneous structure and not have outliers or noise data. In addition, the phase shall allow the data to be ready for analytics to find actionable knowledge, which is revealed in the next section.

Data analytics

Data analytic phase plays a central role in our proposed framework. A wide range of state-of-the-art tasks is integrated into the framework, including fake review detection, sentiment analysis, topic modelling, emotion analysis and recommender system.

Fake review detection. Fake review detection methods can be divided into two categories: *supervised* and *unsupervised* methods. The former requires a set of label data to differentiate between fake and organic reviews to train a machine learning model, while the latter does not. In our framework, we integrate six state-of-the-art methods belonging to both categories. The details of these methods are summarised in Table 1.

We integrate multiple fake review detection methods into our framework, following the 'no winner takes all theorem (Bueno et al., 2022). The theorem implies that no algorithm dominates other algorithms in all scenarios. Therefore, SMEs have multiple options for using state-of-the-art techniques, as an algorithm's selection should depend on the application. Although these integrated algorithms offer state-of-the-art performance, we acknowledge that they have some limitations. Firstly, although these algorithms could yield high accuracy, they lack the declarative information to explain their results due to deep learning techniques' inherent 'black-box' nature. Secondly, they require a relatively large amount of data to train the model; therefore, their performances are affected if we have a small dataset. Finally, their performances are sensitive to the length of distribution of fake reviews. Therefore, besides demonstrating the fake review detecting capability, we equipped our framework with additional analytic tasks that help determine multi-faceted insights and characteristics that distinguish fake reviews from organic reviews, which are revealed next.

Sentiment analysis. Sentiment analysis techniques can be separated into main categories: *(i) machine-learning-based* and *(ii) lexicon-based* approach. We integrate three state-of-the-art sentiment algorithms in our framework including two *lexicon-based* methods and one *machine-learning-based* method, which are outlined below.

1. *Vader* (Hutto & Gilbert, 2014): In Vader sentiment analysis, a ready-built dictionary is leveraged to map lexical features to the intensity level of emotions. Then, we obtain the final sentiment scores by summing up the score of each word in the document.
2. *Textblob* (Loria, 2018): Textblob is a Python library that supports many natural language processing (NLP) tasks such as sentiment analysis and opinion mining. It adopts a lexicon-based approach to building a dictionary, matches the words with polarity scores, and averages these scores to produce the final sentiment.
3. *BERT Sentiment* (Nemes & Kiss, 2021): BERT is Bidirectional Representation for Transformers, which Google AI proposed. It is prevalent in its capacity to perform different NLP tasks with state-of-the-art accuracy.

Topic modelling. Topic modelling block aims to discover the hidden semantics behind a text document. In this work, we adopt the Latent Dirichlet allocation (LDA; Edison & Carcel, 2021) model, which is a state-of-the-art technique for context-based topic modelling. In addition, we use LDAvis to interactively visualise the modelled topics in which users could explore the content of each topic.

Emotion analysis. Besides extracting hidden topics from online reviews, it is essential to identify emotions embedded in fake and organic reviews. To this end, we integrate the lexicon-based emotion analysis – the research conducted by National Research Council Canada (Mohammad & Turney, 2013) – to discover a wide range of emotion types.

Recommender system. We include in the framework a state-of-the-art collaborative filtering recommendation system (Koren & Bell, 2021) to assess how fake reviews affect business activities in terms of product ranking.

Data visualisation

The previous phase extracts hidden and actionable knowledge from data in a series of values. To provide a clearly condensed
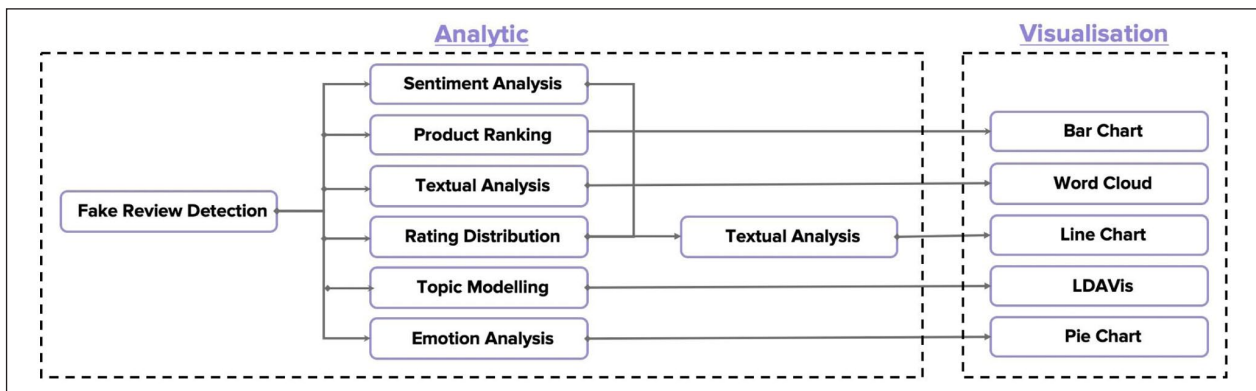
Figure 2. The analytic flow for the case study.

representation of this hidden knowledge, a variety of visualisation techniques have been provided in RAaaS to facilitate the illustration, explanation, and communication of results (Kirk, 2016). The output of visualisation techniques shall be both tables and charts. In addition, different charts such as line charts (e.g. showing time series), bar charts (e.g. comparisons among discrete measurements), pie charts (e.g. comparisons among categorical variables) and topic visualisation (e.g. a webpage interface showing topic-to-document distribution and a local view of how words are associated with each topic using LDAvis) can be selected to provide insights according to the results. The visual analytic elements are then grouped into self-service dashboard, which provide the marketing manager a quick understanding of vast volumes of information in an intuitive and real-time manner to facilitate the decision-making process.

Case study: Fraud review detection and understanding

The impact of online reviews has attracted significant attention in the last years, and it is a crucial marketing factor in determining business success. But unfortunately, the 'phenomenon of fake' is taking over marketing (Salminen et al., 2022). More precisely, only credible reviews significantly impact consumers' purchase decisions, while fake reviews are increasingly destroying the trustworthiness of online systems. For this reason, we select the topic of review fraud understanding to eliminate their effects on business activities as a case study to validate the working mechanism and efficacy of our proposed framework. Towards this goal, in this section, we first define how to set up the framework for the case study (in section 6.1). Then we design the analytic flow (in section 6.2) before specifying the algorithm template (in section 6.3) to fit the flow into our framework.

Framework setup

We use two datasets (https://www.yelp.com/dataset) in this study, that is, Yelpchi (67,393 reviews) and Yelpnyc (358,840 reviews), which are the most prevalent datasets for hotels and restaurants reviews. To simulate live platforms, we create two streams of data – each stream for each dataset – and publish APIs for the data retrieval phase. It is worth noting that although two datasets are employed in this study for demonstration purposes, the framework is quick to extend to additional data sources (if required).

Case study design and analytic flow

With the desired outcomes, this section requires defining the analytic flows and their sequential executions to achieve these goals. The process flow for the case study of fake review detection and understanding is presented in Figure 2. In the flow, we see that, in the *analytic* zone, we can define the necessary analytic steps together with their execution orders. We start with the fake review detection task to assess the detection accuracy, and then various analytic tasks are applied to understand the characteristics of fake and organic reviews. The mappings between outputs of each analytic task and a specific visualisation shall be specified in the *visualisation* zone.

The definition of the algorithm template

As outlined below, an algorithm template needs to be defined using the extensible markup language (XML) to realise the analytic flow.

The template is made available with the framework, and therefore, it is quick for non-analytic experts to modify to fit with different flow and analytic intents. In addition, the tag names associated with the template are straightforward and easy to understand. For example, users can change the data input for analytics by changing the value property (line-10) or mapping to other dataset fields and types (lines 11–17). Considering the data preparation step, users can select which preparation algorithms to apply by including or excluding a task in the task list (lines 23–27). The necessary analytic steps (lines 30–43) and visualisation steps (lines 44–52) can be specified similarly using the task list. Finally, the chronic order of executions shall be defined using a connection tag (lines 53–58). It is worth noting that we omitted some repeated tags in the template due to brevity; however, the whole template will be recovered quickly based on the sequence activities of the analytic flow (as presented in Figure 2).

Results

This section discusses our main findings. Section 7.1 and section 7.2 provide the answer for RQ1 and RQ2, respectively. The remaining sections (7.3–7.6) reveal multifaceted insights between fake and organic reviews and support the answer for RQ3.

Fake review detection

To answer for the RQ1, we evaluate the accuracy of contemporary fake review detection methods, and the results are presented in Table 2.

```xml
1  <?xml version ="1.0" encoding="utf-8"?>
2  <template name ="fake_review_analytic" version="1.0">
3      <metadata>
4          <creator>the_authors </ creator>
5          <property name="Conn String" value="review_transmission_1"/>
6      </ metadata>
7      <!--data input step-->
8      <step name="Input">
9          <process id="data_input1" name="data_input1"type="true">
10             <property name ="DatasetName"value="Yelp"/>
11             <Field List>
12                 <Data Field name = "authorID" data Type="string" index="0"/>
13                 <Data Field name = "productID" data Type="string" index="1"/>
14                 <Data Field name = "authorName" data Type="string" index="2"/>
15                 <Data Field name = "review Text" data Type="string" index="3"/>
16                 <Data Field name = "rating Score" data Type="real" index="4"/>
17                 <Data Field name = "review Time" data Type="datetime" index="5"/>
18             </Field List>
19         </process>
20     </ step>
21     <!--data preparation step-->
22     <step name = "Data Preparation">
23         <process id = "data_preparation" name = "Data Preparation">
24             <TaskList>
25                 <Taskname ="Missing Value" action = "imputation"/>
26                 <Taskname ="Normalisation" action = "rescaling"/>
27                 <Taskname ="OutlierDetection" action = "true"/>
28             </Task List>
29         </process>
30     </step>
31     <!--data analytic step-->
32     <step name = "Data Analytic">
33         <process id = "fake_review_detection" name = "Fake Review Detection">
34             <TaskList>
35                 <Taskname ="Fake Review Detection" action = "supervised"/>
36             </TaskList>
37         </process>
38         <process id = "sentiment_analysis" name = "Sentiment Analysis">
39             <Task List>
40                 <Taskname ="SentimentAnalysis" action ="vader" />
41             </TaskList>
42         </process>
43     <!--...more analytic processes are shorten due to brevity -->
44     </step>
45     <!--data visualisation step -->
46     <step name = "Data Visualisation">
47         <process id = "barchart" name = "Bar Chart Visualisation">
48             <TaskList>
49                 <Task name = "BarChart" action = "horizontal"/>
50             </TaskList>
51         </process>
52     <!--...more visualisation processes are shorten due to brevity -->
53     </step>
54     <!--processes integrating -->
55     <connection from = "data_input1" to = "data_preparation"/>
56     <connection from = "data_preparation" to = "fake_review_detection"/>
57     <connection from = "fake_review_detection" to = "sentiment_analysis"/>
58     <connection from = "sentiment_analysis" to = "barchart_visualisation"/>
59     <!--...more connections are shorten due to brevity -->
60 </template>
```

Overall, we see that FRD-3 achieves much higher accuracy compared to other methods over all datasets. These results are consistent with prior works (Kauffmann et al., 2020), and this improvement originates from the combinations of detecting features, including sentiments, content and rating deviation. In addition, the performance of supervised methods is better than unsupervised methods in general. The difference in performance might result from the feature sets that each method uses in its implementation. The subsequent sections dive into the discrimination of each feature which provides a better understanding of the accuracy of detection methods.

Table 2. Performance of Fake Review Detection Methods.

Category	Method	Accuracy (Yelpchi) (%)	Accuracy (Yelpnyc) (%)
Supervised	FRD-1 (Ott et al., 2011)	87.6	85.3
	FRD-2 (Jeevanandam & Kumaraswamy, 2013)	74.9	76.7
	FRD-3 (Saumya & Singh, 2018)	91.6	90.3
Unsupervised	FRD-4 (Ye & Akoglu, 2015)	81.4	82.3
	FRD-5 (Liu & Pang, 2018)	79.7	78.2
	FRD-6 (Kauffmann et al., 2020)	84.8	86.5
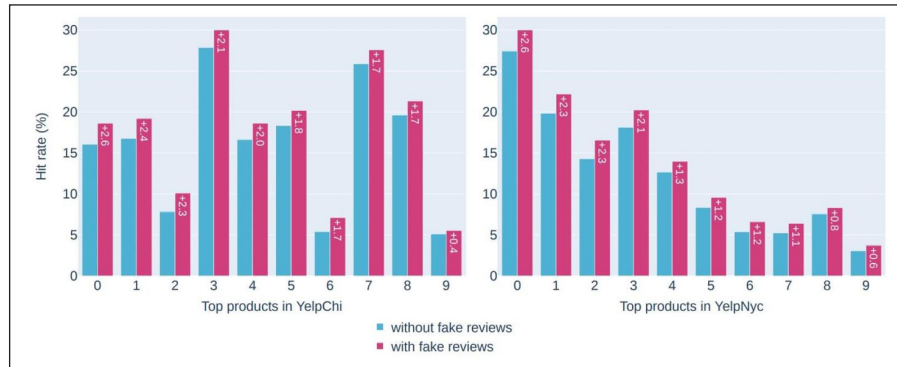
Figure 3. Exposure rate.
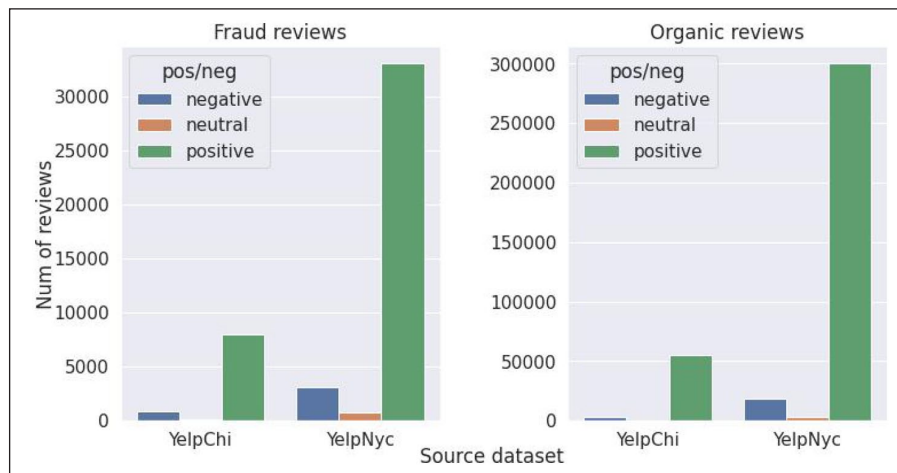
Figure 4. Sentiment analysis.

Exposure rate

This experiment aims to answer for RQ2 of how fake reviews affect business activities. To this end, we employ a recommender system and compare how product ranking changes with and without the existence of fake reviews in the online system. In line with previous research such as Kauffmann et al. (2020), it is revealed that the ranking of the top-10 products with the most reviews is shown in Figure 3. The key finding is that fake reviews can successfully promote a target item to a hit rate (i.e. visibility score) of up to 2.6%. In addition, there is a consistent increase in the product ranking with the existence of fake reviews in the system. The result demonstrates significant adverse effects on the reliability and trustworthiness of business providers.

Sentiment analysis

In this experiment, we apply sentiment analysis to understand the sentiment characteristic between fake and organic reviews, and the results are presented in Figure 4. The sentiment distributions between fraud and organic reviews are different but not significant. The reason could be that review texts recently have been generated by an AI-powered tool (Kreps et al., 2022) that mimics the actual data distribution rather than being written by a human. That is the reason why sentiment features are not sufficient for fraud review detection, and state-of-the-art approaches combine with other features such as rating distribution (Liu & Pang, 2018) or textual features (Saumya & Singh, 2018) to achieve competitive performance.
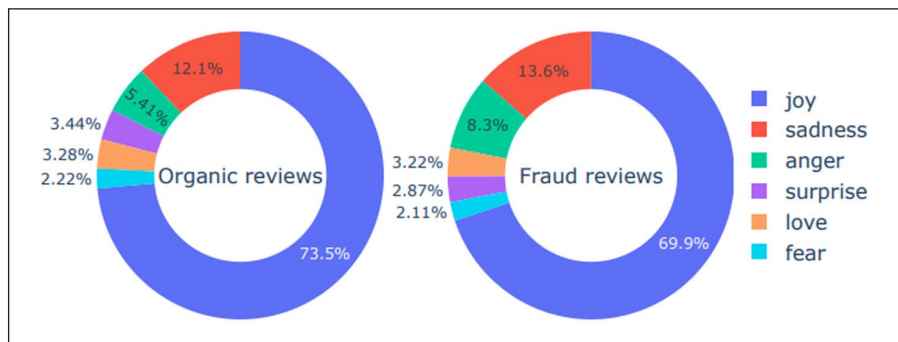
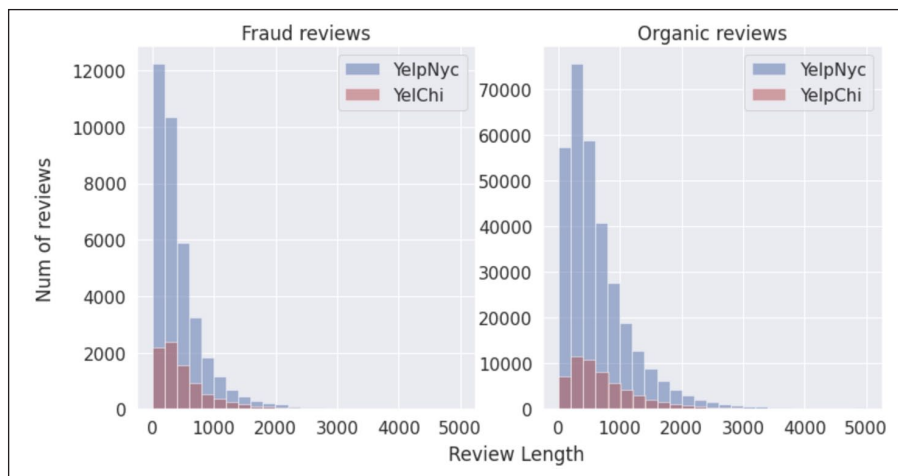**Figure 5.** Emotion analysis (Yelpchi).

Figure 6. Textual analysis.

We further compare and contrast fake and organic reviews on six discrete emotions (i.e. joy, sadness, anger, surprise, love, fear), and Figure 5 shows the results.

Joy is the dominant positive emotion in fake and organic reviews. The second and third dominant in common is negative emotions – sadness and anger. Interestingly, the concentration of all two types of negative emotions is higher in fake reviews than in organic reviews. Our findings of sadness and anger are consistent with prior works as fraudsters adopt fake reviews to depromote competitors' products and services (Luca & Zervas, 2016; Munzel, 2016). Different from previous research, which predominantly focuses on the level of positivity and negativity and overlooks the presence of specific emotions (Kauffmann et al., 2020; Luca & Zervas, 2016), our study shows that the detection of such discriminative features is essential for future explorations about developing fake review detection. Finally, we found that surprise, love, and fear are the least common emotions and have no significant concentration difference between the two types of reviews.

Textual analysis

In this section, we aim to understand the content characteristics of fake and organic reviews. To this end, computerised textual analysis is performed, and the results are presented in Figure 6. An interesting observation is that fake reviews tend to be shorter than organic reviews. Despite being shorter in length, prior research (Bilgihan et al., 2016) indicates that fake reviews that use complex language to demonstrate a high level of expertise may be perceived as credible. We will verify this finding in section 7.5 when we analyse text at the word level. Due to the difference in length, fraud reviews' distribution is more diverse than organic reviews. This finding supports why the method that employed content as detecting features achieves higher detection accuracy (Saumya & Singh, 2018).

As the framework enables the capability of a hierarchy analysis, here we present the result of textual analysis in a sequential order – rating distribution first and then performing the textual analysis. The results are illustrated in Figure 7. We see a decrease in the review view length when the rating score increases. However, the length of the fraud review (versus organic reviews) is consistent with the previous findings.

Word cloud/Qualitative analysis

This experiment provides a closer look into the review content – the word level. The word cloud analysis between fake and organic reviews is presented in Figure 8 for the Yelpchi dataset (findings on Yelpnyc are discarded as they reveal the same characteristics). An interesting finding is that although both reviews employ favourable terms such as 'good' and 'well' to express a positive comment on goods or services, fraud reviews tend to use strong adverbs such as 'really' and 'great' to emphasise the emotions. Future lexical-based fake review detections could take advantage of such discriminative features to improve the detecting accuracy.
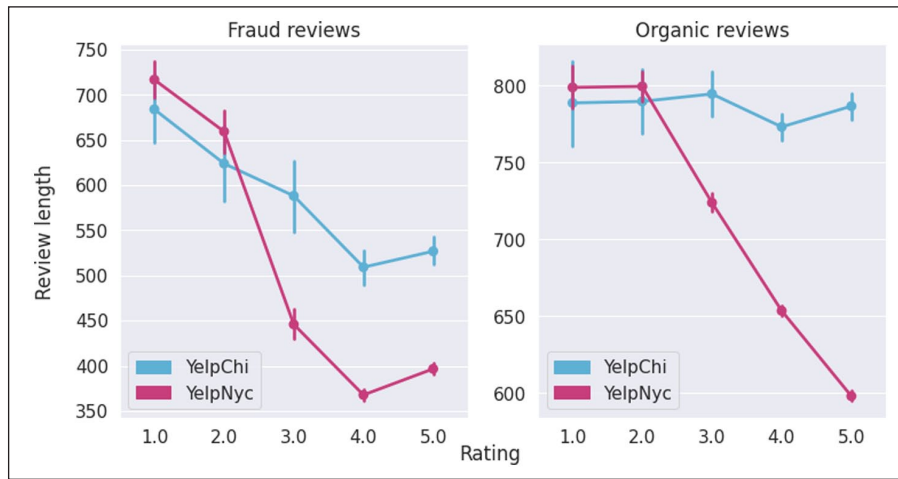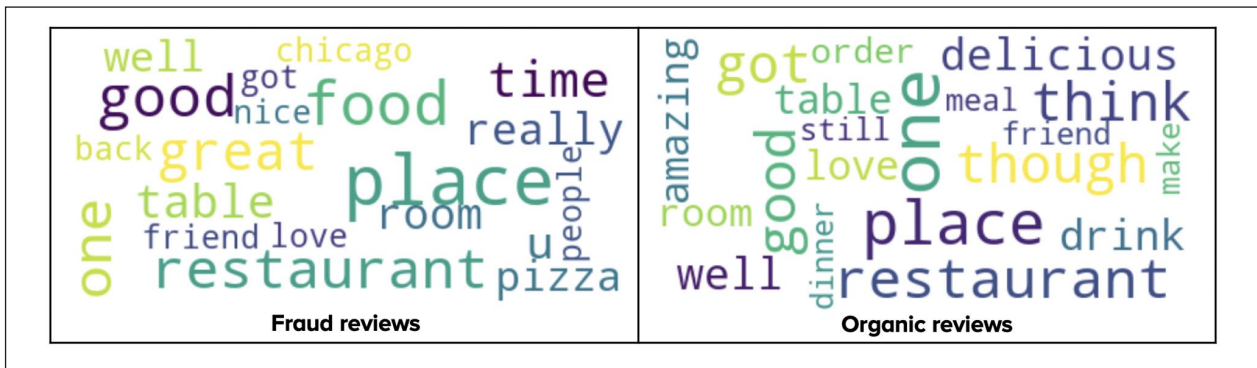
Figure 7. Hierarchy textual analysis.

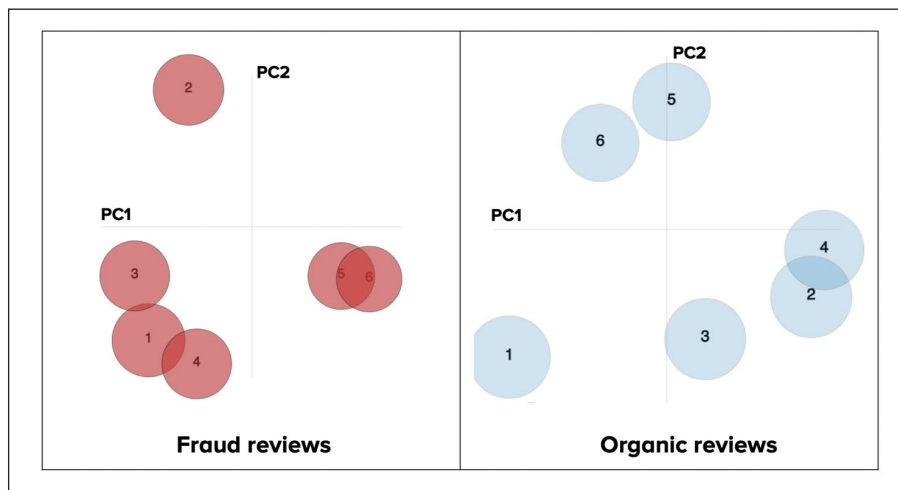**Figure 8.** Wordcloud analysis.

Figure 9. Inter-topic distance map of topic (Yelpchi).

Topic modelling

To investigate the difference between fake and organic reviews in terms of hidden topics, we extract the six most dominant topics in online reviews, scaling them into two-dimensional space where PC1 and PC2 are the two principal components (Abdi & Williams,

2010) of the topics, and visualise them using the LDAVis chart in Figure 9.

The results reveal an inter-topic distance map of topics on the Yelpchi dataset (the result of the remaining dataset is omitted as it exposes the same trend). Each circle represents each topic, and circle size denotes the popularity of those topics in all review

content. The distance between two circles represents the similarities of two hidden topics, and a closer distance suggests a more substantial level of similarity. The critical observation is that the spatial distribution of most popular topics reveals a high contrast between fraud and organic reviews. For example, the most similar topics in fraud reviews are topics 5 and 6, while the most similar topics in organic reviews are topics 2 and 4. Future works on fake review detection and understanding could consider this discrimination in terms of topic spatial distributions to improve the detection capability.

Discussion

Online reviews are an important source of information that can influence customer attitudes and purchase behaviour (Salminen et al., 2022). The reason is that prospective customers like to learn from former customers' experiences before deciding to buy a product (Robertson et al., 2021). Due to the advent of the Internet and the openness of online review platforms (Schindler & Decker, 2013), customers are increasingly sharing their experiences through online reviews (van Esch et al., 2018). Because most customers consult online reviews before engaging with a brand, enterprises must understand their customer's experience to make any necessary changes. On the other hand, fake reviews have become a significant concern in eroding the credibility of online review systems (Luca & Zervas, 2016). Fake reviews send the wrong message to legitimate enterprises and customers, which can distort the ranking system, resulting in negative consequences for genuine organisations, which has been pointed out in prior literature (He et al., 2022). Therefore, it is crucial for enterprises to understand what authentic customers say about their experiences, and how fake reviews could harm a brand's reputation.

Studies on fake detections have gained significant attention in recent years to prevent the adverse effects of fake reviews on society (Mohawesh et al., 2021). The motivation behind these approaches is to eliminate the societal impact caused by fake reviews on the fair-trade environment. Because the significance of fake reviews grows, the prevalence of fake reviews written to appear authentic and informative is increasing, weakening consumers' trust (Moon et al., 2021) in online reviews. Fraudulent reviews do not necessarily reflect the genuine opinion of consumers because they create a wrong impression about products or services. This causes consumers to make poor or incorrect decisions. The dissemination of misleading information has a detrimental effect on social behaviour (Javed et al., 2021). In recent years, many fake review detections have been proposed (Mohawesh et al., 2021) to prevent the negative impacts and restore the consumers' trust.

However, deriving customer insights from online reviews and detecting false reviews required a complicated procedure that goes beyond the capability of SMEs (Kauffmann et al., 2020). To fill the gap, we design a framework that SME marketers can use for detecting and understanding fake reviews. To facilitate online review data analytics, the framework achieves cost-effectiveness by enabling the pay-as-you-go analytic schema (Hung et al., 2019; Nguyen et al., 2020) and can be handled by non-data specialists. Using a case study, we answer the research questions relating to the differences between fake and organic reviews and investigate their differences in multi-perspectives. Using six major summary dimensions (i.e. exposure rate, sentiment, emotion, textual characteristics, word cloud and topic modelling), the analytic results suggest significant differences between fake and organic reviews. Regarding detecting fake reviews, we found that modern detection techniques can identify fake reviews

with high accuracy, while the performance of supervised methods is generally superior to that of unsupervised methods. The key finding regarding *exposure rate* is that fake reviews can effectively promote a target item to a hit rate (i.e. visibility score) of up to 2.6%. Not surprisingly, there is a consistent increase in the product ranking with the existence of fake reviews in the system, which is in line with previous research (Kauffmann et al., 2020).

The substance of the content of these online reviews is of tremendous relevance. Using computerised textual analysis, key differences between fake and organic reviews were found. For *sentiment analysis*, we discovered that the distributions of sentiment between fraudulent and organic reviews are distinct but not statistically significant. The reason could be that recent reviews were generated by an AI-powered technology that matches the actual data distribution (Bakpayev et al., 2022; Feng et al., 2021). For *emotional analysis*, by extending prior research that primarily focuses on the level of positive and negative sentiment (Kauffmann et al., 2020; Luca & Zervas, 2016), we put attention on the existence of specific emotions in the texts and highlight the higher concentration of sadness and anger in fake reviews compared to organic reviews. As seen in the *textual analysis*, it is also found that fake reviews tend to be shorter than organic reviews. Despite being shorter in length, previous work (Bilgihan et al., 2016) suggests that fake reviews employ complicated language to demonstrate a high degree of knowledge that may be viewed as credible. To verify this argument, we further look into the details of *word cloud*, and the finding is consistent – fake reviews tend to make use of strong adverbs. Finally, regarding the *topic modelings*, fake reviews have different patterns in terms of spatial distribution and correlation among the hidden topics. This finding is in line with recent works (Birim et al., 2022; W. Zhang et al., 2022) that employ topic distributions as features for fake review detection.

Research implications

Our study's primary research question is how to harness the critical values of big data and create a proactive platform for non-data specialists to initiate massive review analytics jobs. Besides, we aims to response to additional research questions: *Can review fraud be automatically detected? How do review frauds affect the product ranking system? What are the differences between fake and organic reviews?* The analysis outcomes from computerised textual analytic of more than 400,000 online reviews reveal that fake reviews can be detected with high accuracy by contemporary techniques, the product ranking with the existence of fake review can be biased, and that fake views and organic reviews differ substantially. While we explored the findings of the study in the previous section, we here discuss the implications for marketing theory and practise.

Theoretical implications

As this research aims to propose a cost-effective cloud-based framework for online review analytics, the results of this study will enhance the cloud computing theory (Marinescu, 2022) with practice in marketing. In addition, we demonstrated our framework in a case study on fake review detection and understanding, offering additional important implications to marketing theory. First, marketing-specialised data analytic frameworks have been widely adopted in achieving a variety of marketing goals, including optimal budget allocation (Yang et al., 2022), maintaining corporate reputation (Rantanen et al., 2019), maximising sales revenue and profit (Quach et al., 2022). These optimisations are based on the positive and authentic sides of data (e.g. how to create value from

customers' positive feedback). However, we argue that exploring the other side of the coin – also known as the 'phenomenon of fake' – raises an equally important concern. Rather than maximising a marketing goal, a critical research question has been raised: *how to minimise the value destruction caused by contemporary technology capabilities?* Our study is situated in this context, exploring a specific case of review fraud understanding and alarming the negative impacts on business activities. Extending research on review fraud (Choi et al., 2017; Lee et al., 2022; Salminen et al., 2022), our study is among the first to propose a low-cost, automatic cloud-based framework for reviewing fraud detection and understanding at scale in marketing, which is crucial to scalable information processing theory (Z. Wang et al., 2015). Second, by looking beyond fundamental aspects such as positive versus negative sentiment (Kauffmann et al., 2020; Luca & Zervas, 2016), our findings provide important insights into the impact of review fraud and the characteristics of fake reviews, thereby laying the foundation for developing countermeasures against such fraudulent activities. Although we are not the first to address this concern as it has been highlighted and investigated in recent works (He et al., 2022), we go beyond existing literature by providing a deep understanding of fake reviews with multi-faceted insights. Finally, considering the framework's extensibility, it can be adopted as a basic framework for future analytic tasks or even a future domain such as social feedback (McIntyre et al., 2016), customer engagement (de Ruyter et al., 2018; Ng et al., 2020), and recommender system (Koren & Bell, 2021; Toan et al., 2018). The proposed framework can act as a trigger for forthcoming studies, therefore promising to enhance knowledge in machine learning and AI theory in marketing (Ma & Sun, 2020), specialised for fake review detections.

Practical implications

Although this study's goal is to benefit SMEs in applying review analytics to monitor their brand reputation and customer experience, this study also benefits different stakeholder groups. In this section, we outline substantial practical implications for these stakeholders that are the potential to utilise RAaaS as a core.

SMEs. From the study, businesses, especially SMEs who are constrained by limited financial and technological resources, are encouraged to adopt this framework to monitor their customer's online reviews on a regular basis. This practice is essential as review fraud may quickly escalate into a crisis for a brand's reputation without proper actions (He et al., 2022). Additionally, by understanding the importance of benign reviews and their characteristics, businesses should develop a rewarding programme for their customers willing to contribute their benign reviews to the system. That way, customers shall be encouraged to submit online reviews to increase their loyalty and motivate satisfaction. Businesses should adopt the framework to be transparent about who leaves online reviews, so future customers can see who wrote what – based thereon, the transparency in the processing of online reviews contributes to an essential characteristic for building customer trust.

Review platforms. Reviews platforms are the central stadium that connects buyers and sellers through online reviews. This framework can benefit review platforms in improving the timeliness of reacting to fake behaviours, as eliminating review fraud is central to maintaining transparency and creditability. The reason is that review platforms are in the best position to fix the phenomenon of fake in the online review ecosystem, and they can leverage the framework as a tool to tackle it. The framework can be used to filter out suspicious reviews and suspend or delist companies that exploit fake reviews to manipulate public trust. They can enhance the efficiency of the framework in multiple aspects. For example, platforms should provide simple, effective, and meaningful tools for individuals and legitimate businesses to report fraudulent reviews. Additionally, they can improve the fake review detection technology by hiring more people to label wrong actions and fraud patterns to produce more quality data for training the algorithm. Finally, they can provide more access to the outside research community to explore and enhance the capability of detecting emerging fake patterns to restore the customer's trust.

Customers and society. Fraudulent reviews are characterised as misbehaving customer action that violates socially acceptable norms of conduct (Sigala, 2017). These actions have been assessed as potentially generating psychological or physical harm to society. A recent study (Sigala, 2017) emphasises the potential adverse consequences of untruthful reviews on SMEs and their societal impact situation. The widespread utilisation of cloud-based frameworks for fraudulent reviews monitoring will assist in consumer protection. Additionally, the characteristics and a deep understanding of fake reviews can be used to educate customers about the existence of review fraud (Anderl et al., 2016) and its influence on their purchasing decisions. Thus, consumers' trust will be re-gained, and they will be more likely to utilise these platforms or businesses more consistently.

Conclusion, limitations and future research directions

In this paper, we propose RAaaS – a Review-Analytic-as-a-service framework that specialises in big data analysis and understanding reviews, to facilitate the marketing decision-making process for but not limited to SMEs, review platforms and customers. With the concept of marketing-specialised algorithm template, the framework enables non-data specialists to involve in a complex data analytics process to derive customer experiences from online reviews to facilitate brand engagement. Although the present study provides crucial implications, we acknowledge several limitations that can be considered in future works. The primary limitation of our work is the absence of online reviews written in other languages rather than English. It would be crucial for future research to incorporate our findings in the multi-lingual analysis. Another limitation includes the lack of understanding of review fraud under the propagation time perspective and the correlations between propagating time series. Motivated by the recent success in employing time burst characteristics for fake review detection (N. Wang et al., 2022), we intend to combine multivariate time series representation with graph neural networks (Wu et al., 2020) to perform time-series analysis to discover additional insights in a future study.

Our study also calls for promoting interdisciplinary research to reduce the spread of fake reviews and to address underlying damages they have uncovered. One important finding in our work is that fake reviews can successfully promote a target item to a hit rate (i.e. visibility score) of up to 2.6%. Given the potential impact, it may be prudent for digital marketers to respond first to reviews that are suspected of being fraudulent. Businesses are strongly encouraged to scrutinise and be on top of how customers express their experiences and proactively respond to the case when fake reviews are discovered.

Author's Note

Vinh Khiem Tran is also affiliated to Vietnam National University, Ho Chi Minh City, Vietnam.

Sara Quach and Park Thaichon is now affiliated to Griffith University, Brisbane, QLD, Australia.

Bay Vo is now affiliated to HUTECH University, Ho Chi Minh City, Vietnam.

ORCID iDs

Vinh Khiem Tran [iD] https://orcid.org/0000-0001-7128-2058

Sara Quach [iD] https://orcid.org/0000-0002-0976-5179

Park Thaichon [iD] https://orcid.org/0000-0001-7512-7362

References

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Anderl, E., März, A., & Schumann, J. H. (2016). Nonmonetary customer value contributions in free e-services. *Journal of Strategic Marketing*, *24*(3–4), 175–189.

Bakpayev, M., Baek, T. H., van Esch, P., & Yoon, S. (2022). Programmatic creative: Ai can think but it cannot feel. *Australasian Marketing Journal*, *30*(1), 90–95.

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.

Bastwadkar, M., McGregor, C., & Balaji, S. (2020). *A cloud based big data health-analytics-as-a-service framework to support low resource setting neonatal intensive care unit* [Conference session]. *Proceedings of the 4th international conference on medical and health informatics*, pp 30–36.

Bilgihan, A., Kandampully, J., & Zhang, T. (2016). Towards a unified customer experience in online shopping environments: Antecedents and outcomes. *International Journal of Quality and Service Sciences*, *8*, 102–119.

Birim, Ş. Ö., Kazancoglu, I., Kumar Mangla, S., Kahraman, A., Kumar, S., & Kazancoglu, Y. (2022). Detecting fake reviews through topic modelling. *Journal of Business Research*, *149*, 884–900.

Bueno, I., Carrasco, R. A., Ureña, R., & Herrera-Viedma, E. (2022). A business context aware decision-making approach for selecting the most appropriate sentiment analysis technique in e-marketing situations. *Information Sciences*, *589*, 300–320.

Chen, H. M., Chang, K. C., & Lin, T. H. (2016). A cloud-based system framework for performing online viewing, storage, and analysis on big data of massive BIMs. *Automation in Construction*, *71*, 34–48.

Choi, S., Mattila, A. S., Van Hoof, H. B., & Quadri-Felitti, D. (2017). The role of power and incentives in inducing fake reviews in the tourism industry. *Journal of Travel Research*, *56*(8), 975–987.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort-Martorell, X., & Reis, M. S. (2016). How can SMEs benefit from big data? Challenges and a path forward. *Quality and Reliability Engineering International*, *32*(6), 2151–2164.

De Luca, L. M., Herhausen, D., Troilo, G., & Rossi, A. (2021). How and when do big data investments pay off? The role of marketing affordances and service innovation. *Journal of the Academy of Marketing Science*, *49*(4), 790–810.

Del Vecchio, P., Di Minin, A., Petruzzelli, A. M., Panniello, U., & Pirri, S. (2018). Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. *Creativity and Innovation Management*, *27*(1), 6–22.

de Ruyter, K., Isobel Keeling, D., & Ngo, L. V. (2018). When nothing is what it seems: A digital marketing research agenda. *Australasian Marketing Journal (AMJ)*, *26*(3), 199–203.

de Vries, L., Gensler, S., & Leeflang, P. S. H. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of Interactive Marketing*, *26*(2), 83–91.

Edison, H., & Carcel, H. (2021). Text data analysis using Latent Dirichlet allocation: An application to FOMC transcripts. *Applied Economics Letters*, *28*(1), 38–42.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big data consumer analytics and the transformation of marketing. *Journal of Business Research*, *69*(2), 897–904.

Feng, C. M., Park, A., Pitt, L., Kietzmann, J., & Northey, G. (2021). Artificial intelligence in marketing: A bibliographic perspective. *Australasian Marketing Journal*, *29*(3), 252–263.

Ha, S. H., Bae, S. Y., & Son, L. K. (2015). Impact of online consumer reviews on product sales: Quantitative analysis of the source effect. *Applied Mathematics and Information Sciences*, *9*(2L), 373–387.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science*, *41*, 896–921.

Hung, N. Q. V., Weidlich, M., Tam, N. T., Miklós, Z., Aberer, K., Gal, A., & Stantic, B. (2019). Handling probabilistic integrity constraints in pay-as-you-go reconciliation of data models. *Information Systems*, *83*, 166–180.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*, 216–225.

Javed, M. S., Majeed, H., Mujtaba, H., & Beg, M. O. (2021). Fake reviews classification using deep learning ensemble of shallow convolutions. *Journal of Computational Social Science*, *4*(2), 883–902.

Jeevanandam, D. R., & Kumaraswamy, Y. S. (2013). Opinion mining using decision tree based feature selection through manhattan hierarchical cluster measure. *Journal of Theoretical and Applied Information Technology*, *58*(1), 72–80.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management*, *44*, 1–12.

Kasemsap, K. (2021). Software as a service, semantic web, and big data: Theories and applications. In *Research Anthology on Recent Trends, Tools, and Implications of Computer Programming* (pp 1179–1201). IGI Global.

Kauffmann, E., Peral, J., Gil, D., Ferrández, A., Sellers, R., & Mora, H. (2020). A framework for big data analytics in commercial social networks: A case study on sentiment analysis and fake review detection for marketing decision-making. *Industrial Marketing Management*, *90*, 523–537.

Khan, Z., Anjum, A., Soomro, K., & Tahir, M. A. (2015). Towards cloud based big data analytics for smart future cities. *Journal of Cloud Computing*, *4*(1), 1–11.

Kirk, A. (2016). *Data visualisation: A handbook for data driven design*. SAGE.

Koren, Y., & Bell, R. (2021). Advances in collaborative filtering. In *Recommender systems handbook* (pp. 91–142). Springer.

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117.

Lee, M., Song, Y. H., Li, L., Lee, K. Y., & Yang, S. B. (2022). Detecting fake reviews with supervised machine learning algorithms. *Service Industries Journal*, *42*, 1101–1121.

Li, J., & Zhan, L. (2011). Online persuasion: How the written word drives WOM: Evidence from consumer-generated product reviews. *Journal of Advertising Research*, *51*(1), 239–257.

Liu, Y., & Pang, B. (2018). A unified framework for detecting author spamicity by modeling review deviation. *Expert Systems with Applications*, *112*, 148–155.

Liu, Y., Soroka, A., Han, L., Jian, J., & Tang, M. (2020). Cloud-based big data analytics for customer insight-driven design innovation in SMEs. *International Journal of Information Management*, *51*, 102034.

Loria, S. (2018). Textblob documentation. *Release 0.15*, 2.

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, *62*(12), 3412–3427.

Ma, L., & Sun, B. (2020). Machine learning and ai in marketing – Connecting computing power to human insights. *International Research Journal of Marketing*, *37*(3), 481–504.

Malbon, J. (2013). Taking fake online consumer reviews seriously. *Journal of Consumer Policy*, *36*(2), 139–157.

Manaskasemsak, B., Tantisuwankul, J., & Rungsawang, A. (2023). Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. *Neural Computing and Applications*, *35*, 1169–1182.

Marinescu, D. (2022). *Cloud computing: Theory and practice*. Morgan Kaufmann.

Martens, D., & Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*, *24*(6), 3316–3355.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 60–68.

McIntyre, S. H., McQuarrie, E. F., & Shanmugam, R. (2016). How online reviews create social network value: The role of feedback versus individual motivation. *Journal of Strategic Marketing*, *24*(3-4), 295–310.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake reviews detection: A survey. *IEEE Access*, *9*, 65771–65802.

Moon, S., Kim, M. Y., & Iacobucci, D. (2021). Content analysis of fake consumer reviews by survey-based text categorization. *International Research Journal of Marketing*, *38*(2), 343–364.

Munzel, A. (2016). Assisting consumers in detecting fake reviews: The role of identity information disclosure and consensus. *Journal of Retailing and Consumer Services*, *32*, 96–108.

Nemes, L., & Kiss, A. (2021). Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, *5*(3), 375–394.

Ng, S. C., Sweeney, J. C., & Plewa, C. (2020). Customer engagement: A systematic review and future research priorities. *Australasian Marketing Journal*, *28*(4), 235–252.

Nguyen, T. T., Weidlich, M., Yin, H., Zheng, B., Nguyen, Q. H., & Nguyen, Q. V. H. (2020). *Factcatch: Incremental pay-as-you-go fact checking with minimal user effort* [Conference session]. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2165–2168. ACM.

Olmedilla, M., Martínez-Torres, M. R., & Toral, S. L. (2016). Harvesting big data in social science: A methodological approach for collecting online user-generated content. *Computer Standards & Interfaces*, *46*, 79–87.

Ott, M., Yejin, C., Claire, C., & Hancock, J. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Quach, S., Thaichon, P., Martin, K. D., Weaven, S., & Palmatier, R. W. (2022). Digital technologies: Tensions in privacy and data. *Journal of the Academy of Marketing Science*, *50*, 1299–1323.

Rantanen, A., Salminen, J., Ginter, F., & Jansen, B. J. (2019). Classifying online corporate reputation with machine learning: A study in the banking domain. *Internet Research*, *30*, 45–66.

Robertson, J., Ferreira, C., & Paschen, J. (2021). Reading between the lines: Understanding customer experience with disruptive technology through online reviews. *Australasian Marketing Journal*, *29*(3), 215–224.

Roiger, R. J. (2017). *Data mining: A tutorial-based primer*. Chapman and Hall/CRC.

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.

Saumya, S., & Singh, J. P. (2018). Detection of spam reviews: A sentiment analysis approach. *CSI Transactions on ICT*, *6*(2), 137–148.

Schindler, D., & Decker, R. (2013). Some remarks on the internal consistency of online consumer reviews. *Australasian Marketing Journal (AMJ)*, *21*(4), 221–227.

Schmidt, R., & Möhring, M. (2013). Strategic alignment of cloud-based architectures for big data. In *2013 17th IEEE International Enterprise Distributed Object Computing Conference Workshops* (pp. 136–143). IEEE.

Sigala, M. (2017). How "bad" are you? justification and normalisation of online deviant customer behaviour. In *Information and communication technologies in tourism 2017* (pp 607–622). Springer.

Sultan, N. A. (2011). Reaching for the "cloud": How SMES can manage. *International Journal of Information Management*, *31*(3), 272–278.

Toan, N. T., Cong, P. T., Tam, N. T., Hung, N. Q. V., & Stantic, B. (2018). Diversifying group recommendation. *IEEE Access*, *6*, 17776–17786.

van Esch, P., Arli, D., Castner, J., Talukdar, N., & Northey, G. (2018). Consumer attitudes towards bloggers and paid blog advertisements: what's new? *Marketing Intelligence & Planning*, *36*, 778–793.

Wang, N., Yang, J., Kong, X., & Gao, Y. (2022). A fake review identification framework considering the suspicion degree of reviews with time burst characteristics. *Expert Systems with Applications*, *190*, 116207.

Wang, X., Zhang, X., Jiang, C., & Liu, H. (2018). *Identification of fake reviews using semantic and behavioral features* [Conference session]. 2018 *4th International Conference on Information Management (ICIM)* (pp. 92–97). IEEE.

Wang, Z., Hu, R., Chen, Q., Gao, P., & Xu, X. (2020). Collueagle: Collusive review spammer detection using markov random fields. *Data Mining and Knowledge Discovery*, *34*(6), 1621–1641.

Wang, Z., Tingting, H., Zhun, L., & Song, D. (2015). Spotting fake reviewers using product review graph. *Journal of Computational Information Systems*, *16*(11), 5759–5767.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., & Zhang, C. (2020). *Connecting the dots: Multivariate time series forecasting with graph neural networks* [Conference session]. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 753–763). ACM.

Yang, Y., Feng, B., Salminen, J., & Jansen, B. J. (2022). Optimal advertising for a generalized Vidale–Wolfe response model. *Electronic Commerce Research*, *22*, 1275–1305.

Ye, J., & Akoglu, L. (2015). *Discovering opinion spammer groups by network footprints* [Conference session]. *Joint European conference on machine learning and knowledge discovery in databases* (pp. 267–282). Springer.

Zhang, F., Hao, X., Chao, J., & Yuan, S. (2020). Label propagation-based approach for detecting review spammer groups on e-commerce websites. *Knowledge-Based Systems*, *193*, 105520.

Zhang, W., Xie, R., Wang, Q., Yang, Y., & Li, J. (2022). A novel approach for fraudulent reviewer detection based on weighted topic modelling and nearest neighbors with asymmetric Kullback–Leibler divergence. *Decision Support Systems*, *157*, 113765.

Zulkernine, F., Martin, P., Zou, Y., Bauer, M., Gwadry-Sridhar, F., & Aboul-Naga, A. (2013). *Towards cloud-based analytics-as-a-service (CLAaaS) for big data analytics in the cloud* [Conference session]. 2013 *IEEE international congress on Big Data* (pp. 62–69). IEEE.