Journal of Marketing Analytics
https://doi.org/10.1057/s41270-025-00393-8

ORIGINAL ARTICLE

Decoding deception in the online marketplace: enhancing fake review detection with psycholinguistics and transformer models

Joni Salminen[1] · Mekhail Mustak[2] · Soon-Gyo Jung[3] · Hannu Makkonen[1] · Bernard J. Jansen[3]

Revised: 21 January 2025 / Accepted: 28 February 2025

Abstract

Online reviews significantly influence consumer decision-making in digital marketplaces, yet the proliferation of fake reviews threatens their credibility. This study investigates the psycholinguistic features that differentiate human-written fake reviews from genuine ones and explores how these features, along with distributional semantics, can be leveraged for automatic detection. Using a dataset of 3070 reviews from 307 participants, we analyze linguistic patterns with the Linguistic Inquiry and Word Count tool and train machine learning classifiers to predict review authenticity. Our findings reveal distinct psycholinguistic markers in fake reviews, including heightened cognitive processes and emotional exaggeration, and demonstrate the superior performance of transformer-based models like BERT in fake review detection. This research contributes theoretically by linking psycholinguistic cues with advanced natural language processing techniques and offers practical insights for improving review monitoring systems.

Keywords Fake review · Detection · Psycholinguistics · Natural language processing · Machine learning · Semantics

Introduction

Online reviews play a crucial role in the digital marketplace, significantly influencing consumer decision-making (Ordabayeva et al. 2022; Schoenmueller et al. 2020). Platforms such as Amazon, Yelp, and TripAdvisor provide consumers with valuable insights into product quality and user satisfaction, shaping purchasing behaviors (Choi et al. 2019, 2022). Beyond aiding consumers, online reviews impact businesses by enhancing product visibility and credibility, ultimately driving sales (Petrescu and Krishen 2024; Sahoo et al. 2018). However, the reliability of online reviews is increasingly undermined by fake reviews, which mislead consumers and distort marketplace dynamics (Ananthakrishnan et al. 2020; Moon et al. 2021).

The influence of online reviews is well-documented. Recent studies indicate that 93% of consumers rely on them for purchase decisions, with 91% trusting them as much as personal recommendations (Schoenmueller et al. 2020). Conversely, negative reviews can deter potential buyers and damage brand reputation (Liu et al. 2017; Wang et al. 2024). Notably, 86% of consumers report that negative reviews strongly affect their choices, and 67% would refrain from purchasing a product or service after encountering just one to three negative reviews (Ordabayeva et al. 2022). These findings underscore the importance of online reputation management and the urgent need for mechanisms to detect fraudulent content (Choi et al. 2019; Schoenmueller et al. 2020; Wang, Zhang, and Xu 2024).

Fake reviews—fabricated endorsements or criticisms—pose a serious threat to consumer trust and market fairness (Choi et al. 2022; Moon et al. 2021). Businesses may strategically employ fake positive reviews to boost rankings, while competitors may use fake negative reviews to damage rivals (Burtch et al. 2018). Algorithmic ranking mechanisms further exacerbate the problem, as manipulated reviews influence product visibility and consumer preferences (Mohawesh et al. 2021). Given the low cost of generating fake reviews, both manually and through

✉ Mekhail Mustak
mekhail.mustak@hanken.fi

[1] University of Vaasa, Vaasa, Finland

[2] Department of Marketing, Hanken School of Economics, Arkadiankatu 22, 00100 Helsinki, Finland

[3] Qatar Computing Research Institute, Hamad Bin Khalifa University, Ar Rayyan, Qatar

Published online: 12 March 2025

machine-learning-driven automation, fraudulent activity is escalating (Chen et al. 2022).

In general, fake reviews can be categorized into human-written and computer-generated types (Salminen et al. 2022). This study focuses exclusively on human-written fake reviews because they present unique linguistic and psychological characteristics that differ fundamentally from algorithmically generated text. Human deception involves intentional manipulation of language, including strategic use of emotional appeal, exaggeration, and persuasive techniques that reflect cognitive effort and social influence (Hartmann et al. 2019; Munzel 2016). These psychological elements are absent in computer-generated fake reviews, which rely on automated text production and lack human-like intent or adaptive strategies. Furthermore, existing detection methods for computer-generated reviews benefit from clear algorithmic patterns and known AI-generated markers, whereas human-written fake reviews remain an open challenge due to their variability and contextual adaptability.

Nevertheless, extant studies suggest that deceptive human-written reviews often exhibit distinct psycholinguistic footprints, including differences in tone, authenticity, and confidence. Thus, psychological traits conveyed through language provide valuable indicators of deception (Hartmann et al. 2019; Munzel 2016). This approach offers a promising avenue for identifying fake reviews, yet its potential remains largely untapped in current research. The predictive power of psycholinguistic attribution is also supported by the "psychology of a lie" assumption, also known as *deception theory*, which suggests that spammers leave specific psycholinguistic footprints revealing deception (Mukherjee et al. 2013a; Yoo and Gretzel 2009). Additionally, distributional semantics, which examines the distributional properties of words in large text corpora to understand their meanings and relationships, offers another layer of analysis for identifying fake reviews (Bruni et al. 2014; Mohawesh et al. 2021; Salminen et al. 2022).

However, while theoretical notions support the use of psycholinguistic cues and associated distributional semantics for detecting fake reviews, empirical research applying these methods remains scarce and lacks rigorous validation (Park et al. 2023; Plotkina et al. 2020; Salminen et al. 2022). Existing studies primarily focus on basic linguistic markers, such as sentiment polarity and lexical diversity, but fail to capture the deeper psychological dimensions underlying deceptive language (Wu et al. 2020). Additionally, there is a notable lack of interdisciplinary research combining insights from computer science, psychology, and business to develop more effective detection frameworks (Kumar et al. 2018; Salminen et al. 2022). Without an integrated approach, existing methods remain insufficient for accurately identifying human-driven deception in online reviews.

Against this backdrop, *this research integrates insights from psychology, marketing, and computer science to explore psycholinguistic markers in online reviews for identifying fake reviews.* To achieve this aim, we address the following two research questions (RQs):

RQ1: How do human-written fake reviews differ from genuine reviews in terms of their psycholinguistic features?

RQ2: How can (a) psycholinguistics and (b) distributional semantics be leveraged for the automatic detection of human-written fake reviews?

We address RQ1 by investigating the psycholinguistic characteristics of 3,070 reviews written by 307 individuals in the United Kingdom. Half of these reviews pertain to fake brands, while the other half are based on real consumer experiences within the same product categories. For RQ2, we train ML classifiers using psycholinguistic features and distributional semantics, also known as *word embeddings*, to predict the authenticity of reviews, leveraging natural language processing (NLP) techniques.

This study advances theoretical understanding of fake review detection by integrating psycholinguistic analysis with advanced transformer-based models, providing a granular analysis of linguistic patterns that differentiate human-written fake reviews from genuine ones (Park et al. 2023; Wu et al. 2020). By combining LIWC-derived psycholinguistic insights with deep learning-based distributional semantics, our research extends prior work on deceptive language use and computational deception detection. Methodologically, our findings validate the efficacy of transformer-based NLP models in detecting human deception in online reviews and highlight the importance of interdisciplinary approaches, drawing from psychology, computational linguistics, and machine learning. Practically, this research offers valuable insights for enhancing automated fake review detection systems, equipping platforms and businesses with more effective tools to combat online deception and maintain trust in digital marketplaces.

The remainder of this article is structured as follows: Sect. "Conceptual Underpinnings" shows the conceptual underpinnings, exploring the relevant literature that inform our study. Sect. "Methodology" outlines the methodology employed. In Sect. "Analysis and Results", we present the findings of our analysis, highlighting the performance of various models and the key psycholinguistic features that distinguish fake reviews. Sect. "Discussion and Conclusion" provides a comprehensive discussion of the results, situating them within the broader academic discourse, and articulates the theoretical contributions and practical implications of our study. This section also acknowledges the limitations of the study, and offers detailed suggestions for future research directions.

Conceptual underpinnings

Online reviews and the prevalence of fake reviews

Online reviews serve as a crucial component in the digital marketplace, providing insights from consumers about their experiences with products and services (Burtch et al. 2018; Moon et al. 2021; Wang, Zhang, and Xu 2024). They function as a form of electronic word-of-mouth, enabling potential buyers to assess the quality and performance of products based on the opinions of previous customers (Schoenmueller et al. 2020; Zhang et al. 2016). The influence of online reviews extends to businesses as well, where they play a significant role in shaping a company's online reputation and consumer trust (Burtch et al. 2018; Venkatakrishnan et al. 2024; Wu et al. 2020). The growing significance of online reviews has also contributed to the proliferation of fake reviews.

These "fictitious reviews that have been deliberately written to sound authentic" (Li et al. 2014, p. 1566) have become a prominent concern, with studies suggesting that about 16% of restaurant reviews on Yelp and up to 42% of Amazon product reviews are suspicious (Luca and Zervas 2016). Diverse terminology is employed across fields such as computer science, business, and psychology to describe these fabricated reviews, including fraudulent, false, inauthentic, bogus, deceptive, non-truthful, and incentivized reviews, as well as opinion spam and review spamming (Ananthakrishnan et al. 2020; Costa Filho et al. 2023; Munzel 2016; Salminen et al. 2022). Some estimates indicate that one-third of online reviews might be fake (Salehi-Esfahani and Ozturk 2018), raising concerns given the vast number of reviews published annually. Their prevalence undermines the credibility of online review systems.

Fake reviews are crafted to mislead consumers by falsely boosting a product's reputation with positive reviews or damaging a competitor's standing with negative ones (Ananthakrishnan et al. 2020; Moon et al. 2021). The impact of these reviews is extensive, distorting sentiment analysis and making it difficult for consumers to discern genuine feedback from fraudulent content (Liu et al. 2017; Luca and Zervas 2016). This leads to poor purchasing decisions and erodes trust in both the reviews and the platforms hosting them (Chen et al. 2022; Kumar et al. 2018; Park et al. 2023).

Detection of fake reviews

The effectiveness of deploying fake reviews depends on the reviewers' ability to remain undetected, with more sophisticated fake reviews causing greater harm to the digital marketplace (Plotkina et al. 2020; Zhang et al. 2023). The challenges of detecting them stem from their increasingly sophisticated nature, often crafted to convincingly mimic genuine feedback. Services like ReviewMeta do not even label suspicious reviews as "fake" but identify patterns such as repetitive phrases, multiple reviews from a single author, and high review deletion rates (Hartmann et al. 2019).

The technical challenges in automated fake review detection involve algorithms, datasets, and features (Mohawesh et al. 2021; Ott, Cardie, and Hancock 2013; Salminen et al. 2022). Effective algorithms must detect subtle signals in the data. That necessitates datasets with validated ground truth labels—highly accurate, human-verified classifications of real and fake reviews—to train models effectively (Bell 2020). Features used for detection can include linguistic nuances, sentiment scores, and behavioral patterns, all of which are critical for accurately identifying fake reviews (Abri et al. 2020; Alsubari et al. 2020).

In terms of *algorithms*, fake review detection is generally approached as a binary classification problem, with two available target classes: fake and real. For this purpose, the typical algorithm to apply is logistic regression (Zhao and Wang 2016). Further, support vector machines (SVM) are often used as baseline algorithms for NLP tasks (Ott et al. 2011). *Dataset* variability also presents significant challenges. Larger datasets scraped from public sources offer abundant linguistic examples but often lack reliable ground truth labels. On the other hand, smaller, manually curated datasets, while more accurate, exhibit limited diversity (Mukherjee et al. 2013a). This trade-off complicates the development of universally applicable fake review detection models.

Fake review detection studies categorize *features* into behavioral, text-based, and hybrid approaches (Mohawesh et al. 2021; Mukherjee et al. 2013a). Behavioral features, such as reviewing frequency, account recency, geographic location, and the proportion of positive reviews can have good predictive power (Salminen et al. 2022). But they are often inaccessible due to privacy concerns. Consequently, researchers frequently rely on text-based features, analyzing the structure, content, and style of reviews to detect anomalies (Kauffmann et al. 2020; Wu et al. 2020). Hybrid approaches leverage the strengths of both behavioral and text-based features, providing a more comprehensive and effective means of identifying fake reviews (Mohawesh et al. 2021). Despite limitations, these methods are essential for improving the accuracy and reliability of detection systems.

Psycholinguistic features for fake review detection

Psycholinguistic features delve into the psychological aspects conveyed through language, offering valuable

insights into the writer's state of mind (Hartmann et al. 2019; Munzel 2016) (Hartmann et al. 2019; Munzel 2016). These features include aspects such as word choice, syntax, semantics, and pragmatics, which can reveal underlying psychological traits and intentions. The application of psycholinguistic features in detecting fake reviews is grounded in the premise that deceptive language exhibits distinct patterns that differ from genuine language (Vrij 2000). This approach examines the language choices and patterns employed in a review, potentially revealing underlying motives or inauthenticity (Yoo and Gretzel 2009). By analyzing these features, researchers can develop methods to distinguish between genuine reviews and those crafted with deception in mind (Mukherjee et al. 2013a).

For example, fake reviews often employ exaggerated emotional language or excessive positivity to appear authentic (Yoo and Gretzel 2009). Positive fake reviews may exhibit overly enthusiastic and insincere language, while negative fake reviews frequently rely on exaggerated negativity or emotional manipulation (Munzel 2016). By identifying these emotional discrepancies, researchers can develop algorithms to flag reviews that deviate from the expected emotional patterns of genuine feedback (Hancock et al. 2007). Extant research regarding psycholinguistic features in relation to our study are presented in Table 1.

Recent research has explored the role of psycholinguistic and computational features in detecting fake reviews, employing both LIWC-based linguistic markers and machine learning techniques. (Yoo and Gretzel 2009) were among the first to analyze deception in online reviews, showing that lexical complexity, first-person pronouns, and brand mentions were key indicators of fake content. (Ott et al. 2011; Ott, Cardie, and Hancock 2013) expanded on this by applying LIWC and n-gram analysis, achieving 89.8% accuracy in distinguishing fake from real hotel reviews.

Further research has investigated feature-based detection models across various industries. (Li et al. 2014) analyzed restaurant, hotel, and doctor reviews, finding that first-person pronouns, positive emotions, and perceptual processes were common in fake reviews, whereas negative emotions and punctuation marks were more prevalent in genuine reviews. (Mukherjee et al. 2013a) provided a critical perspective by

Table 1 Use of psycholinguistic and computational features in fake review detection research

Study	Dataset used	Industry	Significant features
Yoo and Gretzel (2009)	TripAdvisor (n = 82)	Hotel	Lexical complexity (F), first-person pronouns (F), positive sentiment (D), brand names (F)
Ott et al. (2011)	AMT (n = 800)	Hotel	LIWC with bigrams resulted in 89.8% accuracy
Ott et al. (2013)	AMT (n = 1600)	Hotel	Expanded dataset confirming LIWC's effectiveness in deception detection
Li et al. (2014)	AMT (n = 800)	Restaurant, Hotel, Doctor	First-person pronouns (F), positive emotions (F), words > 6 letters (F), leisure words (F), future focus (F), perceptual processes (F), commas (F), punctuation (R), numbers (R), "we" (R), negative emotions (R)
Mukherjee et al. (2013a)	Yelp (n = 35,593) & (n = 5,124)	Restaurant, Hotel	Linguistic features were ineffective, while behavioral features were strong predictors of deception
Mukherjee et al. (2013b)	Yelp (n = 5,124) & AMT-generated data	Restaurant, Hotel	Real-world fake reviewers exhibit distinct psycholinguistic patterns compared to AMT workers; n-gram-based models generalize poorly from AMT to real-world deception
Narayan et al. (2018)	AMT (n = 1600)	Hotel	No specific feature importance reported
Alsubari et al. (2020)	Yelp (n = 30,476)	Restaurant	Authenticity (R), analytical thinking (F)
Abri et al. (2020)	Restaurant Dataset (n = 110)	Restaurant	Pausality (F), number of adjectives (F), redundancy (F)
Moon et al. (2021)	AMT (n = 3,236)	Hotel	Present/future time orientation (F), lack of details (F), emotional exaggeration (F), first-person references (F)
Salminen et al. (2022)	Amazon e-commerce dataset	Cross-industry	Generated AI-based fake reviews (ULMFiT, GPT-2); Found that machine classifiers outperform human raters in detecting both AI-generated and human-written fake reviews
Liu et al. (2024)	E-commerce platforms in China	Cross-industry	Determinants of multimodal fake review generation using signaling and actor-network theories

showing that linguistic features were ineffective on real-life Yelp data, while behavioral patterns were strong predictors of deception. (Mukherjee et al. 2013b) further demonstrated that fake reviews written by commercial spammers exhibit distinct psycholinguistic features compared to AMT-generated fake reviews, revealing that n-gram models trained on crowdsourced data fail to generalize to real-world deception.

Recent studies have incorporated advanced NLP and AI-based approaches. (Moon et al. 2021) found that fake reviews tend to exaggerate emotions, lack details, and use more first-person references. (Alsubari et al. 2020) and (Abri et al. 2020) explored authenticity, analytical thinking, pausality, and redundancy as linguistic markers of deception. (Narayan, Rout, and Jena 2018) examined hotel reviews but did not identify significant linguistic patterns.

With the rise of AI-generated fake reviews, research has shifted toward multimodal and deep learning-based detection techniques. (Liu et al. 2024) explored determinants of fake review generation in China's e-commerce platforms, applying signaling and actor-network theories. (Salminen et al. 2022) investigated AI-generated fake reviews using ULMFiT and GPT-2, showing that machine classifiers significantly outperform human raters in detecting both AI-generated and human-written fake reviews.

While these studies have advanced the field, most rely on either psycholinguistic or computational approaches without fully integrating distributional semantics and transformer-based NLP models. Our study extends this work by combining LIWC-based psycholinguistic insights with transformer-based models (e.g., BERT, RoBERTa, XLNet), offering a more robust and scalable approach to fake review detection across multiple product categories.

Beyond sentiment analysis, psycholinguistic features can be extended to examine language complexity and style. Fake reviews may exhibit simpler structures and limited vocabulary, potentially due to attempts at creating generic content (Mohawesh et al. 2021; Ott et al. 2011). Conversely, over-reliance on specific pronouns, adverbs, or intensifiers in fake reviews can be a sign of inauthentic persuasion tactics. Furthermore, these features can delve into a reviewer's mental state. Markers of cognitive load, such as complex sentence structures or unusual vocabulary choices, might indicate the effort required to fabricate a review, aiding in differentiating between genuine and deceptive content (Hartmann et al. 2019; Salminen et al. 2022).

Finally, psycholinguistic features can incorporate the use of distributional semantics. This approach examines the semantic relationships between words within a text corpus, allowing researchers to assess the coherence and topical relevance of language use (Salminen et al. 2022). Fake reviews may exhibit a mismatch between the sentiment expressed and the words used, or they may contain irrelevant or tangential phrases (Chen et al. 2022). Analyzing these semantic relationships can help identify reviews that lack the thematic consistency characteristic of genuine feedback.

Psycholinguistic features demonstrably enhance the effectiveness of ML models in detecting fake reviews (Mukherjee et al. 2013a). By incorporating these features, researchers can train models to identify subtle linguistic cues that differentiate genuine and deceptive content, leading to more robust detection systems (Salminen et al. 2022). This interdisciplinary approach, integrating psychology, linguistics, and computer science, represents a significant step forward in combating online deception. In our study, we leverage state-of-the-art transformer-based models pre-trained on vast amounts of text data, allowing for fine-tuning on a smaller dataset specific to fake review detection (Devlin et al. 2019; Liu et al. 2019; Salminen et al. 2022).

Methodology

Task design

Five product and service categories were selected based on an internal planning session among the research team. These categories were mobile phones, shoes, movies, restaurants, and shampoo. These categories were primarily chosen for their commonness, ensuring that the consumers participating in the study would likely have experiences using products and services from these categories. The selection includes three physical products (mobile phones, shoes, and shampoo) and two services (movies and restaurants). For each category, we assigned fictitious brand names to facilitate participants' task completion. The fake brands and their corresponding categories are shown in Table 2.

We estimated that including more than five categories would risk participants becoming bored or unfocused. Each participant was tasked with writing a review of a fictitious brand in one of the selected categories, followed by writing a review of a real existing brand from the same category. This within-subjects design is illustrated in Fig. 1.

Each participant wrote ten reviews, comprising five fake reviews based on non-existing brands and five real reviews based on authentic product or service encounters. After composing each review, participants indicated the intended

Table 2 Five fake product names created for the task

Fake brand name	Category
Zorkia	Mobile phone
Ogli	Shoe
Lobobo	Restaurant
Scover	Movie
Acterin	Shampoo

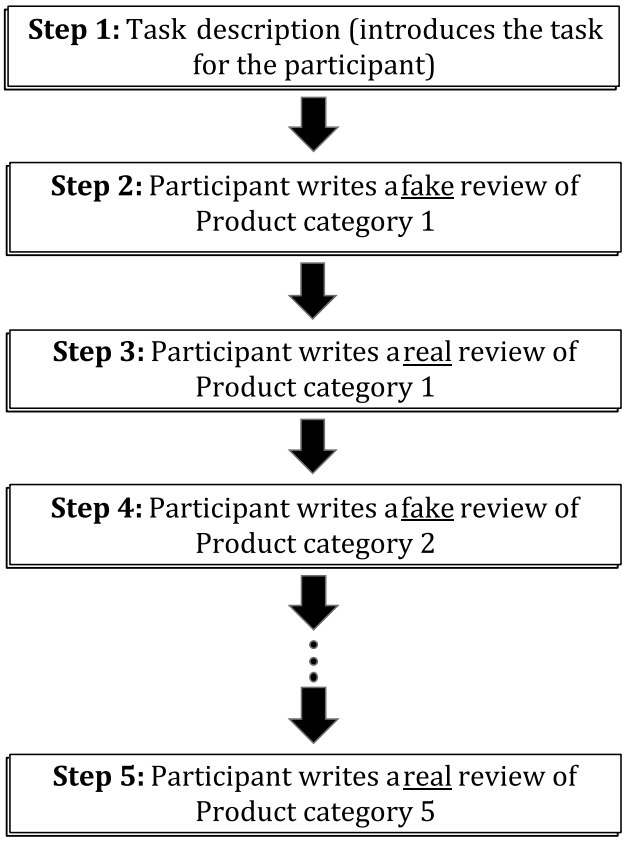

Fig. 1 The research flow of this study (corresponding to a within-subjects design)

sentiment of their review ("My review was generally… positive/negative/neutral"). The text box shows the general instructions provided to the participants. After the general instructions, precise instructions were repeated before the participant wrote their review in a form field (see examples in Table 3).

Please write the FAKE and REAL reviews below according to the instructions

Avoid short and non-informative reviews, like "It's good." Instead, use your creativity to write a proper review that aims to be informative (i.e., contains specific details)

Do NOT copy reviews from the web; write in your own words

Here's an example of a review: "I really liked Nokia 808 phone. It had a great camera and build quality. I didn't have any issues with the operating system, unlike some other people I know."

Operationalizing fake reviews as reviews for non-existent products in research offers several advantages that enhance their comparability to real-world fake reviews. This approach provides a definitive "ground truth," addressing a key challenge in fake review detection by eliminating uncertainty in classification. Additionally, it ensures a controlled research environment, mitigating biases stemming from participants' prior product knowledge and enabling the generation of psycholinguistically authentic reviews.

By maintaining consistency in review quality, this method facilitates the development of reliable detection benchmarks while preventing ethical concerns associated with fabricated reviews influencing real markets. Analyzing deceptive language patterns in the absence of actual product experiences isolates the cognitive processes underlying deception. Furthermore, incorporating diverse product categories enhances the generalizability of findings, aligning with established research methodologies and allowing meaningful comparisons with prior studies.

Participants of the Study

Participants were recruited through Prolific, an online research platform widely validated for data quality in social science research. We employed Prolific's representative sampling option, which aligns demographic distributions with those of a given national population, thereby enhancing the generalizability of our findings. Accordingly, our sample was cross-stratified by age, gender, and ethnicity using a simplified framework based on the Great Britain Census (GB Census), a national survey conducted by the UK government to collect demographic data.

The UK was selected as the sampling country for two primary reasons. First, to ensure that all reviews were written in English, minimizing potential language-related confounds. Second, the UK is among the most advanced nations in e-commerce adoption, making its consumers highly familiar

Table 3 Instructions for writing a (a) fake mobile phone review, and (b) real mobile phone review

(a)	(b)
Write a review about Zorkia, an imaginary cell phone that does not exist. Pretend to be an owner of this phone and write a review about it	Write a review of a REAL cell phone that you have owned. Choose freely any cell phone
NOTE: avoid short and non-informative reviews, like "It's good." Instead, use your creativity to write a proper review that aims to be informative (i.e., contains specific details)	NOTE: avoid short and non-informative reviews, like "It's good." Instead, use your creativity to write a proper review that aims to be informative (i.e., contains specific details)

with online reviews. Based on Prolific's recommendations, the target sample size was at least 300 participants, and we ultimately recruited 307 participants.

Participants were directed from Prolific to the Qualtrics survey platform, where they completed the task by writing their reviews in multiple form fields. The collected reviews were then exported for research analysis and modeling. The sample comprised 153 males (49.8%) and 154 females (50.2%), with no non-binary participants. The average age of the participants was 44.7 years (SD = 15.5), with a median age of 43.5 years. On average, participants took 24.3 min (SD = 27.4) to complete the task of writing ten reviews, averaging 2.4 min per review. The median task completion time was 16.7 min. Participants were compensated based on the UK minimum hourly wage at the time of data collection.

Data exploration

In total, 3070 reviews were obtained, with each participant writing ten reviews, consisting of five fake and five real reviews as previously described. The dataset is perfectly balanced, with 50% (n = 1535) representing fake reviews and the other 50% (n = 1535) representing real reviews. Each product or service category contains 614 reviews, which constitutes 20% of the total dataset, equally split between fake and real reviews.

To ensure data quality, we implemented a two-step verification process. First, we used Turnitin plagiarism detection software to verify that participants had not copied their reviews from the web, leveraging Turnitin's extensive database, which includes billions of web pages. Second, two researchers manually screened each review for coherence and plausibility, ensuring that the text was not gibberish or randomly generated and that it appeared authentic. This thorough review process confirmed that the data met quality standards, and no reviews were omitted from the study. Table 4 presents descriptive statistics on the length of the collected reviews, comparing real and fake reviews in terms of minimum and maximum length, median, mode, interquartile range (IQR), and relative standard deviation (RSD).

Table 4 Descriptive statistics of the reviews' length (in characters)

	Real (n = 1535)	Fake (n = 1535)
Minimum	37	32
Maximum	894	835
Median	191	166
Mode	150	118
IQR	125	118
RSD	51.8%	54.1%

RSD relative standard deviation

In analyzing the dataset, several key findings emerged. First, real reviews (M = 211.7 characters) were significantly longer on average than fake reviews (M = 188.9 characters), $t(1534) = 10.3$, $p < 0.001$. This finding contrasts with previous research (e.g., Abri et al. 2020), which suggested that fake reviews tend to be longer. Second, there is a clear negativity bias in fake reviews, as evidenced in Table 5. Fake reviews were more likely to be negative (40.7%), whereas real reviews had a much lower proportion of negative sentiment (13.2%). Conversely, real reviews exhibited a strong positivity bias, with 75.9% being positive, compared to only 49.3% of fake reviews. Neutral reviews constituted a small portion of both categories, with real reviews being 10.9% neutral and fake reviews 10.0% neutral.

Further analysis reveals that these sentiment differences persist across all product categories (see Fig. 2). This suggests that when participants fabricated reviews, they tended to emphasize negative aspects, potentially due to the absence of real experiences to elaborate on. Table 5 provides a detailed breakdown of sentiment distributions in real and fake reviews.

A chi-squared test of independence confirms that fake reviews are significantly more likely to be negative ($\chi^2(1, N = 2{,}747) = 302.6$, $p < 0.00001$). This analysis excludes neutral reviews (n = 323) to focus on the contrast between positive and negative reviews. The distribution of neutral reviews is consistent across both categories, with 168 neutral real reviews and 155 neutral fake reviews, collectively representing only 10.5% of the dataset. This low proportion suggests that when participants fabricate reviews, they are more inclined to adopt strongly positive or negative sentiments rather than neutrality. This finding aligns with prior research indicating that deceptive content often involves emotional exaggeration rather than neutral expression (Munzel 2016; Yoo and Gretzel 2009).

Table 6, which focuses on the mobile phone category, presents examples of reviews across different sentiment categories, illustrating these observed patterns. In each pair, the upper review reflects the original text written by a participant, while the lower review demonstrates the brand name replacement process. Brand names were substituted with randomly generated strings to enhance the model's generalizability beyond specific brands. This modification aimed to increase the task's difficulty for machine learning models,

Table 5 Number of reviews in each sentiment class

	Positive	Negative	Neutral	Total
Real	1165 (75.9%)	202 (13.2%)	168 (10.9%)	1535
Fake	756 (49.3%)	624 (40.7%)	155 (10.0%)	1535
Total (%)	*1921 (62.6%)*	*826 (26.9%)*	*323 (10.5%)*	*N = 3070*

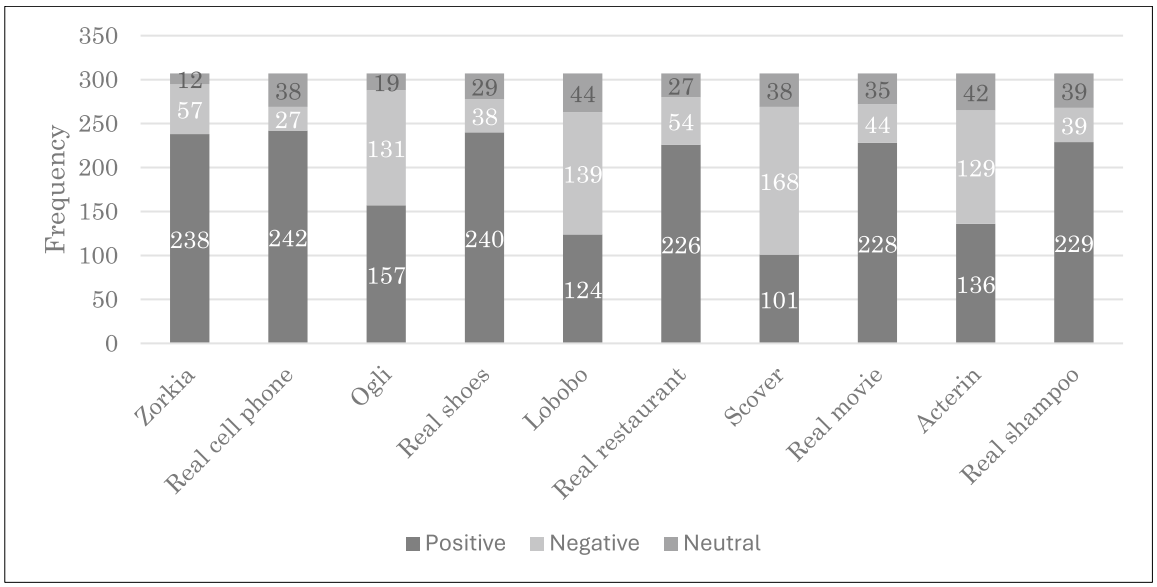

Fig. 2 Review sentiment differences between fake and real reviews across product and service categories

Table 6 Examples of fake and real reviews in the mobile phone category

Fake	Real
Zorkia is a phone that I was instantly attracted to due to how affordable the price point was, be that as it may it failed to deliver on essential features that I use regularly. I guess you get what you pay for	The Motorola is a well built 4G phone with an excellent camera and 128 Gb of storage. Android 11 and high spec. Well worth the money with a good guarantee and 2 years of upgrades
btcmwb is a phone that i was instantly attracted to due to how affordable the price point was, be that as it may it failed to deliver on essential features that i use regularly. i guess you get what you pay for	The qntlinrr is a well built 4 g phone with an excellent camera and 128 gb of storage. gkksrdug11 and high spec. well worth the money with a good guarantee and 2 years of upgrades

ensuring robustness in distinguishing between fake and real reviews across various contexts.

Analysis and results

Psycholinguistic distinctions between fake and real reviews: an LIWC analysis

We employed the Linguistic Inquiry and Word Count (LIWC) method to examine how human-written fake reviews differ from genuine reviews in terms of their psycholinguistic features. LIWC is a widely used tool in computational linguistics and NLP for analyzing psychological aspects of text. It utilizes multiple manually curated dictionaries to infer psychological, cognitive, and social processes based on word usage (Tausczik and Pennebaker 2010). This technique has been extensively applied in machine learning and NLP research to examine various social science topics (Tausczik and Pennebaker 2010). In our analysis, we utilized the full spectrum of 93 LIWC categories to evaluate 3,070 reviews used in our study.

Our analysis comprised two main steps. First, we conducted descriptive analyses using t-tests to compare the mean LIWC scores between fake and real reviews. Second, we employed logistic regression models to evaluate the predictive power of LIWC features in distinguishing fake reviews from real ones.

Given the high number of predictors, we initially checked for redundancy by examining a cross-correlation matrix and subsequently removed highly correlated variables using a correlation cutoff of 0.7. The variables excluded were word count (“wc”), pronouns (“pronoun”), personal pronouns (“ppron”), positive emotions (“posemo”), and informal text (“informal”). Additionally, we applied a square root-based scaling function using the R programming language to normalize the varying scale ranges of the LIWC variables. Descriptive analyses included frequency tables for categorical variables and mean (standard deviation, SD) for continuous variables. Associations between variables were assessed

using the Chi-square test for independence for categorical variables and t-tests for continuous variables, with a 95% confidence level for significance testing.

The descriptive results, summarized in Table 7, reveal significant associations between review type (fake vs. real) and several LIWC categories. The full results are available from the authors upon request.

Table 7 highlights key psycholinguistic differences between fake and real reviews, revealing significant distinctions across various Linguistic Inquiry and Word Count (LIWC) categories. Real reviews tend to exhibit higher levels of analytical thinking ($M = 0.91$ vs. $M = 0.85$, $p = 0.002$), clout ($M = 0.87$ vs. $M = 0.76$, $p < 0.001$), emotional tone ($M = 0.99$ vs. $M = 0.79$, $p < 0.001$), and word complexity, including longer sentences (words per sentence: $M = 0.94$ vs. $M = 0.86$, $p < 0.001$) and more words with six letters or more ($M = 0.94$ vs. $M = 0.88$, $p < 0.001$). They also show greater use of third-person singular pronouns ($M = 0.18$ vs. $M = 0.10$, $p = 0.023$), numbers ($M = 0.52$ vs. $M = 0.43$, $p = 0.004$), social processes ($M = 0.78$ vs. $M = 0.72$, $p = 0.019$), mentions of the male gender ($M = 0.22$ vs. $M = 0.12$, $p = 0.006$), and affiliation-related words ($M = 0.48$ vs. $M = 0.40$, $p = 0.013$), indicating more detailed and socially-oriented content.

Conversely, fake reviews are characterized by higher cognitive processes ($M = 0.89$ vs. $M = 0.83$, $p = 0.003$), discrepancies ($M = 0.58$ vs. $M = 0.49$, $p = 0.005$), perceptual processes ($M = 0.74$ vs. $M = 0.67$, $p = 0.016$), and a focus on the future ($M = 0.45$ vs. $M = 0.38$, $p = 0.033$). They also display more frequent use of certain linguistic features such as punctuation marks ($M = 0.92$ vs. $M = 0.87$, $p = 0.003$), impersonal pronouns ($M = 0.81$ vs. $M = 0.73$, $p < 0.001$), auxiliary verbs ($M = 0.94$ vs. $M = 0.87$, $p < 0.001$), negations ($M = 0.72$ vs. $M = 0.47$, $p < 0.001$), and expressions of negative emotions like anxiety ($M = 0.36$ vs. $M = 0.15$, $p < 0.001$), anger ($M = 0.29$ vs. $M = 0.13$, $p < 0.001$), and sadness ($M = 0.36$ vs. $M = 0.23$, $p < 0.001$). These findings underscore the nuanced differences in language use between fake and real reviews, providing valuable insights into the distinct psychological and linguistic patterns that can help in distinguishing between the two.

Psycholinguistics and distributional semantics in fake review detection

Leveraging psycholinguistics for detecting human-written fake reviews

Following the descriptive analysis, we constructed logistic regression models to evaluate the ability of LIWC features to predict whether a review is fake or real. Drawing inspiration from related research on fake review detection that employed feature selection techniques (Abri et al. 2020), we implemented a feature selection procedure. Specifically, we compared two feature selection methods: *group LASSO* (Least Absolute Shrinkage and Selection Operator) and forward/backward stepwise selection, both widely used in statistical learning (Hastie et al. 2015). LASSO penalizes the absolute size of the regression coefficients to shrink the coefficients of irrelevant variables to zero, while stepwise selection uses p-values to include or exclude variables.

Through this process, 57 variables were selected using LASSO, and 45 variables were selected using stepwise modeling. We then compared the resulting logistic models and calculated the odds ratios to interpret the results. Model performance was evaluated using standard metrics: accuracy, precision, recall, F1 score, area under the curve (AUC), and Kappa (Bell 2020). Accuracy measures the proportion of correctly predicted samples, precision indicates the correctness of positive predictions, recall assesses the proportion of actual positives correctly predicted, and F1 score is the harmonic mean of precision and recall. AUC represents the model's ability to distinguish between classes. Kappa measures the agreement between predicted and true labels, adjusted for chance.

We found that both models perform similarly based on the *Akaike Information Criterion* (AIC)—with AIC scores of 3799.8 for the stepwise model and 3819.3 for the LASSO model. However, when it comes to overall performance comparison, shown in Table 8, the stepwise model is slightly superior.

Overall, the results indicate that despite numerous significant LIWC predictors, the best model's accuracy is relatively modest at 67.5%, which is only 35% better than a random guess (50%). Kappa metrics ($k_{stepwise} = 0.350$ and $k_{LASSO} = 0.309$) suggest a fair agreement between the classifiers and the true labels (McHugh 2012). Although the performance serves as a reasonable baseline for more advanced approaches, it is not adequate to reliably distinguish fake reviews from real ones, as the classifier would be incorrect approximately one in every three cases. Therefore, while several significant psycholinguistic indicators contribute valuable information to the fake review detection problem, their predictive ability alone is insufficient to build robust classifiers. Consequently, more advanced language representations are needed, which we address in the next subsection.

Leveraging distributional semantics for detecting human-written fake reviews

To examine how distributional semantics can be leveraged for the automatic detection of human-written fake reviews, we tested a variety of ML algorithms, including both traditional data science models and advanced transformer architectures. The traditional algorithms utilized were Support Vector Machine (SVM), XGBoost (XGB), and multi-layer

Table 7 Results for testing the statistical significance of LIWC features between fake and real reviews

	Fake review (0, N = 1535)	Real review (1, N = 1535)	p-value
Higher for real reviews			
Analytical thinking (analytic)	0.85 (0.48)	0.91 (0.46)	0.002**
Length (len)	0.83 (0.45)	0.93 (0.48)	< 0.001***
Clout (clout)	0.76 (0.57)	0.87 (0.58)	< 0.001***
Emotional tone (tone)	0.79 (0.50)	0.99 (0.38)	< 0.001***
Words per sentence (wps)	0.86 (0.41)	0.94 (0.45)	< 0.001***
Words longer than six letters (sixltr)	0.88 (0.41)	0.94 (0.42)	< 0.001***
Third-person singular (shehe)	0.10 (0.81)	0.18 (1.14)	0.023*
Numbers (number)	0.43 (0.86)	0.52 (0.90)	0.004**
Social processes (social)	0.72 (0.66)	0.78 (0.66)	0.019*
Mentions of male gender (male)	0.12 (0.84)	0.22 (1.11)	0.006**
Affiliation (affiliation)	0.40 (0.85)	0.48 (0.94)	0.013*
Reward (reward)	0.60 (0.75)	0.68 (0.78)	0.008**
Leisure (leisure)	0.46 (0.86)	0.53 (0.87)	0.025*
Home (home)	0.16 (0.89)	0.24 (1.06)	0.015*
Death (death)	0.07 (0.73)	0.17 (1.20)	0.003**
Nonfluencies (nonflu)	0.26 (0.88)	0.34 (1.02)	0.028*
Parentheses (parenth)	0.08 (0.77)	0.20 (1.17)	0.001**
Third-person plural (they)	0.34 (0.85)	0.47 (0.97)	< 0.001***
Conjuctions (conj)	0.86 (0.45)	0.93 (0.44)	< 0.001***
Quantifiers (quant)	0.57 (0.76)	0.67 (0.81)	< 0.001***
Present focus (focuspresent)	0.81 (0.53)	0.90 (0.50)	< 0.001***
Higher for fake reviews			
Cognitive processes (cogproc)	0.89 (0.52)	0.83 (0.49)	0.003**
Discrepancies (discrep)	0.58 (0.90)	0.49 (0.78)	0.005**
Perceptual processes (percept)	0.74 (0.74)	0.67 (0.67)	0.016*
Hear (hear)	0.47 (1.00)	0.37 (0.80)	0.002**
Future focus (focusfuture)	0.45 (0.98)	0.38 (0.83)	0.033*
Money (money)	0.56 (0.86)	0.50 (0.84)	0.044*
Assent (assent)	0.24 (1.19)	0.15 (0.71)	0.008**
Punctuation marks (all_punc)	0.92 (0.48)	0.87 (0.42)	0.003**
Periods (period)	0.88 (0.55)	0.83 (0.50)	0.005**
Dictionary words count (dic)	1.00 (0.08)	0.99 (0.09)	< 0.001***
Function words (function)	1.00 (0.16)	0.97 (0.16)	< 0.001***
Impersonal pronouns (ipron)	0.81 (0.65)	0.73 (0.62)	< 0.001***
Auxiliary verbs (auxverb)	0.94 (0.43)	0.87 (0.42)	< 0.001***
Negations (negate)	0.72 (0.90)	0.47 (0.67)	< 0.001***
Verbs (verb)	0.98 (0.33)	0.91 (0.32)	< 0.001***
Negative emotions (negemo)	0.67 (1.05)	0.31 (0.59)	< 0.001***
Anxiety (anx)	0.36 (1.18)	0.15 (0.68)	< 0.001***
Anger (anger)	0.29 (1.18)	0.13 (0.71)	< 0.001***
Sadness (sad)	0.36 (1.09)	0.23 (0.80)	< 0.001***
Differentiation (differ)	0.77 (0.76)	0.63 (0.66)	< 0.001***
Risk (risk)	0.38 (1.07)	0.24 (0.80)	< 0.001***
Past focus (focuspast)	0.83 (0.75)	0.58 (0.64)	< 0.001***
Exclamation marks (exclam)	0.40 (1.10)	0.24 (0.76)	< 0.001***

Only significant results are shown

N = 3070 for all tests

Table 8 Logistic regression results with different feature selection techniques

	Accuracy	Precision	Recall	F1	Kappa
LASSO	0.654	0.638	0.715	0.674	0.309
Stepwise modeling	0.675	0.659	0.724	0.690	0.350

perceptron (MLP), commonly employed in text classification tasks and as baselines for comparison (Abri et al. 2020; Moon et al. 2021). The advanced transformer models included BERT, RoBERTa, and XLNet, representing state-of-the-art NLP techniques. These models were pre-trained on extensive text corpora and fine-tuned for binary classification (fake vs. real reviews) through transfer learning. All models were implemented using Python.

Features, or numerical representations of review texts, varied between the models. For baseline models, we used *Term Frequency-Inverse Document Frequency* (TF-IDF) technique to weigh words based on their commonness across the dataset. Transformer models, on the other hand, utilized word embeddings, representing words, sentences, and reviews as n-dimensional vectors that capture semantic relationships. Transformer experiments were conducted in the *Google Colab* environment using a nVIDIA® Tesla® K80 Accelerator. To enhance generalizability and prevent models from identifying reviews based on brand names, we replaced all brand names with random strings before preprocessing.

We employed supervised machine learning (ML) training algorithms using k-fold cross-validation (k = 5), which involves splitting the data into five subsets and training the model five times, each time using a different subset as the validation set and the remaining subsets for training. The data was divided into training, testing, and validation splits of 75%, 15%, and 10%, respectively (Krogh & Vedelsby, 1995). To fine-tune the initial models, we used a technique called grid search, which systematically tests different combinations of model parameters to find the optimal settings. For the transformer models, we utilized the Adam optimizer, an algorithm that adjusts the learning rate during training to improve performance. The learning rate was set to 0.0005, and the models were trained in batches of 32 samples at a time, ensuring efficient and effective parameter updates. This approach helped us maximize the accuracy and robustness of our detection algorithms. Further technical details and computational notebooks are available upon request from the authors.

The results, summarized in Table 9, indicate that transformer models generally outperform traditional models, with a macro-average accuracy of 81.9% compared to 63.8% for traditional models. Transformer models also show substantial agreement with true labels (k = 0.61–0.80), while traditional models only show fair agreement (k = 0.21–0.40). Among transformers, BERT performed best, achieving the highest scores in four out of six evaluation metrics, although the differences between transformer models were minor, typically around 5% or less.

Further analysis, illustrated in Fig. 3, reveals that reviews with negative sentiment are more challenging for the classifier, likely due to their lower proportion in the dataset. Additionally, all product categories achieved an F1 score above 0.7, indicating good model generalizability across different categories. Movies and shoes had the highest accuracy, whereas mobile phones had the lowest.

These findings underscore the effectiveness of transformer models in detecting fake reviews. They highlight the importance of advanced feature representations in enhancing model performance. Overall, the results indicate that while transformer models perform well across various categories, further refinements are needed to address specific challenges associated with different product types and sentiments.

Discussion and conclusion

General discussion

Fake reviews, often crafted to artificially enhance or damage a product's reputation, distort perceptions of product quality, leading to poor purchasing decisions and undermining the

Table 9 Model performance averaged across five test folds

	Traditional models			Transformers		
	SVM	XGB	MLP	BERT	RoBERTa	XLNET
Accuracy	65.3%	63.3%	62.7%	**82.6%**	82.0%	81.0%
Precision	0.660	0.645	0.656	0.855	**0.870**	0.864
Recall	0.632	0.594	0.550	**0.785**	0.755	0.734
F1	0.645	0.618	0.584	**0.818**	0.808	0.794
AUC	0.718	0.686	0.689	0.908	**0.910**	0.898
Kappa	0.306	0.266	0.255	**0.651**	0.641	0.619

Higher is better

The highest scores are bolded

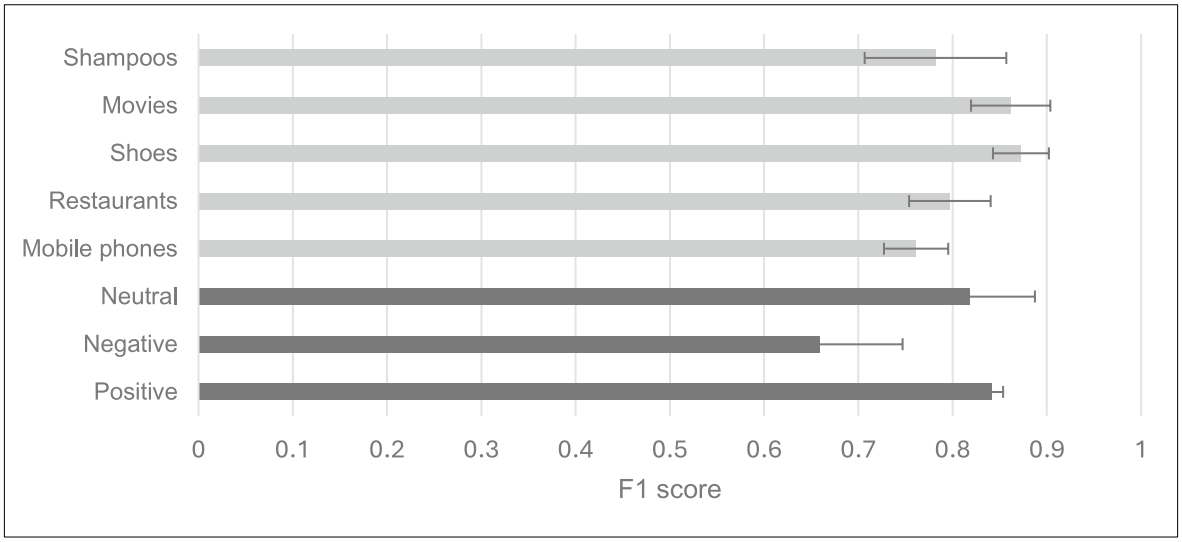

Fig. 3 F1 scores by product category and sentiment class when using the BERT classifier (i.e., the best model). Error bars indicate standard deviation across the five folds

credibility of online platforms (Salminen et al. 2022; Wu et al. 2020). This study examines the detection of human-written fake reviews by leveraging psycholinguistic features and advanced machine learning models to improve the reliability of online reviews (He et al. 2020; Wu et al. 2020).

We address two key issues: (1) the psycholinguistic differences between fake and genuine reviews and (2) the effectiveness of psycholinguistics and distributional semantics in automated detection. Our findings reveal significant psycholinguistic distinctions between fake and real reviews and demonstrate the superior performance of transformer-based models, such as BERT, in detecting deceptive content (Alsubari et al. 2020; Mukherjee et al. 2013b).

Our results regarding the LIWC features show deviations from previous findings. Alsubari et al., (2020) found higher authenticity and lower analytical thinking in real reviews compared to fake ones, whereas our dataset showed no significant difference in authenticity and higher analytical thinking in real reviews. Li et al. (2014) indicated that leisure words and longer words were more predictive of fake reviews, but we found these features more prominent in real reviews. Additionally, while previous studies found punctuation to be more predictive of real reviews, our results showed the opposite. Out of the 12 influential psycholinguistic characteristics, only one-third overlapped with our findings, indicating that future orientation, perceptual processes, and negative emotions are indicative of fake reviews, while numbers are more associated with real reviews. These inconsistencies suggest that psycholinguistic characteristics alone may not be sufficient for robust fake review detection without supplementary features. Moreover, logistic regression models based on LIWC features achieved modest accuracy, highlighting the complexity of detecting fake reviews in larger, more varied datasets.

In terms of distributional semantics, our results demonstrate the superior performance of transformer-based models, such as BERT, RoBERTa, and XLNet, over traditional ML algorithms. This aligns with recent advancements in NLP that have shown transformers' effectiveness in various text classification tasks (Devlin et al. 2019; Liu et al. 2019). Our study extends prior research that primarily utilized simpler models like logistic regression and support vector machines. Additionally, our study situates its findings within the broader literature by confirming that fake reviews often display exaggerated emotional tones and overly positive language, consistent with the psychological traits of deception (Munzel 2016; Yoo and Gretzel 2009).

Based on the findings and the above discussions, we can characterize several psycholinguistic features (PFs) regarding online reviews. Concerning real reviews, we discern that:

- PF 1: *Real reviews exhibit more analytical thinking (e.g., reflection) than fake reviews*.
- PF 2: *Real reviews are longer on average, contain more words per sentence, and have longer words than fake reviews*.
- PF 3: *Real reviews are more precise, with a higher use of quantifiers and numbers compared to fake reviews*.
- PF 4: *Real reviews focus more on the present, whereas fake reviews focus more on the future or the past*.
- PF 5: *Real reviews include more third-person singular and plural references, while fake reviews use more impersonal pronouns*.

- PF 6: *Real reviews contain more words related to rewards, whereas fake reviews contain more words related to risks*.
- PF 7: *Real reviews are less fluent than fake reviews, showing higher non-fluency*.
- In turn, the PFs (psycholinguistic features) we deduce for fake reviews are as follows:
- PF 8: *Fake reviews have more punctuation marks than real reviews, especially exclamation marks and periods*.
- PF 9: *Fake reviews mention money more frequently than real reviews*.
- PF 10: *Fake reviews contain more expressions of approval or agreement (assent) than real reviews*.
- PF 11: *Fake reviews include more negations, anxiety, and expressions of negative emotions (anger, sadness) than real reviews*.
- PF 12: *Fake reviews contain more verbs than real reviews*.

Our results also identified that reviews with negative sentiment are more challenging for classifiers to detect accurately, which contrasts with the common perception that negative information is more diagnostic (Shoham et al. 2017). This detail highlights the need to consider how sentiment is spread out when creating detection algorithms.

Theoretical contributions

This study contributes into fake review detection by revealing the critical role of psycholinguistic features and distributional semantics. Online reviews hold immense sway over consumer decisions and business outcomes (Schoenmueller et al. 2020; Wang et al. 2024). Yet, much of the existing research has focused on broader dimensions, such as the volume and valence of reviews (Zhang et al. 2023). This study goes a step further, delving deeply into the linguistic patterns that distinguish fake reviews from genuine ones. It identifies specific language cues that signal deception, providing a richer understanding of this pervasive issue.

This study also contributes by validating advanced transformer-based models, such as BERT, which excel in detecting subtle linguistic nuances critical for distinguishing fake reviews (Devlin et al. 2019; Liu et al. 2019). By integrating psycholinguistic features with cutting-edge NLP techniques, this research bridges psychological theories of deception with computational approaches. This synergy enhances the robustness of detection frameworks, offering a holistic method for analyzing deceptive content (Mukherjee et al. 2013b; Yoo and Gretzel 2009).

In doing so, the study further advances deception theory, often referred to as the "psychology of a lie." This theory posits that deceptive language leaves psychological traces, such as heightened cognitive effort and emotional exaggeration. The research empirically validates these principles by identifying linguistic markers, including discrepancies in focus and exaggerated emotions, which reflect the mental strain involved in fabricating information. By linking these insights to scalable computational tools, the study not only deepens theoretical understanding but also provides a practical, data-driven framework for detecting deception in digital contexts. While prior research, such as Moon et al. (2021), has successfully applied LIWC to detect human-generated fake reviews, our study expands this approach by integrating psycholinguistic insights with distributional semantics and transformer-based models, offering a more robust detection framework. Additionally, our use of multiple product categories and a demographically representative dataset enhances the generalizability of our findings.

While previous research, such as Moon et al. (2021), has effectively applied LIWC for fake review detection, our study advances beyond this work in several key ways. First, Moon et al. (2021) primarily relied on Amazon Mechanical Turk (AMT) participants to generate fake and genuine hotel reviews. In contrast, our study collected reviews using Prolific, a platform designed to provide a more demographically representative sample, ensuring broader generalizability of findings. Second, Moon et al. (2021) focused on a single domain—hotels—whereas our study expands the scope to multiple product categories, including mobile phones, shoes, restaurants, movies, and shampoo, allowing for a more comprehensive examination of deceptive linguistic patterns across industries. Third, while Moon et al. (2021) identified psycholinguistic markers of deception, such as lack of details, present/future time orientation, and emotional exaggeration, their approach relied exclusively on linguistic feature analysis. In contrast, our study integrates distributional semantics and transformer-based NLP models (e.g., BERT, RoBERTa, XLNet) to complement LIWC analysis, providing a more robust, scalable, and context-aware detection framework. By combining psycholinguistic insights with deep learning models, our research offers a more advanced methodology for identifying human-written fake reviews across diverse product categories.

The interdisciplinary nature of this study adds further value. It merges insights from psychology (Mukherjee et al. 2013b), computational linguistics (Li et al. 2014; Taboada et al. 2011), and consumer behavior to create a comprehensive framework for understanding deception (Banerjee 2022; Banerjee and Chua 2021). This approach not only strengthens theoretical models but also highlights the interplay between linguistic and psychological dimensions in fake reviews. Such a perspective provides a solid foundation for advancing research in this area.

In summary, this research not only deepens theoretical insights into fake review detection but also strengthens connections between psychological theories and computational

methods (Tausczik and Pennebaker 2010). It lays the groundwork for more effective interdisciplinary approaches to addressing online deception, paving the way for a more trustworthy digital ecosystem.

Practical implications of this study

Based on the findings, we draw the following practical implications of this study. Companies should:

- Implement psycholinguistic features in review monitoring systems to better analyze linguistic cues that differentiate fake reviews from genuine ones.
- Integrate transformer-based models like BERT, RoBERTa, and XLNet into the marketing analytics framework. These models enhance the detection of fake reviews, allowing for more precise identification and filtering of deceptive content.
- Incorporate sophisticated NLP techniques into information systems, leading to more accurate identification of deceptive content and thereby improving the reliability of systems that process and analyze user-generated data.
- Recognize that reviews with negative sentiment are more challenging for classifiers to detect accurately, necessitating consideration of sentiment distribution in review analysis and monitoring systems. Addressing negative sentiments can have a substantial impact on consumer trust and decision-making.

Through more accurate identification, companies can better filter out deceptive content, ensuring that business decisions are based on authentic consumer feedback and more reliable data. Consequently, this not only improves the effectiveness of marketing campaigns but also nurtures greater consumer trust in online reviews, thereby enhancing overall brand reputation and consumer loyalty. Further, for review writers, the study highlights the importance of transparency and authenticity in review writing. Reviewers are encouraged to provide genuine feedback, knowing that their honest opinions contribute to a trustworthy review ecosystem.

Limitations of this study and suggestions for future research

Despite its contributions, this study has several limitations. First, although the dataset is diverse, it is restricted to English-language reviews, limiting its applicability to the broader linguistic diversity of online reviews. Second, the study focuses exclusively on human-written fake reviews, excluding AI-generated fake reviews, which are becoming increasingly sophisticated and pose a growing concern. Third, while transformer models demonstrated superior performance, their computational complexity and resource demands may hinder their practical implementation in real-time systems, particularly for small businesses with limited technological infrastructure.

Future research should address several key areas to build on the findings and overcome the limitations identified in this study. These suggestions are presented in Table 10.

A key avenue for future research is the exploration of multilingual and cross-cultural datasets to enhance the generalizability of fake review detection. Expanding beyond English-language reviews would allow for a more comprehensive understanding of linguistic and cultural variations in deceptive content. Researchers should examine how psycholinguistic features differ across languages and cultural contexts and assess the effectiveness of transformer-based models in detecting fake reviews in multilingual settings. Addressing challenges in adapting these models for cross-linguistic applications will be crucial for developing more universally applicable detection systems.

Future studies should also focus on integrating psycholinguistic features with advanced NLP techniques to enhance detection accuracy. Combining traditional linguistic analysis with modern approaches—such as deep learning, ensemble methods, and hybrid models—can improve the ability to differentiate between fake and genuine reviews. Research should explore the synergy between psycholinguistics, contextual semantics, and deep learning architectures to identify the most effective strategies for deception detection.

Methodologically, longitudinal studies should be conducted to analyze the evolution of fake review patterns over time. Tracking how deceptive strategies change, particularly with the rise of AI-generated fake reviews, can provide insights into emerging fraud tactics. Additionally, evaluating the long-term effectiveness of detection algorithms will help in designing adaptive and resilient detection systems. Using larger and more diverse datasets will further strengthen the robustness of detection frameworks.

Expanding research beyond e-commerce to include diverse online platforms, such as social media, travel review sites, and service platforms, is another critical area. Each platform has unique user behaviors and review dynamics, which can impact the nature of fake reviews and the effectiveness of detection models. Understanding platform-specific characteristics will enable the development of versatile, context-aware detection techniques that can be applied across different online environments.

Interdisciplinary theoretical perspectives should also be incorporated into future research. Insights from psychology, sociology, and communication studies can deepen our understanding of social influence, deception strategies, and identity construction in fake reviews. Applying these frameworks can help develop nuanced detection models that account for the complex social and psychological processes underlying review manipulation.

Table 10 Suggestions for future research

Key areas	Main issues to focus on	Specific research questions
Multilingual and Cross-Cultural Analysis	Enhancing generalizability across languages; Investigating cultural variations in deception strategies	1. How do psycholinguistic features of fake reviews vary across different languages and cultural contexts? 2. Can advanced transformer models be effectively applied to multilingual datasets while accounting for cultural differences? 3. How do cultural norms influence deception strategies in fake reviews?
Integration of Psycholinguistic Features with Advanced NLP Techniques	Combining linguistic analysis with newer ML approaches; Exploring hybrid detection models	1. What are the most effective strategies for integrating psycholinguistic features with transformer-based NLP models? 2. How can hybrid models improve fake review detection accuracy? 3. What novel machine learning techniques can be incorporated into psycholinguistic-based detection frameworks?
Methodology: Longitudinal Studies	Examining the evolution of fake review patterns over time; Assessing AI-generated deception	1. How do fake review patterns evolve over time, particularly with the rise of AI-generated content? 2. What are the emerging trends and techniques used by fraudsters, and how can detection models adapt dynamically? 3. How resilient are current detection algorithms in maintaining accuracy against evolving deception tactics?
Contextual Expansion to Different Platforms	Analyzing fake reviews across diverse online environments; Studying unique platform-specific behaviors	1. What are the distinguishing characteristics of fake reviews on various online platforms (e.g., e-commerce, social media, travel sites)? 2. How should detection models be adapted to different platform architectures and user behaviors? 3. What platform-specific features can enhance the effectiveness of fake review detection?
Use of Interdisciplinary Theoretical Perspectives	Integrating insights from psychology, sociology, and communication studies; Applying theories related to social influence and deception	1. How can theories from psychology, sociology, and communication studies enhance our understanding of fake reviews? 2. What insights can social influence and communication theories provide for developing more robust detection models? 3. How can interdisciplinary theoretical frameworks be integrated into computational models for improved detection?
Practical Implementation of Transformer Models	Addressing trade-offs between accuracy, scalability, and explainability; Ensuring real-world applicability	1. What are the trade-offs between detection accuracy, computational efficiency, and scalability in real-time review monitoring systems? 2. How can explainable AI (XAI) techniques be incorporated to enhance transparency and trust in fake review detection? 3. What are the ethical considerations and challenges in deploying transformer-based fake review detection in commercial settings?

Finally, practical implementation remains a key challenge. Future research should address the trade-offs between detection accuracy, computational efficiency, and real-time applicability in monitoring systems. While transformer-based models have demonstrated superior performance, their high computational demands may limit widespread adoption. Investigating scalable and resource-efficient solutions—particularly for small businesses—will be essential. Additionally, integrating explainable AI (XAI) techniques can enhance transparency and user trust in detection systems. Collaboration with industry stakeholders to test and refine these models in operational environments will provide valuable insights into their feasibility and impact.

Acknowledgements Mekhail Mustak & Joni Salminen gratefully acknowledge the support of Liikesivistysrahasto (Foundation of Economic Education), Finland.

Funding Open Access funding provided by Hanken School of Economics. The first and the second authors acknowledge financial support from Liikesivistysrahasto (Foundation for Economic Education), Finland.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. The authors have no conflicts of interest to declare that are relevant to the content of this article. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no financial or proprietary interests in any material discussed in this article.

Use of generative AI During the preparation of this work the authors used ChatGPT 4o and Gemini Advanced (2.0 Flash) in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Abri, Faranak, Luis Felipe Gutierrez, Akbar Siami Namin, Keith S. Jones, and David R. W. Sears. 2020. Linguistic Features for Detecting Fake Reviews. In *Proceedings of the IEEE International Conference on Machine Learning Applications (ICMLA'20)*.

Alsubari, Saleh, Mahesh Shelke, and Sachin Deshmukh. 2020. Fake reviews identification based on deep computational linguistic features. *International Journal of Advanced Science and Technology* 29 (8): 3846–3856.

Ananthakrishnan, Uttara M., Beibei Li, and Michael D. Smith. 2020. A tangled web: should online review portals display fraudulent reviews? *Information Systems Research* 31 (3): 950–971.

Banerjee, Snehasish. 2022. Exaggeration in fake vs. authentic online reviews for luxury and budget hotels. *International Journal of Information Management* 62: 102416.

Banerjee, Snehasish, and Alton Y. K. Chua. 2021. Calling out fake online reviews through robust epistemic belief. *Information & Management* 58 (3): 103445.

Bell, Jason. 2020. *Machine learning: hands-on for developers and technical professionals*. Wiley.

Bruni, E., N.K. Tran, and M. Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49: 1–47.

Burtch, Gordon, Yili Hong, Ravi Bapna, and Vladas Griskevicius. 2018. Stimulating online reviews by combining financial incentives and social norms. *Management Science* 64 (5): 2065–2082.

Chen, Jianqing, Zhiling Guo, and Jian Huang. 2022. An economic analysis of rebates conditional on positive reviews. *Information Systems Research* 33 (1): 224–243.

Choi, Angela Aerry, Daegon Cho, Dobin Yim, Jae Yun Moon, and Oh. Wonseok. 2019. When seeing helps believing: the interactive effects of previews and reviews on e-book purchases. *Information Systems Research* 30 (4): 1164–1183.

Choi, HanByeol Stella., Oh. Wonseok, Chanhee Kwak, Junyeong Lee, and Heeseok Lee. 2022. Effects of online crowds on self-disclosure behaviors in online reviews: a multidimensional examination. *Journal of Management Information Systems* 39 (1): 218–246.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv.

Filho, Costa, Diego Nogueira Murilo, Lucia Salmonson Rafael, Guimarães Barros, and Eduardo Mesquita. 2023. Mind the fake reviews! Protecting consumers from deception through persuasion knowledge acquisition. *Journal of Business Research* 156: 113538.

Hancock, Jeffrey T., Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. 2007. On lying and being lied to: a linguistic analysis of deception in computer-mediated communication. *Discourse Processes* 45 (1): 1–23.

Hartmann, Jochen, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing* 36 (1): 20–38.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical learning with sparsity: the lasso and generalizations*. New York: Chapman and Hall/CRC.

He, Jiaxiu, Xin Wang, Mark B. Vandenbosch, and Barrie R. Nault. 2020. Revealed preference in online reviews: purchase verification in the tablet market. *Decision Support Systems* 132: 113281.

Kauffmann, Erick, Jesús Peral, David Gil, Antonio Ferrández, Ricardo Sellers, and Higinio Mora. 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.

Kumar, Naveen, Deepak Venugopal, Liangfei Qiu, and Subodha Kumar. 2018. Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems* 35 (1): 350–380.

Li, Jiwei, Myle Ott, Claire Cardie, and Eduard Hovy (2014), "Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1566–76.

Liu, Yang, Juan Feng, and Xiuwu Liao. 2017. When online reviews meet sales volume information: is more or accurate information always better? *Information Systems Research* 28 (4): 723–743.

Liu, Chunnian, Xutao He, and Lan Yi. 2024. Determinants of multimodal fake review generation in China's e-commerce platforms. *Scientific Reports* 14 (1): 8524.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019), "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv.

Luca, Michael, and Georgios Zervas. 2016. Fake it till you make it: reputation, competition, and yelp review fraud. *Management Science* 62 (12): 3412–3427.

McHugh, Mary L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22 (3): 276–282.

Mohawesh, Rami, Xu. Shuxiang, Son N. Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. 2021. Fake reviews detection: a survey. *IEEE Access* 9: 65771–65802.

Moon, Sangkil, Moon-Yong. Kim, and Dawn Iacobucci. 2021. Content analysis of fake consumer reviews by survey-based text categorization. *International Journal of Research in Marketing* 38 (2): 343–364.

Mukherjee, Arjun, Vivek Venkataraman, Bing Liu, and Natalie Glance 2013a, What yelp fake review filter might be doing?. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Mukherjee, Arjun, Vivek Venkataraman, Bing Liu, and Natalie Glance 2013b, Fake review detection: Classification and analysis of real and pseudo reviews, *UIC-CS-03–2013. Technical Report*.

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

Narayan, Rohit, Jitendra Kumar Rout, and Sanjay Kumar Jena. 2018. Review spam detection using opinion mining. In *Progress in intelligent computing techniques: theory, practice, and applications. Advances in intelligent systems and computing*, ed. P.K. Sa, M.N. Sahoo, M. Murugappan, Y. Wu, and B. Majhi, 273–279. Singapore: Springer.

Ordabayeva, Nailya, Lisa A. Cavanaugh, and Darren W. Dahl. 2022. The upside of negative: social distance in online reviews of identity-relevant brands. *Journal of Marketing* 86 (6): 70–92.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. "Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint* arXiv:1107.4557.

Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, 497–501.

Park, Sungsik, Woochoel Shin, and Jinhong Xie. 2023. Disclosure in incentivized reviews: does it protect consumers? *Management Science* 69 (11): 7009–7021.

Petrescu, Maria, and Anjala S. Krishen. 2024. Marketing analytics in 2024 conferences: AI and data-driven decision-making. *Journal of Marketing Analytics* 12 (4): 743–745.

Plotkina, Daria, Andreas Munzel, and Jessie Pallud. 2020. Illusions of truth—experimental insights into human and algorithmic detections of fake online reviews. *Journal of Business Research* 109: 511–523.

Sahoo, Nachiketa, Chrysanthos Dellarocas, and Shuba Srinivasan. 2018. The impact of online product reviews on product returns. *Information Systems Research* 29 (3): 723–738.

Salehi-Esfahani, Saba, and Ahmet Bulent Ozturk. 2018. Negative reviews: formation, spread, and halt of opportunistic behavior. *International Journal of Hospitality Management* 74: 138–146.

Salminen, Joni, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-Gyo. Jung, and Bernard J. Jansen. 2022. Creating and detecting fake reviews of online products. *Journal of Retailing and Consumer Services* 64: 102771.

Schoenmueller, Verena, Oded Netzer, and Florian Stahl. 2020. The polarity of online reviews: prevalence, drivers and implications. *Journal of Marketing Research* 57 (5): 853–877.

Shoham, Meyrav, Sarit Moldovan, and Yael Steinhart. 2017. Positively useless: irrelevant negative information enhances positive impressions. *Journal of Consumer Psychology* 27 (2): 147–159.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37 (2): 267–307.

Tausczik, Yla R., and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29 (1): 24–54.

Venkatakrishnan, Jeeva, Ravikumar Alagiriswamy, and Satyanarayana Parayitam. 2024. Disentangling the relationship between trust, online buying, and customer satisfaction: A three-way interaction model. *Journal of Marketing Analytics* 12 (4): 806–828.

Vrij, Aldert. 2000. *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice, Wiley series in psychology of crime, policing and law*. Chichester: Wiley.

Wang, Xiaoli, Chenxi Zhang, and Xu. Zeshui. 2024. A product recommendation model based on online reviews: improving PageRank algorithm considering attribute weights. *Journal of Retailing and Consumer Services* 81: 104052.

Wu, Yuanyuan, Eric WT. Ngai, Wu. Pengkun, and Wu. Chong. 2020. Fake online reviews: literature review, synthesis, and directions for future research. *Decision Support Systems* 132: 113280.

Yoo, Kyung-Hyan., and Ulrike Gretzel. 2009. Comparison of deceptive and truthful travel reviews. In *Information and communication technologies in tourism 2009*, ed. W. Höpken, U. Gretzel, and R. Law, 37–47. Vienna: Springer.

Zhang, Dongsong, Lina Zhou, Juan Luo Kehoe, and Isil Yakut Kilic. 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.

Zhang, Dong, Wenwen Li, Baozhuang Niu, and Wu. Chong. 2023. A deep learning approach for detecting fake reviewers: exploiting reviewing behavior and textual information. *Decision Support Systems* 166: 113911.

Zhao, Jun, and Hong Wang. 2016. Detection of fake reviews based on emotional orientation and logistic regression. *CAAI Transactions on Intelligent Systems* 11 (3): 336–342.

Joni Salminen holds a Ph.D. in Marketing from Turku School of Economics and is working as an Associate Professor (tenure track) at the University of Vaasa, Finland. His research focuses on using machine learning for marketing applications, such as automatic persona generation and classification of social media content.

Mekhail Mustak is an Assistant Professor of Marketing (tenure track) at the Hanken School of Economics, Finland. He is also an adjunct professor at the University of Eastern Finland. His research focuses on the application of artificial intelligence in marketing, B2B marketing, and services marketing. Before joining academia, he was a Senior Executive at A.P. Moller–Maersk, where he was involved in the international

supply chain management for Nike, Puma, JCPenney, and Tesco Stores. Mekhail is also a mountaineer.

Soon-Gyo Jung, MSc is a Research Associate at the Qatar Computing Research Institute working in the area of computational social science. He has a background in web applications and software development and holds a master's degree in Electrical and Computer Engineering from Sungkyunkwan University in South Korea. He has published several articles in areas including information dissemination and audience segmentation.

Hannu Makkonen is a Professor of Marketing at the School of Marketing and Communication at the University of Vaasa. His research interests lie in the areas of innovation management, innovation ecosystems, and value creation logics in industrial networks and relationships. His previous work has been published in e.g. Technology Analysis & Strategic Management, Industrial Marketing Management, Journal of Business Research, Journal of Business & Industrial Marketing, Management Decision, Marketing Theory, Journal of Business Market Management, and Journal of Financial Services Marketing.

Bernard J. Jansen is a Principal Scientist in the artificial intelligence center of the Qatar Computing Research Institute. He is a graduate of West Point and has a Ph.D. in computer science from Texas A&M University. Professor Jansen is editor-in-chief of the journal Information Processing & Management (Elsevier).