Information Sciences 638 (2023) 118953

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Measuring semantic gap between user-generated content and product descriptions through compression comparison in e-commerce

Carlos A. Rodriguez-Diaz [b], Sergio Jimenez [a], Daniel Bejarano [a], Julio A. Bernal-Chávez [a], Alexander Gelbukh [b,*]

[a] Instituto Caro y Cuervo, Bogotá, Colombia
[b] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico City, Mexico

ARTICLE INFO

Keywords:
Lexical-semantic gap
Customer-vendor communication
Semantic drift
Social commerce language
Product reviews ambiguity
Product description ambiguity

ABSTRACT

The significance of user-generated content as a source for business intelligence and analytics has been on the rise since the inception of electronic commerce platforms and has been solidified in the wake of the pandemic due to the prominence of electronic commerce as a sales channel. The prevailing approach to harnessing unstructured data involves the utilization of Artificial Intelligence; however, there exist simpler alternatives capable of yielding valuable information. This article introduces a methodology grounded in information theory to quantify the semantic disparity between the consumer community and product descriptions.

This disparity can result in potential misunderstandings in the dialogue among consumers, and incidental costs in the dialogue between consumers and vendors. One plausible explanation for this disparity is that the terminology employed by consumers may possess different meanings compared to that utilized by product description writers. Our methodology employs large corpora of consumer reviews and product descriptions to quantify this semantic disparity across multiple electronic commerce domains through the implementation of random word exchanges and compression. Furthermore, we utilize neural word embeddings to identify specific words exhibiting the greatest semantic drift between reviews and descriptions, thereby providing lexical examples of these gaps. Our findings indicate that lower levels of lexical-semantic gap are associated with better consumer satisfaction.

1. Introduction

Electronic commerce can be considered as business transactions facilitated through a dialogue concerning products on the Internet between suppliers and consumers. Initially, this dialogue is unidirectional, with the supplier providing multimedia information to the consumer regarding products description. This information is utilized by the consumer to make informed purchasing decisions and establish expectations [16].

Conversely, in social e-commerce, consumers engage in dialogue with one another to exchange experiences pertaining to the products and services they have purchased or intend to purchase [36,49,27]. Within these dialogues, written language is of paramount importance in product descriptions, consumer reviews, consumer-to-consumer conversations, customer service and negotiations [21].

* Corresponding author.
E-mail address: gelbukh@cic.ipn.mx (A. Gelbukh).

https://doi.org/10.1016/j.ins.2023.118953
Received 21 November 2022; Received in revised form 11 April 2023; Accepted 13 April 2023

Available online 19 April 2023

The function of language in electronic commerce can be likened to that of Plain Language, a discipline dedicated to diminishing the linguistic disparity between public administration and citizens [1]. The challenges encountered in Plain Language are analogous to those that arise between consumers and suppliers. Firstly, the population of citizens greatly exceeds that of individuals responsible for writing public communications. Secondly, specialized language is often employed as administrators, technocrats, lawyers or politicians are tasked with communicating with citizens. Thirdly, the sociodemographic and linguistic differences between the two groups significantly impact the efficacy of the communication process. Moreover, research has demonstrated [1] that readiness was among the foremost research clusters from 2002 to 2012 in the e-government domain. The same motivations that render Plain Language desirable in public administration are applicable to the electronic commerce sphere; however, minimal research has been conducted on this subject [24].

It is commonly assumed that consumers possess a comprehensive understanding of the information provided by the supplier; however, there exist several reasons why this may not be the case. Firstly, the population of users is typically larger and more diverse than that of individuals responsible for writing product descriptions [35]. Secondly, depending on the domain, product descriptions may contain specialized language to varying degrees that may not be fully comprehended by users [46]. Thirdly, due to the trend towards the globalization of commerce, the sociodemographic and linguistic differences between consumers and suppliers can be substantial [7]. These factors, in conjunction with others such as the inherently ambiguous nature of language, and the widespread utilization of automatic translators [5], suggest that this issue may impact e-commerce. Nevertheless, this issue has received little attention from the scientific community.

Within a specific e-commerce domain, product descriptions and user reviews serve distinct purposes and are consequently dissimilar in nature. The former constitutes the initial means of communication between the supplier and consumer, while the latter represents communication among consumers. Moreover, there exists a general consensus that high-quality descriptions and favorable reviews are advantageous to the business. Both types of text address the same topics and share a considerable portion of vocabulary. If semantic issues arise within this shared vocabulary, the objectives of these texts may be hindered. There exist several scenarios in which a potential lexical-semantic gap between the language of users and that of suppliers can give rise to problematic situations. Firstly, if descriptions are accurate and comprehensive but users fail to recognize the terminology or meanings employed, then the purpose of the descriptions may be compromised. Secondly, if consumer expectations regarding a product are erroneous due to this semantic gap, consumer satisfaction may be jeopardized. Thirdly, consumers engaging in social e-commerce may encounter misunderstandings that undermine the efficacy and benefits of this social mechanism. The identification, quantification, and characterization of this gap represent the initial step towards providing potential solutions to mitigate these problematic situations.

From a marketing standpoint, the correlation between consumer satisfaction and business benefits has been well-established [41] and recently reaffirmed in e-commerce scenarios [28]. Theories aimed at modeling and elucidating consumer satisfaction are primarily predicated on economic and behavioral factors pertaining to users and businesses [39]. Conversely, empirical studies have associated consumer satisfaction with factors such as service quality [8], logistics [20], and the fulfillment of consumer expectations [16]. The linguistic aspects of consumer satisfaction in e-commerce have been examined from the perspective of analyzing the content of user reviews; that is, identifying which words and expressions denote satisfaction or dissatisfaction in consumer discourse [48]. This user-centric approach is most advantageous for consumer profiling, recommendation, and product development [18]. However, few studies have investigated language usage from the perspective of providers in a holistic manner, and the potential factors associated with consumer satisfaction, such as semantics, style, readiness, and complexity [38]. Given that language usage by consumers represents only an observable factor while language usage on the part of providers constitutes a relatively controllable factor, the potential relationships between the latter and consumer satisfaction would offer managerial perspectives to benefit businesses.

From a linguistic perspective, *semantic drift* refers to the alteration or divergence in the meaning of words between two populations distinguished by usage, dialect, time, demography, nature, or purpose [6]. A word exhibiting semantic drift in e-commerce has a predominant sense usage by consumers that differs from its usage in textual content produced by suppliers. For instance, within the domain of sports products, for consumers, the word ‘core’ denotes a group of muscles surrounding the human torso, whereas in product descriptions, the predominant meaning refers to the internal material or component from which a specific product is constructed. While both parties may be cognizant of the dual senses and may disambiguate depending on context, if the preferred meanings of each party differ significantly, it is conceivable that this factor may become a source of miscommunication. Furthermore, if numerous words within a domain exhibit this behavior, the efficacy of written communication may be impacted.

The manual identification of relevant words exhibiting semantic drift within a specific domain represents a overwhelming task for linguists and lexicographers due to the diversity and specificity of domains and the vast quantity of texts requiring analysis. Concordance analysis enables linguists to compare the contexts of usage of a word within two sets of texts. However, even when utilizing concordancing software, given that the size of a community’s vocabulary is often substantial, the detection of words exhibiting semantic drifts constitutes a tedious, overwhelming, and costly undertaking.

Alternatively, automatic approaches have been explored to identify and quantify semantic drift. These approaches are based on the notion of ascertaining the meaning of words through the linguistic distributional hypothesis, which asserts that ‘you shall know a word by the company it keeps’ [15]. To this end, it is typically necessary for each speech community to be represented by a corpus of considerable size in order to identify statistically significant differences between the contexts of usage of the words. Traditional methods for achieving this include the vector space model, random indexing, latent semantic indexing, and others, which have demonstrated relative efficacy but possess limitations in terms of vocabulary length and the quantity of texts supported [42]. Recently, there has been a paradigm shift in natural language processing towards the utilization of neural networks for the semantic analysis of words [45]. Specifically, *word embeddings* [29] facilitate the acquisition of vector representations for words

within geometric spaces possessing semantic properties of similarity, analogical and compositional reasoning. Word embeddings surmount the limitations of previous methods, enabling large quantities of text (billions of words) to be processed efficiently.

The drift of dialects, whether diachronic or diatopic, frequently encompasses aspects of a phonetic, morphological, lexical, syntactic, and semantic nature. Among these, phonetics exhibits the lowest incidence in the written modality [13]. Morphological variation is often associated with alterations in declension or marking, with relatively minor incidence in the semantic aspect. With regard to syntactic variation, these arise from changes in the order of phrase or sentence constituents, causing semantic differences in selected cases [4]. Conversely, lexical aspects possess a greater incidence in semantics, generally entailing the alteration of meaning or the use of new words [6]. Additionally, word frequency represents a differentiating factor between dialects or speech communities [31]. These differences have proven useful in identifying words that characterizes groups [43]. However, particularly in gender studies, there exists debate as to whether or not this difference impacts communication [9]. Therefore, among these factors, lexico-semantic variation possesses significant potential to affect communication between sociolects and technolects.

The textual data employed in this study endeavors to be representative of e-commerce and social commerce scenarios. Product reviews authored by consumers constitute the fundamental mechanism through which the consumer community disseminates information regarding their shopping experiences [50]. We will utilize these reviews under the assumption that other forms of communication among consumers in social commerce possess analogous linguistic features. Additionally, we will employ the textual descriptions of products to represent the language of suppliers, assuming that they share analogous linguistic features with other textual components on their web pages and in other communications between consumers and vendors [38].

However, there exists a significant limitation in the e-commerce scenario with regard to the application of approaches based on word embeddings. While the quantity of textual content generated by users in their social interactions is substantial and suitable to this approach, the quantity of text describing products is considerably smaller. This imbalance results in the production of high-quality vector representations for words within the reviews corpus, while those obtained from product descriptions yield noisy vectors.

In this study, we propose a novel approach to measure the semantic gap between two highly size-unbalanced corpora by utilizing Shannon's information theory. Our method employs a regular file compression algorithm (i.e., bzip2) and a randomization mechanism to combine reviews and descriptions. The compressed sizes of this combined corpora are then used to quantify the semantic gap. We evaluated our approach on a large dataset comprising 28 product domains, where each review is accompanied by a rating ranging from 1 to 5. To validate our method, we applied word embeddings to both the large corpus of product reviews and the small corpus of product descriptions for each domain. This resulted in a set of words exhibiting considerable semantic drift but with a high error rate due to the marked imbalance in corpus size. We manually identified the words with true semantic drift by analyzing their nearest neighbors, and used these as ground truth to assess the validity of our compression-based method. Our experiments explored the potential relationship between the lexical-semantic gap and customer satisfaction as inferred from ratings.

This study makes contributions to both the field of e-commerce and computational linguistics. In the context of e-commerce, our research elucidates the relationship between the degree of semantic gap and lexical ambiguity between customers and vendors, and customer satisfaction. In the field of computational linguistics, we introduce a conceptually simple yet robust method to measure this gap despite imbalances in the data. Further details on these contributions and other perspectives of this study are discussed in Sections 6.3 and 6.4.

The remainder of this article is structured as follows: Section 2 provides a review of related work on the analysis of product reviews and descriptions and their relationship to customer satisfaction. Additionally, we introduce background concepts for readers unfamiliar with computational linguistics to facilitate their understanding of our hypotheses and methods. Section 3 describes the methods and data used in our study. In Section 4, we present our results and evaluations. Section 5 discusses these results in detail. Finally, Section 6 concludes the article by summarizing our findings, discussing their implications, and outlining future research directions.

2. Literature review

This section aims to provide context for this study. Firstly, subsection 2.1 reviews previous research that has examined the potential relationships between the textual content of product descriptions and user reviews and commercial factors such as customer satisfaction and profit. Subsections 2.2 and 2.3 provide context for the methods presented in subsection 2.4 (Materials and Methods), allowing for a better understanding of these methods and the study's hypothesis. Finally, subsection 2.4 presents the hypothesis of this study.

2.1. Related work

Francis et al. (2002) [16] conducted a survey of e-commerce customers and found that the quality of information in product attribute descriptions positively influences customer satisfaction. Similarly, Ludin (2014) [28] discovered a significant positive relationship between the quality of product information and customer satisfaction, implying a transitive relationship with e-loyalty. They concluded that high-quality information assists customers in making better purchasing decisions, leading to increased satisfaction. Patrada (2021) [36] reaffirmed these findings, demonstrating the perceptible effect of product information quality on customer satisfaction. It is assumed that product information includes textual content, where "quality" refers to well-written, informative, complete, accurate, precise, and coherent text. Mou (2019) [33] also found positive relationships between the quality of written descriptions and cognition-related involvement, product affectivity, and customer platform. In conclusion, the aforementioned studies represent a sample of the evidence supporting the positive effect of high-quality written content related to products on e-commerce. However,

current evidence is limited to the informational content of descriptions; there is no evidence linking linguistic characteristics of descriptions to customer reactions.

In the field of artificial intelligence, Pryzant et al. (2017) [38] provided evidence of a strong correlation between product descriptions and sales. They trained a deep learning neural network on sequences of words in product descriptions to identify lexical patterns associated with sales. They found that, even after controlling for confounding factors such as product type, brand, and price, textual descriptions were highly predictive of sales, with an R2 value of approximately 0.80. While this does not establish a causal relationship, it supports previous research in the field of marketing that has demonstrated the positive impact of high-quality product descriptions on sales [26,22].

Several studies have examined the relationship between product descriptions and customer reviews. Wang (2018) [47] found that for certain product characteristics in descriptions, it is possible to automatically identify related emotional factors in customer reviews to develop products with affective responses. In response to the issue of inadequate or low-quality product descriptions in some domains, researchers have explored the automatic generation of product descriptions from customer reviews [35,10]. These descriptions, derived from crowd-sourced material, have proven to be reliable content for customers [12]. However, despite the sound rationale for these initiatives, there is no empirical evidence to support the claim that these synthetic product descriptions have a positive effect on customer satisfaction.

In light of these findings, the present study aims to enhance our understanding of the semantic textual features in product descriptions and their potential impact on customer satisfaction. Specifically, we will investigate the degree of semantic alignment between the meanings of words used in product descriptions and those used in customer reviews. This meta-linguistic characteristic is independent of the language or vocabulary used and is associated with a universal aspect of human language: ambiguity. In contrast, previous research has identified specific words in the English language, within particular domains, that are predictive of sales or satisfaction [38]. However, these results are difficult to generalize to other languages or domains. Our contribution is to provide a language-agnostic method for measuring the semantic gap between descriptions and reviews, and to empirically test the hypothesis that this gap is related to customer satisfaction.

2.2. Corpus information by compression

Cilibrasi and Vitanyi (2003) [11] demonstrated that the Kolmogorov complexity of a data file, defined as the size of its theoretical minimum compressed version, can be obtained using a theoretical normal compressor. This compressor, denoted as C, must satisfy the properties of *idempotency* $C(xx) = C(x)$, *monotonicity* $C(xy) \geq C(x)$, *symmetry* $C(xy) = C(yx)$, and *distributivity* $C(xy) + C(z) \leq C(xz) + C(yz)$. Most file compressors, such as zip, rar, and bzip2, asymptotically approximate the properties of a normal compressor and thus Kolmogorov complexity. As a result, this notion of information is associated with the size of the compressed data file and is applicable to any type of data. It is particularly useful for large textual corpora.

A *lossless* file compressor analyzes a sequence of numerical data to identify patterns and recognize redundancy preserving all information. This allows the compressor to "understand" the data sequence and reproduce it from a condensed version. As a result, a sequence of random signs is not compressible, while any sequence with some degree of order can be compressed to some extent. Measuring the compressed size of a corpus is useful for quantifying the amount of information in a specific feature of the corpus. For instance, consider a textual corpus x of English text with an uncompressed size of $|x|$ bytes and a compressed size of $C(x)$ bytes. Let us create a variant x' by replacing all occurrences of the word "that" with "this". The uncompressed size $|x|$ is equal to $|x'|$, but $C(x) > C(x')$ because x contains more information than x' due to the removal of one word from its vocabulary. The difference $\delta = C(x) - C(x')$ represents the amount of information conveyed by the word "that" in corpus x, which is the measured feature in this example.

Montemurro and Zanette (2011) [32] applied the principle of compression to measure the amount of information associated with the order of words in a corpus. They created a variant of an original corpus by shuffling the words within each sentence, resulting in two corpora of identical size. The difference between the compressed sizes of these two corpora represents the amount of information associated with the order of words. They discovered that this difference is universal across different language families. The effectiveness of this method relies on the relationship between the factor being randomized and the linguistic feature it represents. In this case, word order serves as a proxy for syntax, meaning that when words are shuffled within sentences, any notion of syntax is lost. As a result, the difference in the amount of information (in an information-theoretic sense) between the randomized corpus (i.e., without syntax) and the original corpus measures the amount of information due to syntax. Although syntax is a multidimensional factor, this unidimensional measure quantifies not syntax itself but rather the information due to syntax. One advantage of this approach is that if the relationship between the factor being randomized and the linguistic feature is well established, it is intuitive that other factors do not influence the measurement.

In this study, we present a method that manipulates text corpora produced by customer and provider speaking communities. The difference in size between the two versions of the corpora represents the semantic gap between the two communities. Our randomization mechanism is based on the concept of synonymy as defined by Miller (1990) [30], which states that two words are synonyms if they are interchangeable in many contexts. Extending this idea, two semantically equivalent words should be interchangeable in any context. To test the semantic equivalence of two words using a corpus, their occurrences can be randomly swapped without affecting the semantic content of the corpus. Consequently, the compressed size of the corpus should remain unchanged or change minimally. If it does change significantly, this difference serves as a quantitative measure of the semantic difference between the two words. Thus, we use random interchangeability as a proxy for semantic difference, analogous to the random shuffling of words within sentences being used as a proxy for syntax.

2.3. Measuring word drifts by word embeddings

Neural word embeddings are a set of techniques for obtaining vector representations of words in a high-dimensional Euclidean space [45,29,19]. Given a large textual corpus as input, the output is a matrix of size $n \times m$ where each row represents a vector of dimension m associated with each of the n words in the corpus vocabulary. In this m-dimensional space, each word is assigned coordinates that reflect its semantic properties. The primary characteristic of this space is that the distance between words represents their degree of similarity or semantic relationship. For example, words such as 'car' and 'automobile' would be positioned close to each other, while 'car' and 'cucumber' would be farther apart. Another property of this semantic space is compositionality, whereby the vector sum of the representations of words such as 'doctor' and 'heart' would be close to that of 'cardiologist'. Additionally, this space enables analogical reasoning. As a prototypical example, consider the relative distance and direction between words such as 'king' and 'queen', which would be similar to those between 'prince' and 'princess'. This similarity enables the solution of analogies such as 'king'→'queen': 'prince'→X. These word representations are obtained from the weights of connections between neurons in an artificial neural network. The network is trained on a word co-occurrence prediction task using a large corpus as training data.

The semantic drifts of words can be determined by geometrically comparing their vector representations obtained using two large corpora representing distinct speech communities or dialects. Word drift detection and the complementary task of identifying semantic equivalences can be performed using word embeddings [40]. Given two corpora, corpus A and corpus B, representing distinct speech communities or dialects as input, the output is a single semantic Euclidean space where words are labeled with an identifier indicating their source corpus. The distances between pairs of labeled words in this space are then compared (e.g., $house_A$ and $house_B$). Words that are close to their pair are considered semantic equivalences (i.e., 'house' has the same meaning in both A and B), while the farthest pairs are considered semantic drifts. For example, if corpus A contains real estate descriptions (where 'house' refers to a building) and corpus B contains descriptions of luxury items (where 'house' refers to a brand), the word 'house' exhibits potential semantic drift. Approaches for obtaining a single semantic space from two corpora include geometric alignment [3], conditional probability [19], and meta-algorithms applied to current word embedding algorithms [40].

To evaluate the quality of results for words identified as exhibiting semantic drift, a gold standard is required. This is typically constructed manually by professional linguists and often produces only a limited number of instances [17]. For instance, in the DIACR-Ita challenge [2], only 18 words were available in the gold standard for detecting semantic drift in diachronic corpora (i.e., one corpus from the past and one from the present). In this challenge, Angel et al. (2020) [2] demonstrated that it is not necessary to obtain a single semantic space to identify semantic drifts; this can be achieved using word embeddings obtained separately for each corpus. To determine whether a word exhibits semantic drift, its corresponding sets of k neighboring words in each semantic space are collected and compared using the Jaccard index. Target words with the lowest index values exhibit the highest degree of semantic drift. This approach yielded good results and was robust to imbalances in corpus size, such as when there are few texts from the past and many from the present. In this study, we apply this idea to address imbalances between the amounts of text produced by customers and product descriptions.

2.4. Research question and hypothesis

This study aims to propose a method for measuring the semantic gap between the language used by customers and suppliers in an e-commerce context. We seek to determine whether such a gap exists and to quantify its magnitude across different market domains. Additionally, we aim to investigate whether this gap is related to customer satisfaction. Our research question and hypothesis are as follows:

- RQ: In which e-commerce domains is the lexical-semantic gap between vendors (i.e., product descriptions) and customers (i.e., user reviews) the largest and smallest?
- Hypothesis (H): Within a set of e-commerce domains, larger lexical-semantic gaps are linked to lower levels of customer satisfaction.
- Null Hypothesis (H_0): There is no relationship between the lexical-semantic gap and customer satisfaction.

3. Materials and methods

The methodology presented in this section aims to achieve two objectives. First, to introduce a novel method for measuring the degree of the semantic gap between the language used by customers and suppliers. Second, to provide a framework for testing the hypothesis that the magnitude of this gap is related to customer satisfaction in e-commerce. To accomplish these objectives, a substantial amount of representative text is required, categorized at the domain level to enable observation and analysis of variations across domains. Additionally, to evaluate our hypothesis, associated metadata providing quantitative information on customer satisfaction (e.g., customer ratings) is necessary. The primary methodological challenge is designing a measure of the semantic gap for highly unbalanced data, such as the large volume of customer-generated product reviews compared to the limited quantity of product descriptions available. Since established methods (e.g., neural word embeddings) require a relatively balanced corpus, we propose an information theory-based measure of the semantic gap. To validate this method, we applied neural word embeddings, widely recognized for their suitability for this task on balanced data, and performed manual curation of the results to construct a gold standard. By comparing this gold standard with the results obtained using our proposed method, we aim to validate and obtain

Table 1
Word count for each domain and the ratio of word counts between reviews and descriptions.

Category	Reviews' words	Descriptions' words	Word ratio
All Beauty	13,454,743	1,285,534	10:1
Amazon Fashion	22,235,491	5,919,402	4:1
Appliances	19,211,478	4,566,216	4:1
Arts Crafts & Sewing	74,693,032	27,578,534	3:1
Automotive	226,834,970	161,065,374	1:1
CDs & Vinyl	386,158,660	21,973,448	18:1
Cell Phones & Accessories	349,903,420	72,093,498	5:1
Clothing Shoes & Jewelry	866,541,485	329,871,943	3:1
Digital Music	55,810,053	1,554,864	36:1
Electronics	1,024,197,307	135,120,799	8:1
Gift Cards	2,372,944	84,021	28:1
Grocery & Gourmet Food	149,018,135	26,804,856	6:1
Home & Kitchen	759,164,935	192,811,582	4:1
Industrial & Scientific	54,799,737	16,627,174	3:1
Kindle Store	382,079,930	3,420,666	112:1
Luxury Beauty	21,696,560	1,040,486	21:1
Magazine Subscriptions	3,610,776	133,637	27:1
Movies & TV	469,363,734	18,061,999	26:1
Musical Instruments	69,072,246	18,739,313	4:1
Office Products	191,650,532	49,943,042	4:1
Patio Lawn & Garden	186,284,611	45,539,118	4:1
Pet Supplies	269,510,943	32,330,942	8:1
Prime Pantry	10,415,289	1,048,980	10:1
Software	33,571,934	4,916,812	7:1
Sports & Outdoors	468,858,744	91,971,650	5:1
Tools & Home Improvement	327,711,025	107,013,154	3:1
Toys & Games	273,351,609	82,249,226	3:1
Video Games	179,005,493	11,711,784	15:1

a fully automated approach to this task. Finally, we will analyze gap measures qualitatively and quantitatively to compare them with customer satisfaction scores.

3.1. The data

We utilized the *Amazon Review Data 2018* [34], which comprises approximately 180 million reviews and 12 million product descriptions distributed across 28 domains. We used the text of the reviews and extracted text from the product data fields *title*, *tech1*, *description*, *feature*, and *similar item*. Each review includes a rating on an ordinal-discrete scale *ranging from* 5 (best) to 1 (worst), *which* we used as a quantitative indicator of customer satisfaction. Sentence boundaries were identified using the Natural Language Toolkit,[1] and rows from tabular descriptions were treated as sentences. All non-alphabetical tokens were subsequently removed. The sizes of the resulting review and description corpora for each domain are shown in Table 1.

3.2. Semantic gap by compression

Inspired by Montemurro and Zanette (2011) [32], we produced two versions of the combined reviews and description corpora for each domain, both of equal size. One version contained the original sentences (*True* corpus), while the other featured random word shifts (*Rand* corpus). These shifts were designed such that the difference between the compressed sizes of the two corpora would reflect the semantic gap between customers and product description writers. In computational linguistics, two words can be considered synonyms if they are interchangeable in many contexts [30]. The Rand corpus represents the null hypothesis that a particular word w (or set of words) has the same or similar meaning in both reviews and descriptions because random interchanges should not affect the null hypothesis of no effect due to context interchangeability. If there is little or no difference in the compressed sizes of the two corpora, the null hypothesis is accepted.

First, for each domain, we created a mixed corpus of reviews and descriptions by alternating between inserting a sentence from each source. This approach was motivated by the fact that most current compressors are stream-based and incrementally encode reviews and descriptions data, favoring the distributivity property of compression (see subsection 2.2). Since the reviews corpus contains more sentences than the descriptions corpus, we repeated from the first description as many times as necessary once all descriptions had been inserted. Next, we selected the vocabulary W to be tested for semantic drift by removing from the common vocabulary between reviews and descriptions the 500 most frequent words and words that appeared fewer than 50 times. This was motivated by the fact that the most frequent words are typically functional words with little or no ambiguity. Similarly, at the other end of the frequency spectrum, words with few occurrences provide limited evidence for any automatic analysis.

[1] Natural Language Toolkit at https://www.nltk.org/.

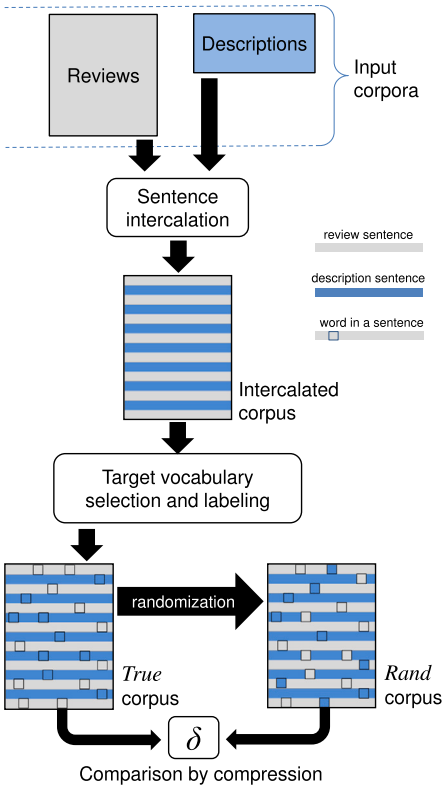

Fig. 1. Diagram illustrating the process for obtaining the semantic gap measure δ by combining and randomizing input corpora of user reviews and product descriptions for a specific domain.

Next, for each word w in W, we labeled each occurrence as w_R if w appeared in a review and w_D if it appeared in a description. We refer to this as the *True* corpus. To produce from it the *Rand* corpus, we randomly swapped the labels R and D on each target word with a probability of $P = 0.5$ (i.e., by 'flipping a coin'). This intermediate value was justified by analyzing the extreme values of $P = 0$, which results in no changes, and $P = 1$, which results in a complete label exchange that also produces no changes. The motivation for this randomization is that if w has the same meaning in both reviews and descriptions, then w_R and w_D can be interchanged in any context. Conversely, if w_R and w_D have different meanings, random exchanges would place words in unexpected contexts, increasing the corpus's information content. Fig. 1 illustrates this process graphically: gray represents text from reviews, blue represents text from descriptions, horizontal stripes represent sentences, small squares represent target words, and the fill color of these squares indicates their labels. Finally, we measured the size of the two compressed corpora using the *bzip2*[2] compressor, the same tool used by Wikipedia administrators to store and distribute backups of their encyclopedia.

Table 2 presents an example of the *True* and *Rand* corpora for the target word 'nature' in the Electronics domain. These corpora result from alternating between inserting a sentence from the review corpus and one from the description corpus. In the *True* corpus, the tags of the target word 'nature' match the source corpus of the sentence, while in the *Rand* corpus, tags are randomly assigned. If the target word is used in the same sense in both corpora, most contexts would be similar and there would be little difference between the compressed sizes of the *True* and *Rand* corpora. Otherwise, contexts would differ and random shifts would result in the tagged target word being placed in many inappropriate contexts, increasing the information content of the *Rand* corpus. In the example shown in Table 2, only a single target word is considered, but this tagging process is performed simultaneously for all target words that are neither very frequent nor very rare in the corpus. In particular, the word 'nature' exhibits drift in the Electronics domain. Its dominant meaning in descriptions is "the outer world of living beings", while in reviews it refers to "the basic character or quality". However, in other domains such as CD & Vinyl, 'nature' is associated with idealism, spirituality, and serenity in both reviews and descriptions, i.e., no drift.

Following Montemurro and Zanette (2011) [32], the relative difference in information content between the *True* and *Rand* corpora can be approximated using the sizes of the compressed files:

$$\delta = \frac{C(Rand) - C(True)}{C(Rand)}$$

[2] http://www.bzip.org/.

Table 2
Example of the *True* and *Rand* mixed review/description corpora for the sample word 'nature'. A R label is attached to target words from the review corpus and a D label is attached to target words from the description corpus.

True label	True corpus	Rand. label	Random corpus
R	That is the nature_R of digital.	D	That is the nature_D of digital.
D	Improve your mood in nature_D euphony.	D	Improve your mood in nature_D euphony.
R	The nature_R of wireless is noise.	R	The nature_R of wireless is noise.
D	Bushnell nature_D view plus x compact binocular.	R	Bushnell nature_R view plus x compact binocular.
R	Just the nature_R of the beast.	D	Just the nature_D of the beast.
D	Timex cd clock radio with nature_D sounds	R	Timex cd clock radio with nature_R sounds.
R	Love the retractable nature_R of this cord.	R	Love the retractable nature_R of this cord.
D	Separate nature_D pictures for each individual verse.	R	Separate nature_R pictures for each individual verse.
R	I rarely use the nature_R sounds.	D	I rarely use the nature_D sounds.
D	Used and praised by nature_D photographers.	R	Used and praised by nature_R photographers.
R	That is just the nature_R of privacy filters.	D	That is just the nature_D of privacy filters.
D	Perfect for nature_D sports and surveillance photos.	D	Perfect for nature_D sports and surveillance photos.

Here, δ is always positive because $C(Rand) > C(True)$. The measure δ is scaled by $C(Rand)$ in the denominator to enable comparison of its values for different corpus sizes of *True* and *Rand*. This is relevant in our scenario, as shown in Table 1, where several domains exhibit considerable variations in corpus data sizes. Thus, δ represents an approximation of the amount of information attributable to the aggregate semantic difference of the words being tested for semantic drift. This is because the randomization applied to the *Rand* corpus removes any notion of differential context from the words being tested. However, another factor that affects our method is variation in the size imbalance between the review and description corpora. Since the descriptions corpus is repeated multiple times to match the size of the review corpus, variations in redundancy can affect δ through the compression ratio of *True*. We anticipate that the final semantic gap measure should depend on both δ and the compression ratio defined as $cr(True) = \frac{|True|}{C(True)}$, where $|True|$ is the size of the uncompressed *True* corpus. In Section 4, we will adjust the measure δ by this factor based on experimental results.

3.3. Word embeddings for word drift

To identify words with distinct meanings between review and description corpora within each domain, we adopted the approach outlined by Angel et al. (2020) [2]. This involved training word embeddings for each corpus and comparing the neighboring words of each target word. We utilized the Gensim[3] implementation of the *word2vec* algorithm [29], employing the CBOW model and a 5-word window, which are the default parameters. Given that the review corpus was larger than the corresponding description corpus for all domains, we selected a vector size of 200 for reviews and 50 for descriptions to reduce the total number of parameters to be learned for description embeddings. Furthermore, we only considered words that appeared at least 50 times in the review corpus and at least ten times in the description corpus. Ultimately, two sets of word embeddings were obtained for each domain by training for five epochs (i.e., passes of the corpus) on the review corpus and ten epochs on the description corpus, as larger vector sizes typically necessitate more epochs.

Subsequently, for each word shared between the two sets of word embeddings within a domain, we identified the 30 closest neighbors in both the review and description corpora. Each shared word was then assigned a score using the following formula:

$$S(w) = \log(\min(f_{w_r}, f_{w_d})) \times (1 - Jaccard(w_r, w_d))^p \quad (1)$$

In this formula, f_{w_r} represents the frequency of word w in the review corpus, while f_{w_d} represents its frequency in the description corpus. The sets w_r and w_d denote the closest neighboring words of w in the review and description word embeddings, respectively. The Jaccard coefficient is represented by $J()$, and p is a parameter for the multiplicative combination set to 5. It is worth noting that in computational linguistics, logarithmic scaling is often applied to word frequencies within a corpus due to Zipf's Law. The objective of this scoring function is to identify the most frequent words in both corpora that share few neighboring words. It should be emphasized that the frequencies f_{w_r} and f_{w_d} are derived from the artificially balanced *True* corpus rather than the original corpus, rendering their values comparable.

[3] https://radimrehurek.com/gensim/models/word2vec.html (retrieved April 2023).

Table 3
Examples of words exhibiting semantic drifts obtained from word embeddings trained separately on reviews and descriptions. See the entire list with 250 entries at https://tinyurl.com/mhk5a4d4. Neighbor words were obtained using cosine distance.

Domain	Word	Neighbors in reviews	Neighbors in descriptions
All Beauty	**sexy**	feminine classy elegant girly manly sparkly	lingerie tights underwear sleepwear lolita
Arts Crafts	**friendly**	efficient intuitive satisfying responsive	safe conscious biodegradable recyclable
Automotive	**cool**	badass sexy sweet cute nice snazzy neat classy	warm toasty refreshing cockpit blazing cozy
Arts Crafts & Sewing	**hands**	fingers finger wrist thumb fingertip	professionalism expertly handwash machine tlc
CDs & Vinyl	**lovers**	enthusiast connoisseur fanatic fan collector	woman man lady fool baby letter girl kisses
Cell Phones & Accessories	**hard**	difficult harder tough easy tricky	tpu silicone silicon hardshell tortoise
Clothing Shoes & Jewelry	**left**	leaving cracked rubbed leaves corner	click sizing exploded front length forearm
Digital Music	**number**	handful slew couple several consecutive	top labels period shelf seam
Electronics	**bus**	buses train plane subway trains	nics uart quatech epp fsb
Kindle Store	**love**	adore loved enjoy loving loves	passion hope true trusts

3.4. Building a ground truth for the word drift task

To establish an initial ground truth for identifying words with the greatest semantic drift between reviews and descriptions, we ranked the words within each domain in descending order according to their $S()$ score (Eq. (1)). We then manually analyzed the two lists of neighboring words to determine the first ten words exhibiting semantic drift. The domains of Gift Cards and Magazine Subscriptions were excluded due to the inferior quality of their descriptions word embeddings, which resulted from their small corpus size. However, we observed that the semantic relationships among the nearest neighboring words were better represented in the word embeddings derived from product reviews than those from descriptions. Numerous unrelated words had to be removed from the set of neighboring words in descriptions in order to discern their meaning within that corpus. This task was performed by a professional linguist. For most domains, the first ten words with clear semantic drift were found among the top 200 words ranked by $S()$, yielding a total of 280 instances in the ground truth. Table 3 presents examples of these words for several domains. The complete manually curated ground truth is available at https://tinyurl.com/mhk5a4d4.

4. Results

In this section, we present the outcomes of our method for measuring the semantic gap (subsection 3.2) in each of the 28 domains of the Amazon dataset. Table 4 displays the file sizes of the uncompressed corpus $|True|$ (equals to $|Rand|$), compressed $C(True)$, and compressed $C(Rand)$ in the first three columns, while the last two columns show the compression ratio $cr(True)$ and the value of δ. As expected, $C(Rand) > C(True)$ for all domains, indicating that randomizing the target word labels systematically increases the information of each domain corpus. Moreover, the table reveals that large variations in corpus sizes between domains do not impact the compression ratio. For instance, the largest corpus (Electronics) and the smallest (Gift Cards) have nearly identical compression ratios. Similarly, the variation of δ is roughly uniform. This finding demonstrates that the compressor and δ effectively control the impact of corpus size.

Fig. 2 displays a scatter plot of $cr(True)$ against δ. The figure illustrates that the δ scores depend on the compression ratio of the *True* corpus. As $cr(True)$ and $cr(Rand)$ are highly collinear ($r = 0.99$), we only consider the relationship between δ and $cr(True)$. The trendline in the plot indicates that δ scores systematically increase as $cr(True)$ does. Therefore, controlling this effect is necessary to obtain a reliable score for the lexical-semantic gap. One possible way to remove the effect of $cr(True)$ from δ is to assume linearity in their relationship, which can be expressed as follows:

$$Gap = \delta - (\alpha \times cr(True) + \beta)$$

The slope of the trendline in Fig. 2 is $\alpha = 0.0127$. To prevent negative gap scores, we adjusted the value of the intercept β to -0.1. As these scores will be correlated with other variables and the used correlation coefficients are independent of translation or displacement of the data, the values of β have no impact on the results. To display the Gap scores for each domain in descending order, we have listed them in Table 5.

To evaluate the relationship between Gap and customer satisfaction, we computed the average product rating on a 5-star scale associated with the reviews having the attribute "verified" in the dataset for each domain. We present the number of verified reviews along with the mean (μ) and standard deviation (σ) of the average rating in Table 5. Additionally, we report the signal-to-noise ratio of user ratings.

The variable Gap is continuous and normally distributed, as verified by passing the D'Agostino and Pearson test[4] with $p = 0.904 > 0.05$. Similarly, the variable "Avg. Rating" is continuous since it represents the average of numerous discrete ratings in each domain. It is also normally distributed, as demonstrated by the same test with $p = 0.257 > 0.05$. Therefore, we utilized the Pearson correlation to measure the relationship between these variables. For $n = 28$ (domains), we obtained a correlation coefficient of $r = -0.39$ with $p = 0.041 < 0.05$, which leads us to accept the hypothesis H. According to Cohen's guideline for behavioral sciences, a medium effect

[4] We utilized the following implementation of the test: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html.

Table 4
Sizes in bytes of the uncompressed combined corpora ($len(True)$), the compressed versions, along with the compression ratio of *True* corpus and the δ measure.

Category domain	$\|True\|$	$C(True)$	$C(Rand)$	$cr(True)$	δ
Clothing Shoes & Jewelry	1,217,914,495	186,406,038	204,659,594	6.53:1	8.92%
Automotive	2,832,375,046	467,371,563	500,261,257	6.06:1	6.57%
Industrial & Scientific	705,511,438	137,303,333	146,208,774	5.14:1	6.09%
Arts Crafts & Sewing	934,203,313	184,218,214	195,594,044	5.07:1	5.82%
Amazon Fashion	266,728,085	47,729,007	50,626,066	5.59:1	5.72%
Cell Phones & Accessories	4,311,334,335	794,017,456	840,087,426	5.43:1	5.48%
Sports & Outdoors	5,962,494,741	1,234,920,172	1,305,708,659	4.83:1	5.42%
Tools & Home Improv.	4,117,764,937	782,820,428	827,427,177	5.26:1	5.39%
Pet Supplies	3,268,259,330	659,406,610	696,922,671	4.96:1	5.38%
Home & Kitchen	9,456,062,039	1,881,635,230	1,987,848,441	5.03:1	5.34%
Electronics	12,883,129,803	2,576,736,801	2,719,114,064	5.00:1	5.24%
Prime Pantry	129,819,718	27,345,889	28,820,273	4.75:1	5.12%
Office Products	2,377,268,914	450,370,239	474,488,208	5.28:1	5.08%
Luxury Beauty	274,992,763	57,509,347	60,553,319	4.78:1	5.03%
Patio Lawn & Garden	2,314,589,274	464,598,990	488,915,211	4.98:1	4.97%
Appliances	238,906,119	38,473,936	40,380,485	6.21:1	4.72%
Toys & Games	3,351,805,288	692,410,324	726,553,150	4.84:1	4.70%
Musical Instruments	858,896,733	175,982,737	184,153,567	4.88:1	4.44%
Software	432,607,252	91,797,240	95,769,822	4.71:1	4.15%
Video Games	2,252,218,353	502,378,978	522,860,137	4.48:1	3.92%
Grocery & Gourmet Food	1,853,386,781	364,717,134	379,243,446	5.08:1	3.83%
Gift Cards	27,601,419	4,970,840	5,160,772	5.55:1	3.68%
All Beauty	168,319,060	39,190,328	40,575,551	4.29:1	3.41%
Kindle Store	4,962,574,129	1,156,436,586	1,188,970,025	4.29:1	2.74%
Movies & TV	5,995,578,632	1,535,411,622	1,567,645,327	3.90:1	2.06%
Magazine Subscriptions	45,939,401	10,387,532	10,601,108	4.42:1	2.01%
Digital Music	694,748,383	184,387,269	187,983,057	3.77:1	1.91%
CDs & Vinyl	4,877,298,653	1,248,562,068	1,270,177,461	3.91:1	1.70%

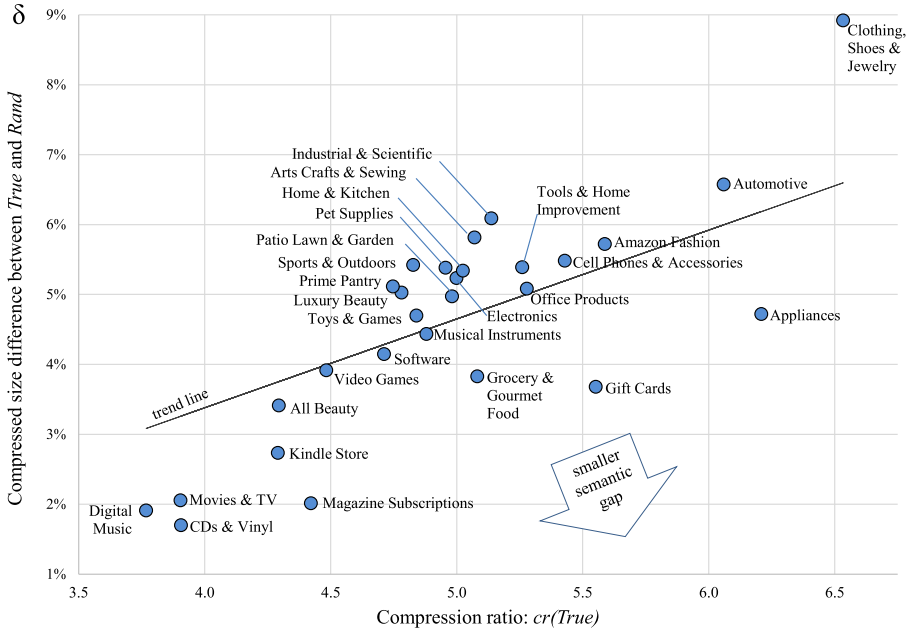

Fig. 2. Relative differences between the compressed sizes of *Rand* and *True* corpora (δ) versus the compression ratio of *True* corpus ($cr(True)$) for 28 e-commerce domains.

size is available with $0.30 < r \leq 0.50$. Furthermore, R^2 indicates that Gap and customer satisfaction share 15.1% of the variance. We also examined the correlation between Gap and SNR and obtained a correlation coefficient of $r = -0.47$ with $p = 0.013 < 0.05$. This outcome provides further evidence for accepting H.

Although the validity of the presented method is theoretically based on the distributional hypothesis for word semantics, information theory, and the soundness of the compression method, additional experimental validation is necessary. We utilized the word

Table 5
Results of our semantic gap function along with the number of ratings, average rating (standard deviation in parentheses), rating signal/noise ratio (SNR) and average Jaccard index for the words with the largest drift (until the 10th ground truth word was retrieved).

Category domain	*Gap*	Ratings	Avg. rating	SNR	J@10
Magazine Subscriptions	0.0640	58,654	4.21(1.29)	3.26	n.a.
Gift Cards	0.0663	138,237	4.71(0.88)	5.33	n.a.
CDs & Vinyl	0.0674	2,578,257	4.60(0.88)	5.22	2.68
Appliances	0.0684	563,870	4.35(1.23)	3.53	6.88
Movies & TV	0.0710	6,731,296	4.31(1.17)	3.70	1.57
Digital Music	0.0713	1,217,667	4.69(0.79)	5.97	1.93
Kindle Store	0.0729	4,036,164	4.28(1.11)	3.86	1.94
Grocery & Gourmet Food	0.0738	4,437,360	4.35(1.21)	3.59	1.87
All Beauty	0.0796	322,473	4.11(1.36)	3.02	1.53
Software	0.0816	309,345	3.75(1.56)	2.41	2.09
Video Games	0.0822	1,948,309	4.10(1.37)	3.00	0.26
Musical Instruments	0.0824	1,346,124	4.28(1.22)	3.51	0.63
Office Products	0.0838	5,055,152	4.22(1.30)	3.24	1.09
Toys & Games	0.0855	7,231,522	4.26(1.25)	3.40	0.67
Cell Phones & Accessories	0.0859	9,209,864	3.94(1.46)	2.69	2.21
Amazon Fashion	0.0863	828,699	3.90(1.42)	2.75	2.60
Patio Lawn & Garden	0.0865	4,799,516	4.15(1.35)	3.07	0.99
Tools & Home Improvement	0.0871	8,299,068	4.25(1.27)	3.35	0.56
Automotive	0.0888	7,561,414	4.27(1.27)	3.35	1.08
Electronics	0.0889	18,597,092	4.11(1.36)	3.02	1.32
Luxury Beauty	0.0895	504,542	4.25(1.29)	3.31	2.15
Home & Kitchen	0.0896	20,015,942	4.22(1.29)	3.28	0.74
Prime Pantry	0.0909	382,686	4.38(1.15)	3.80	1.10
Pet Supplies	0.0909	5,979,680	4.17(1.32)	3.15	1.01
Sports & Outdoors	0.0929	11,913,059	4.26(1.23)	3.47	0.48
Arts Crafts & Sewing	0.0938	2,682,955	4.33(1.20)	3.60	3.29
Industrial & Scientific	0.0957	1,636,564	4.32(1.22)	3.54	0.65
Clothing Shoes & Jewelry	0.1062	30,286,706	4.19(1.23)	3.42	0.73

embedding method, as described in subsection 2.3, which is already known to be suitable for measuring semantic distances in word drifts, provided that two large and balanced corpora are available. To validate the variable *Gap*, we used the ground truth described in subsection 3.4. For each domain, we calculated the average Jaccard scores (from equation (1)) for the leading words according to $S()$ until all ten words in the ground truth were retrieved. These values are reported in Table 5 in the column labeled "J@10". We compared this with *Gap* and checked for other potential confounders, such as the size of the corpus of reviews or descriptions and their differences, to avoid spurious correlation. Since "J@10" failed the normality test ($p = 4.5 \times 10^{-8}$), we opted for the Spearman correlation as a non-parametric alternative and obtained $\rho = -0.40$, $n = 26$ ($p = 0.041$). The correlation is negative because "J@10" is a similarity measure (combination of cosine and Jaccard's measures), while *Gap* is a "distance" measure. This result provides evidence for the validity of the *Gap* measure, indicating that it correlates with the aggregation of word drifts identified semi-automatically using a method commonly used for word-drift identification combined with manual curation.

5. Discussion

The results show that domains related to language, such as Magazine Subscriptions (articles), CDs & Vinyl (lyrics), Kindle Store (books), and Movies & TV (scripts and transcripts), have a small gap. On the other hand, more technical domains like Industrial & Scientific, Electronics, and Sports & Outdoors have a large semantic gap, probably due to the extensive use of specialized jargon. However, other technical domains like Software and Video Games have an intermediate gap, indicating a better alignment between the language used by customers and that used by the description writers. This finding highlights that the quality of information alone does not determine customer satisfaction, but rather, the use of semantically close language is also associated with customer satisfaction. It is worth noting that the Appliances domain, which has a long tradition of providing quality and user-friendly product descriptions [48], has a small semantic gap (4th rank) and a large customer satisfaction (4th rank). This example demonstrates the relationship between a well-semantically aligned language and customer satisfaction. The relationship between customer satisfaction and the quality of descriptions has been extensively studied in the literature [26,22,25]. However, our result suggests that the use of language that is semantically close to that of customers is also associated with customer satisfaction.

Overall, we consider that the semantic gaps obtained for each domain are accurate, indicating the effectiveness of the method proposed in subsection 3.2 as a viable solution to the problem. Furthermore, our findings suggest that the information-theoretical approach is robust, and the use of a commonly used file compressor like *bzip2* is appropriate for this task.

The observed correlation between the semantic gap and the average review ratings is valid as user ratings can be associated with overall customer satisfaction [14]. Hence, the average customer ratings reflect their level of satisfaction, whether it is for receiving precise recommendations, fulfilling expectations, convenient user interface, product quality, description quality, or other factors. Additionally, the stronger correlation between the semantic gap and SNR of user ratings shows that our measure not only correlates with domain ratings but with a "noise-less" signal of user ratings. In other words, it correlates with domains having low variance

(noise or error) in ratings, indicating a higher degree of agreement between user opinions. Consequently, our *Gap* measure correlates with both customer satisfaction and agreement.

Our findings suggest that the lexical-semantic gap could be one of the factors that explains or predicts customer satisfaction, which in turn can positively affect e-commerce revenues [8]. As the semantic gap is potentially controllable from the provider's perspective, this opens up opportunities to investigate its possible causal effects on customer satisfaction and business profits. The relationship between linguistic-semantic alignment and customer satisfaction, along with the significant variation observed across different domains, provides tools for managers to make revenue-weighted intervention decisions. Managerial decisions aimed at improving customer satisfaction are often differentiated by customer type to optimize resources [23]. The semantic gap measure can provide information to support such decisions, either for different domains or other categorizations of the input corpus. Possible measures to reduce the semantic gap include following the recommendations of the "plain language" movement [37], which involves tailoring documents to the intended audience of each customer-facing document. From a theoretical standpoint, the direct relationship between the absence of a semantic gap and customer satisfaction provides a valuable resource for developing behavior models, where the use of language by customers is an independent and observable variable, while on the supplier side, it is controllable.

Regarding the relationship between this semantic gap and the quality of customer-to-customer communication in social e-commerce, our results do not provide direct evidence of its existence. However, it is plausible to infer that customer satisfaction is also related to better communication due to less ambiguous language. An indirect evidence of that is the stronger relationship found between semantic gap and ratings signal-noise ratio [44], meaning that such gap is related to lower variance in ratings. This variance in ratings has been associated with niche products, i.e. products that are either loved or hated by a small segment of customers [44], which could be a confounding factor.

With respect to the observed relationship between our semantic gap function and the ground truth ("J@10" in Table 5), this provides a quantitative indication that the information-theoretical method presented measures the semantic gap due to word drift and ambiguity. Thus, it can be concluded that the method is effective in measuring the semantic gap between customers and suppliers in e-commerce, although it does have the limitation of not identifying particular words. Nonetheless, the manually extracted ground truth is a necessary resource for developing and evaluating methods to identify these words.

The proposed method for measuring the semantic gap between customers and suppliers has an advantage in that it is not based on specific features of the English language. Therefore, it can be applied to any language in which speech can be segmented into words. The Amazon data used in our study provided a wide range of corpus sizes, but even the smallest corpus (as shown in Table 1) can be considered large in comparison to those that can be obtained from smaller e-commerce sites or domains. While we did not investigate the reliability of the semantic gap measurements for small corpora, we believe that a corpus of reviews with more than 1 million words and a corpus of product descriptions with at least 100,000 words for each domain should be sufficient for a reliable measurement.

6. Conclusions, limitations, and implications

We introduced a new method for measuring the lexical-semantic gap between product reviews and descriptions in various e-commerce domains. Our approach is based on distributional semantics and information theory, and is inspired by the effective application of file compression in other domains. Our method requires few assumptions, few parameters, and is language-independent. We have also shown that our approach overcomes the limitations of current methods when comparing corpora of vastly different sizes. Additionally, we were assisted by word embeddings to manually construct a ground truth dataset for our study, which allowed us to identify words with semantic drift between product reviews and descriptions. This new resource provided additional validation for our method, as we found that the semantic gap measures for each domain were significantly correlated with the average semantic drift of the words in the semi-automatically obtained model. In conclusion, our method is an effective and valuable tool for measuring lexical-semantic gaps in e-commerce domains, and our ground truth dataset represents a valuable new resource for future research in this area.

The most significant finding of this study is that there is an inverse relationship between the semantic gap of various e-commerce domains and the quantitative ratings given by customers on a 5-star scale for the products they purchase. This suggests that the semantic differences in the language used by customers and suppliers are related to customer satisfaction. Although it may seem counter-intuitive, this study has shown that the language used in product descriptions and user reviews is connected through the ambiguities in their shared vocabulary. Specifically, we found that a greater semantic gap results in lower customer satisfaction, as represented by their ratings. Our observations were made using the proposed method on a large corpus consisting of 6.9 billion words of user reviews and 1.5 billion words of product descriptions from 28 marketing domains on Amazon. We believe that this corpus is likely the largest and most representative textual e-commerce data publicly available. Our results show a significant correlation between the semantic gap and customer satisfaction, with an observed correlation coefficient of $r = -0.39$ ($p \leq 0.041$).

6.1. Theoretical contribution

Customer satisfaction is a complex and widely studied phenomenon in e-commerce, but there have been relatively few studies focused on the role of language. In social e-commerce environments, where customers engage in written dialogue, language plays an even more prominent role than in traditional e-commerce settings. We have introduced the concept of the *semantic gap*, which refers to the differences in the meaning of words used by customers and suppliers. Despite the limitations of this scenario, we have demonstrated that the semantic gap can be effectively measured and is negatively related to customer satisfaction. This finding is

significant because semantics is likely the main factor associated with the effectiveness of human communication. Our study suggests that language not only serves a functional-utilitarian role in the customer-provider relationship, but also has implications for the quality and effectiveness of the business. Therefore, understanding and managing the semantic gap in e-commerce environments may be an important factor in improving customer satisfaction and overall business success.

The semantic gap is a novel variable that adds to the existing factors known to explain customer satisfaction, such as logistics quality, product quality, customer service, personalization, and others. This new variable has the potential to enhance the predictive power of customer satisfaction models since it represents a domain external to the current management research domain, which is not typically considered in current theory.

Our contribution has practical implications for any social e-commerce environment where there is a significant amount of customer-generated text, such as product reviews and textual descriptions. In such scenarios, the semantic gap, provides a valuable quantitative indicator that may help to identify subsets of customers with low satisfaction.

6.2. Limitations

From a linguistic perspective, the method we have presented for measuring the semantic gap has few limitations that would prevent its application to other types of texts or languages. In practice, the method only requires the ability to segment texts at the word level, perform random manipulations at this level, and then use a standard file compressor. In agglutinative languages, such as Finnish, word segmentation may be necessary to control excessive vocabulary size. Moreover, we anticipate that the method could be applied to other types of documents representing users, such as complaints and requests, and providers, such as manuals and regulations. We believe that our approach based on Shannon's information theory, as represented by our hypothesis H, may be compatible with these types of data. Similarly, the semantic gap method could be used with transcripts of oral dialogues, which opens up new possibilities for research on the semantic differences between spoken and written language.

A limitation of our study is that its applicability is primarily relevant for large retail companies with a considerable number of domains to compare. In e-commerce environments based on a single domain, the usefulness of the semantic gap measure may be limited if it cannot be compared with other domains, as it is relative to factors such as the size of the corpus and the compression ratio. However, relative semantic gap measures could be used in customer groupings other than by product domain, such as demographic or behavioral factors, among others.

While it is plausible to generalize that a small semantic gap, implying good communication, could be positively related to customer satisfaction, our results show that this relationship, while statistically significant, is relatively weak. Therefore, caution should be exercised when applying this relationship to other domains that are dissimilar from those studied, such as finance or real estate. Similarly, the potential causality from the semantic gap to customer satisfaction seems reasonable, given the apparent impossibility of an implication in the opposite direction. However, this consideration should also be approached with caution, as our results suggest possible confounding factors, such as the degree to which language is involved with the product, as in the case of magazines and movies. Therefore, it is important to continue exploring the relationship between the semantic gap and customer satisfaction in other domains and contexts, as well as investigating possible confounding variables that may impact this relationship.

6.3. Implications

The main contributions and implications of this study can be summarized as follows:

1. For managers, it is important to recognize that different marketing domains exhibit varying degrees of differences in meaning between the words used by customers and those used in product descriptions, (i.e., the semantic gap). Our study has shown that this difference is related to customer satisfaction: the smaller the semantic gap, the higher the customer satisfaction. This relationship was obtained using linguistic data associated with each domain, with almost no assumptions or additional information required. Therefore, it is reasonable to expect that these findings can be transferred beyond the 28 Amazon domains studied. To measure the semantic gap, all that is required is a relatively large corpus of texts produced by customers, such as product reviews, and another corpus with product descriptions. These two corpora are compared using a simple pre-processing method (accessible to any professional computer programmer) and a standard file compressor, such as bzip2. Our experiments revealed an inverse relationship between the semantic gap and customer satisfaction. While causality is not proven and there may be confounding factors, using plain language with a vocabulary containing few ambiguities is an intuitively convenient option for better communication. Any action aimed at reducing the semantic gap in the written discourse of product descriptions is likely to have a beneficial effect on customer satisfaction and, ultimately, business profits.
2. For academics, it is widely recognized that the quality of information in product descriptions is a critical factor in determining customer satisfaction. Our study has the concept of "quality of information" by introducing the notion of "quality of communication" between customers and suppliers, specifically when product descriptions are used as a communication channel. In this context, quality of communication refers to the degree of ambiguity in the shared vocabulary used by customers and suppliers.
3. For computational linguists, the method we have presented to measure the semantic gap between two corpora by compression is a novel approach that has proven to be particularly effective in cases where there is a significant imbalance between the size of the corpus being compared. Our method can be easily adapted to measure the semantic gap in subsets of vocabulary words, and even to assess the semantic drift of individual words. Additionally, our experiments showed that our method produced results similar to those obtained when comparing distances using neural word embeddings.

4. For system designers and developers, measuring the δ measure for the semantic gap as a means of monitoring the quality of communication with users in an electronic commerce platform. By continuously measuring the semantic gap, and comparing the texts of recent user reviews against the product descriptions of each domain, it is possible to detect anomalies or sudden changes in the semantic gap. These changes may arise due to shifts in the user population resulting from the incursion into new markets or geographic locations, among other factors. In such cases, creators of product descriptions can take corrective actions to preserve the semantic quality of communication with their users.
   Similarly, keeping updated a word embeddings model trained with a corpus of reviews and descriptions combined with the method presented (see section 3.2) is the input for building writing assistants for user communication and product descriptions. Each word in these texts can be verified for its degree of semantic drift in the embedding model to alert the writer about possible misunderstandings. In this way, writers can take corrective measures such as including clarifications in parentheses, adding notes, or avoiding using these words in their writing. Indeed, one of the limitations of this usage scenario is the imbalance between the size of the user reviews corpus versus the product description corpus, which produces noisy word drifts. One way to mitigate this problem would be to augment product texts with other sources, such as user manuals and product documentation.

6.4. Future work

Several research perspectives are opened from the methods and resources presented in this study:

1. A perspective that opens from this study is to experimentally evaluate the possible causal relationship between the semantic gap in a controlled commercial scenario.
2. It is interesting to investigate how the lexical-semantic gap relates other factors of e-commerce, such as customer service, trust, loyalty, and others.
3. It is also interesting to study whether the semantic gap also relates customer-customer communications in social commerce and its implications.
4. The users' reviews have other features in addition to their sentiment polarity (i.e., rating), such as intensity and usefulness, that can serve to explore the semantic gap and thus expand the understanding of it and its relationships with other factors of marketing.
5. Research on the possible transfer of the methods presented to other analogous domains, such as those of the citizen-government or patient-healthcare system, becomes relevant.

CRediT authorship contribution statement

Carlos A. Rodriguez-Diaz: Conceptualization, Methodology, Software, Formal analysis, Writing, Vizualization. **Sergio Jimenez:** Conceptualization, Methodology, Software, Formal analysis, Writing. **Daniel Bejarano:** Conceptualizacion, Writing. **Julio A. Bernal-Chávez:** Conceptualization, Writing, Funding acquisition. **Alexander Gelbukh:** Conceptualization, Methodology, Formal analysis, Writing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in the article is an already public dataset

References

[1] Emad Abu-Shanab, Yousra Harb, E-government research insights: text mining analysis, Electron. Commer. Res. Appl. 38 (2019) 100892.
[2] Jason Angel, et al., NLP-CIC@ DIACR-Ita: POS and neighbor based distributional models for lexical semantic change in diachronic Italian corpora, arXiv preprint, arXiv:2011.03755, 2020.
[3] Mikel Artetxe, Gorka Labaka, Eneko Agirre, A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings, arXiv preprint, arXiv:1805.06297, 2018.
[4] William J. Ashby, The drift of French syntax, Lingua 57 (1) (1982) 29–46.
[5] Meritxell Fernández Barrera, et al., Enhancing cross-border EU E-commerce through machine translation: needed language resources, challenges and opportunities, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC'16, 2016, pp. 4550–4556.
[6] Lisa Beinborn, Rochelle Choenni, Semantic drift in multilingual representations, Comput. Linguist. 46 (3) (2020) 571–603.
[7] Prasad Bingi, Ali Mir, Joseph Khamalah, The challenges facing global e-commerce, Inf. Syst. Manag. (2006).
[8] Karl Markos Biswas, Mohammed Nusari, Abhijit Ghosh, et al., The influence of website service quality on customer satisfaction towards online shopping: the mediating role of confirmation of expectation, Int. J. Manag. Sci. Bus. Adm. 5 (6) (2019) 7–14.
[9] Deborah Cameron, The Myth of Mars and Venus, Oxford University Press, USA, 2007.
[10] Qibin Chen, et al., Towards knowledge-based personalized product description generation in e-commerce, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 3040–3050.
[11] Rudi Cilibrasi, Paul Vitanyi, Clustering by compression, arXiv preprint, arXiv:cs/0312044, 2003.

[12] Guy Elad, Kira Radinsky, Benny Kimelfeld, Generating personalized product descriptions from user reviews, PhD thesis, Computer Science Department, Technion, 2019.
[13] T. Mark Ellison, Luisa Miceli, Distinguishing contact-induced change from language drift in genetically related languages, in: Proceedings of the Workshop on Computational Models of Language Acquisition and Loss, 2012, pp. 1–9.
[14] Tobias H. Engler, Patrick Winter, Michael Schulz, Understanding online product ratings: a customer satisfaction model, J. Retail. Consum. Serv. 27 (2015) 113–120.
[15] John R. Firth, A synopsis of linguistic theory, in: Studies in Linguistic Analysis, 1957, pp. 1930–1955.
[16] Julie E. Francis, Lesley White, et al., PIRQUAL: a scale for measuring customer expectations and perceptions of quality in internet retailing, in: K. Evans, L. Scheer (Eds.), AMA Winter Educators' Conference, 2002, pp. 263–270.
[17] Lea Frermann, Mirella Lapata, A bayesian model of diachronic meaning change, Trans. Assoc. Comput. Linguist. 4 (2016) 31–45.
[18] Negin Ghasemi, Saeedeh Momtazi, Neural text similarity of user reviews for improving collaborative filtering recommender systems, Electron. Commer. Res. Appl. 45 (2021) 101019.
[19] Rujun Han, et al., Conditional word embedding and hypothesis testing via bayes-by-backprop, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4890–4895.
[20] Wei Hua, Zhou Jing, An empirical study on e-commerce logistics service quality and customer satisfaction, in: WHICEB Proceeding, 2015, pp. 269–275.
[21] Cuixia Jiang, et al., The impact of soft information extracted from descriptive text on crowdfunding performance, Electron. Commer. Res. Appl. 43 (2020) 101002.
[22] Ling Alice Jiang, Zhilin Yang, Minjoon Jun, Measuring consumer perceptions of online shopping convenience, J. Serv. Manag. (2013).
[23] Timothy L. Keiningham, et al., Does customer satisfaction lead to profitability? The mediating role of share-of-wallet, Manag. Serv. Qual. (2005).
[24] Hsiangchu Lai, Wan-Jung Lin, Gregory E. Kersten, The importance of language familiarity in global business e-negotiation, Electron. Commer. Res. Appl. 9 (6) (2010) 537–548.
[25] Eun-Ju Lee, Soo Yun Shin, When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo, Comput. Hum. Behav. 31 (2014) 356–366.
[26] Moez Limayem, Mohamed Khalifa, Anissa Frini, What makes consumers buy from internet? A longitudinal study of online shopping, IEEE Trans. Syst. Man Cybern., Part A, Syst. Hum. 30 (4) (2000) 421–432.
[27] Pu Liu, et al., The effects of social commerce environmental characteristics on customers' purchase intentions: the chain mediating effect of customer-to-customer interaction and customer-perceived value, Electron. Commer. Res. Appl. 48 (2021) 101073.
[28] Izyan Hizza Bt Hila Ludin, Boon Liat Cheng, Factors influencing customer satisfaction and e-loyalty: online shopping environment among the young adults, Manag. Dyn. Knowl. Econ. 2 (3) (2014) 462.
[29] Tomas Mikolov, et al., Distributed representations of words and phrases and their compositionality, in: Proceedings NIPS, 2013, pp. 3111–3119.
[30] George A. Miller, et al., Introduction to WordNet: an on-line lexical database, Int. J. Lexicogr. 3 (4) (1990) 235–244.
[31] Burt L. Monroe, Michael P. Colaresi, Kevin M. Quinn, Fightin'words: lexical feature selection and evaluation for identifying the content of political conflict, Polit. Anal. 16 (4) (2008) 372–403.
[32] Marcelo A. Montemurro, Damián H. Zanette, Universal entropy of word ordering across linguistic families, PLoS ONE 6 (5) (2011) e19875.
[33] Jian Mou, Wenlong Zhu, Morad Benyoucef, Impact of product description and involvement on purchase intention in cross-border e-commerce, Ind. Manag. Data Syst. (2019).
[34] Jianmo Ni, Jiacheng Li, Julian McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019, pp. 188–197.
[35] Slava Novgorodov, et al., Generating product descriptions from user reviews, in: The World Wide Web Conference, 2019, pp. 1354–1364.
[36] Rolan Patrada, Erna Andajani, Effect and consequence e-customer satisfaction for e-commerce users, IPTEK J. Proc. Ser. 1 (2021) 219–227.
[37] Roslyn Petelin, Considering plain language: issues and initiatives, Corp. Commun. 15 (2) (2010) 205–216.
[38] Reid Pryzant, Youngjoo Chung, Dan Jurafsky, Predicting sales from the language of product descriptions, in: eCOM@ SIGIR, 2017.
[39] Mario Rese, Relationship marketing and customer satisfaction: an information economics perspective, Mark. Theory 3 (1) (2003) 97–117.
[40] Sebastian Ruder, Ivan Vulić, Anders Søgaard, A survey of cross-lingual word embedding models, J. Artif. Intell. Res. 65 (2019) 569–631.
[41] Roland T. Rust, Anthony J. Zahorik, Customer satisfaction, customer retention, and market share, J. Retail. 69 (2) (1993) 193–215.
[42] Sahlgren Magnus, The word-space model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, PhD thesis, Stockholm University, 2006.
[43] H. Andrew Schwartz, et al., Personality, gender, and age in the language of social media: the open-vocabulary approach, PLoS ONE 8 (9) (2013) e73791.
[44] Monic Sun, How does the variance of product ratings matter?, Manag. Sci. 58 (4) (2012) 696–707.
[45] Joseph Turian, Lev Ratinov, Yoshua Bengio, Word representations: a simple and general method for semi-supervised learning, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 384–394.
[46] Ravichandran Vengadasamy, Azhar Jaludin, Afendi Hamat, Characteristics of written text in e-commerce websites, Internet J. e-Lang. Learn. Teach. 1 (2) (2004) 15–32.
[47] Wai Ming Wang, et al., Extracting and summarizing affective features and responses from online product descriptions and reviews: a Kansei text mining approach, Eng. Appl. Artif. Intell. 73 (2018) 149–162.
[48] Yuren Wang, Xin Lu, Yuejin Tan, Impact of product attributes on customer satisfaction: an analysis of online reviews for washing machines, Electron. Commer. Res. Appl. 29 (2018) 1–11.
[49] Jing Zhang, Xingchen Lu, Dian Liu, Deriving customer preferences for hotels based on aspect-level sentiment analysis of online reviews, Electron. Commer. Res. Appl. 49 (2021) 101094.
[50] Xiaolin Zheng, Shuai Zhu, Zhangxi Lin, Capturing the essence of word-of-mouth for social commerce: assessing the quality of online e-commerce reviews by a semi-supervised approach, Decis. Support Syst. 56 (2013) 211–222.