Expert Systems With Applications 209 (2022) 118309

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Understanding customer satisfaction via deep learning and natural language processing

Ángeles Aldunate [a], Sebastián Maldonado [b,d], Carla Vairetti [a,d,*], Guillermo Armelini [c]

[a] Universidad de los Andes, Chile, Facultad de Ingeniería y Ciencias Aplicadas, Santiago, Chile
[b] Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Santiago, Chile
[c] Universidad de los Andes, Chile, ESE Business School, Santiago, Chile
[d] Instituto Sistemas Complejos de Ingeniería (ISCI), Chile

ARTICLE INFO

Keywords:
Analytics
Customer satisfaction
Customer feedback
Natural language processing
Deep learning
BERT

ABSTRACT

It is of utmost importance for marketing academics and service industry practitioners to understand the factors that influence customer satisfaction. This study proposes a novel framework to analyze open-ended survey data and extract drivers of customer satisfaction. This is done automatically via deep learning models for natural language processing. According to 11 drivers acknowledged by the marketing literature to determine customer experience, the data is cast into a multi-label classification problem. This expert system not only supports the automatic analysis of new data but also ranks the drivers according to their importance to various service industries and provides important insights into their applications. Experiments carried out using 25,943 customer survey responses related to 39 service companies in 13 different economic sectors show that the drivers can be identified accurately.

1. Introduction

The growth of artificial intelligence (AI) is rapidly changing how marketing decisions are taken in the service industry (Huang & Rust, 2018, 2021). Several AI-based expert systems have been developed in recent years, addressing the latest challenges such as digital content marketing (DCM) and messaging (Gregoriades et al., 2021), prediction of online shopping behavior from clickstream data (Koehn et al., 2020), or social network targeting (Robles et al., 2020).

AI-based expert systems have enormous potential in customer experience (CX) (Huang & Rust, 2018, 2021). In a March 2020 MIT Technology Review Insights survey of more than 1000 business leaders, the results indicate that customer service departments are most likely to use AI (73%), followed by sales and marketing departments (59%). The survey also predicts customer service will be the leading department by 2022 (McCauley, 2020). According to McAfee et al. (2012), organizations that successfully manage customer feedback data are, on average, 6% more profitable and 5% more productive than other organizations; AI is achieving excellent results in this area.

One of the main challenges in the service industry is to manage customer experience, i.e., all the interactions that a customer might have with the firm during the purchasing process. Given that customer perception is subjective, it is important to gather his/her opinion on the last experience to improve the service offering. In doing so, firms usually survey random customers asking the level of satisfaction and their feedback (open-ended questions), and then they must process all these insights in order to improve. Therefore, automatically extracting useful insights from verbatim CSAT responses is extremely relevant because customers constantly change their behavior, making it important to weigh each driver of CSAT accordingly. Furthermore, companies need to collapse insights about customers into key performance indicators that reveal critical customer opinions (McColl-Kennedy et al., 2019).

To address this challenge, we propose a novel framework for extracting knowledge from customer feedback data using deep learning (DL). This popular AI approach has become the de facto strategy for extracting knowledge from text (Manning, 2015). DL has shown excellent results in marketing, improving classification results when dealing with textual resources (Ma & Sun, 2020). Our main contributions are:

- In line with other AI-based expert systems (Gregoriades et al., 2021; Koehn et al., 2020), this study makes a relevant contribution to the empirical literature on marketing analytics. We present a novel AI task, which is the automatic identification of drivers of CSAT according to open-ended responses. We conduct a thorough literature review to find 11 factors/drivers that determine CSAT:

* Corresponding author at: Universidad de los Andes, Chile, Facultad de Ingeniería y Ciencias Aplicadas, Santiago, Chile.
E-mail addresses: adaldunate@miuandes.cl (Á. Aldunate), sebastianm@fen.uchile.cl (S. Maldonado), cvairetti@uandes.cl (C. Vairetti), garmelini.ese@uandes.cl (G. Armelini).

https://doi.org/10.1016/j.eswa.2022.118309
Received 22 October 2021; Received in revised form 7 June 2022; Accepted 26 July 2022
Available online 30 July 2022

price, reliability, ease of use, responsiveness, empathy, assurance, convenience, brand, assortment, atmosphere, and past CX (Grewal et al., 2009; Parasuraman et al., 1988; Verhoef et al., 2009). The success of our novel predictive approach is an important contribution for marketing researchers and practitioners. We gain important insights into the application.

- We propose a novel four-step methodology for understanding customer feedback data via descriptive analytics and AI. Our goal is to gain managerial insights from customer ratings and drivers of CSAT, hidden in open-ended responses, that DL identifies. We also benchmark a state of the art DL method such as BERT (Devlin et al., 2018) against established classification models. We collect a data set of 25,943 customer responses related to 39 Chilean companies and use it to construct DL models that accurately identify whether responses include specific drivers.
- We consider state-of-the-art DL technologies for natural language processing (NLP). Several text mining and NLP studies have applied DL models – including studies related to customer satisfaction and polarity detection (Kumar & Zymbler, 2019) – but we do not know of any research that applies these techniques to survey data for customer feedback. Surveys, unlike online customer reviews, allow firms to assess the perception of specific customer segments about certain topics. While reviews reflect what the customer wants to say in a specific context about certain topics, sometimes companies need answers to specific questions that consumers might not be addressed in their reviews. For example, the NPS methodology considered in this study seeks to understand how likely a consumer would recommend a firm's product/service/brand to her/his colleagues. In this case, the instrument (survey) wants to assess the level of recommendation a specific company has according to a random sample.
- Although our contribution is mainly applied, we propose a novel classification strategy for NLP in which we combine the TF-IDF and BERT embeddings as an alternative for the learning process.

This paper is organized as follows: Section 2 discusses prior studies relevant for this paper. The proposed framework for analyzing CSAT data using deep learning is presented in Section 3. Section 4 provides the results obtained by using real-world datasets from Chilean companies of several sectors of the service industry, also discussing managerial insights for decision-making. The main conclusions are summarized in Section 5, addressing possible directors for future developments.

2. Prior work

This section first provides an overview of prior studies on CSAT, describing the NPS measure (Section 2.1). Next, Section 2.2 introduces deep learning and its application in CSAT.

2.1. Customer satisfaction and the Net Promoter Score (NPS)

The service and marketing literature devotes much attention to the antecedents and consequences of customer satisfaction, defined as "the consumer's judgment that a product or service meets or falls short of expectations" (Gupta & Zeithaml, 2006). CSAT is a holistic and subjective metric - i.e., it reflects the individual's perception of the whole experience - (Guenther & Guenther, 2020; Verhoef et al., 2009). Modeling customer satisfaction has also become an important research topic for machine learning researchers, which aim to extract relevant information automatically from customer reviews and other data sources to improve decision-making (Kumar & Zymbler, 2019).

Scholars developed tools, such as the American Customer Satisfaction Index (Anderson & Fornell, 2000), to assess the construct. However, firms seek simpler methods because long surveys tend to overwhelm customers and thus evoke low response rates. For this reason, practitioners have proposed new metrics for measuring customer satisfaction, and NPS is among the most popular (De Haan et al., 2015; Reichheld & Markey, 2011). It measures customer experience and predicts business growth, gaining insights on how the company can obtain more promoters and fewer detractors (Reichheld & Markey, 2011).

NPS data consists of a score on a 0 to 10 scale for the following question: How likely is it that you would recommend our company/product/service to a friend or colleague? This question is followed by an open-ended request for elaboration, soliciting the reasons for the rating. The promoters (9s and 10s), passives (7s and 8s), and detractors (0s through 6s) are obtained based on the answer to the first question via a simple binning process. Then, the answer to the open-ended "why?" question provides the verbatim considered for the learning process in the framework proposed in this study.

The service quality literature proposes several drivers that ultimately explain satisfaction ratings. The SERVQUAL model (Parasuraman et al., 1988) was the first model to measure service quality; it includes five dimensions of service quality: reliability, responsiveness, assurance, empathy, and tangibility. Several empirical studies have tested its scales (Zeithaml & Parasuraman, 2004). In service marketing, scholars also have proposed additional drivers, such as employees' empathy (Zeithaml et al., 1996) and identified factors that affect CSAT in retail environments (Table 1), such as the atmosphere (Grewal et al., 2009), product or service availability (i.e., assortment), price (e.g., policy, promotions, loyalty programs), and brands.

2.2. Text mining applied to CSAT

Text mining, also known as text analytics, is a discipline that uses various techniques to extract valuable information from text automatically (Manning et al., 1999). The automatic analysis of the content of various text sources provides several advantages to companies in terms of scalability and reliability, thus discouraging preconception and biases (Yu et al., 2011). Text mining offers companies high potential for insight because approximately 80% of corporate information is available in textual data formats (Ur-Rahman & Harding, 2012). This approach is strongly related to NLP, which concerns the interactions between human language and machines, aiming at processing and analyzing natural language data (Liu et al., 2019).

Typical text analytics tasks include text categorization, text clustering, concept/entity extraction, document summarization, and sentiment analysis. Customer feedback research mainly has focused on sentiment analysis, that is, the extraction of information and patterns related to opinions or sentiments (Goldberg & Zhu, 2006; Lin & He, 2009; Ye et al., 2009). Our study is one of the first to seek to understand customer feedback via text categorization (text classification), which, by assigning predefined categories to free-text documents (Manning et al., 1999) provides a powerful tool for text pattern analysis. A limitation of text classification, however, is that it needs a large amount of manually labeled data to learn from past examples, especially when using DL approaches (Gargiulo et al., 2019; Manning, 2015). In these supervised learning approaches, the machine learning process is guided by the outcome of past examples, unlike descriptive approaches such as data clustering, topic models, and association rules (Medhat et al., 2014).

There are several text-based resources, such as reviews, complaints, websites, books, emails, and articles, that are useful to companies (Kraus et al., 2020; Ma & Sun, 2020). Arguably, reviews are the most valuable resource for customer feedback; several studies have been devoted to analyzing this data source, and many companies use customer reviews to improve their business strategies (Decker & Trusov, 2010). Recently, researchers have applied text mining to various contexts, such as online hotel reviews (Gregoriades et al., 2021; Miguéis & Nóvoa, 2017), and restaurant reviews (Vairetti et al., 2020). A study of automobile reviews, by Ramaswamy and DeClerck (2018), is especially relevant because it uses semantic tagging for DL. Semantic tagging

Table 1
Drivers of customer experience indicated in the literature.

Variable	*Description*	*Reference*
ASSORTMENT	Variety, uniqueness and quality of the offer.	Grewal et al. (2009), Huffman and Kahn (1998) and Janakiraman et al. (2006)
ASSURANCE	Knowledge and courtesy of employees and their ability to convey trust and confidence.	Jerger and Wirtz (2017), Parasuraman et al. (1988) and Zeithaml et al. (1996)
ATMOSPHERE	Consumers attend to design, social, and ambient environment cues when evaluating stores, because they believe these cues offer reliable information about product-related attributes.	Baker et al. (2002), Grewal et al. (2009), Kaltcheva and Weitz (2006) and McColl-Kennedy et al. (2019)
BRAND	Extent to which brand names affect perception of CX	Keller (2003) and Warren et al. (2019)
CONVENIENCE	Offering choice, consistency, and timeliness at the channel level.	Lemon and Verhoef (2016) and Verhoef et al. (2009)
EASE OF USE	How easy is it to do business with the focal company? How much effort should a customer invest to interact with the company?	Monsuwé et al. (2004), Soudagar et al. (2011) and Zeithaml et al. (1996)
EMPATHY	Caring, individualized attention to its customers; capacity to recognize feelings that are being experienced by others.	Jerger and Wirtz (2017), Parasuraman et al. (1988) and Zeithaml et al. (1996)
PAST CX	All interactions between an organization and a customer during customer's lifetime as a client.	Van Doorn and Verhoef (2008) and Verhoef et al. (2009)
PRICE	Pricing strategies play key role in CX, either as a value proposition or in other tools (i.e., loyalty programs) that also affect pricing.	Gauri et al. (2008), Noble and Phillips (2004) and Yang et al. (2019)
RELIABILITY	Ability to perform promised service dependably and accurately; "Do what you say you're going to do when you said you were going to do it".	McColl-Kennedy et al. (2019), Parasuraman et al. (1988) and Soudagar et al. (2011)
RESPONSIVENESS	Respond quickly, promptly, rapidly, immediately, or instantly.	Parasuraman et al. (1988) and Zeithaml et al. (1996)

describes a document, allowing graphic representations that are useful for summarizing information from customer perceptions. However, it is an unsupervised task that – unlike our study – does not support the categorization of reviews.

We would like to emphasize the differences between previous studies that analyze customer reviews automatically and the proposed framework. On the one hand, supervised learning has been useful for classifying the polarity of a given review based on its rating and content (Gregoriades et al., 2021). On the other hand, unsupervised learning can unveil the topics hidden in survey data, classifying reviews according to these topics (Pietsch & Lessmann, 2019). However, this is the first study that develops a learning machine designed to identify drivers of customer experience in a supervised manner, which has the advantage of guiding the learning process with the exact concepts practitioners would like to seek in open-ended responses. In other words, we are presenting a completely new predictive task in marketing analytics, which we address via DL and multilabel classification.

Customer complaints also are relevant resources. Coussement and Van den Poel (2008) propose an automatic email classification system for distinguishing between complaints and non-complaints. In a supervised manner, the authors use Adaboost as a classifier in combination with latent semantic indexing (LSI) and singular value decomposition (SVD) for dimensionality reduction. Joung et al. (2019) analyze customer complaints using text mining to identify customers' true needs from complaint data and thereby develop market-oriented products. Text mining literature has discussed the task of analyzing customer feedback and the requirements for product design; it has focused on products such as home appliances (Wang et al., 2018), automobiles (Aguwa et al., 2017), mobile phones (Decker & Trusov, 2010), and real estate (Wang & Tseng, 2015). Such text mining analyses typically use data from product reviews or surveys.

It is important to notice that textual data is not the only valuable source for marketing decisions in the digital era. Recent AI-based expert systems go beyond the traditional sociodemographic factors and purchase history, incorporating sources such as graph theory features created from social networks (Robles et al., 2020) or clickstream data (Koehn et al., 2020).

3. Proposed methodology for understanding CSAT

In this section, we propose a novel framework that uses DL to extract knowledge from customer feedback surveys. DL extends traditional artificial neural network (ANN) architectures for dealing with unstructured data, such as images or text (LeCun et al., 2015). Our main goal is to create a machine that interprets verbatim comments, classifying them automatically into 11 factors that the marketing and service literature acknowledges are influencers of CSAT.

We propose a four-step methodology for analyzing customer feedback data as a sequence of steps for value creation. This approach is a customized version of the cross-industry standard process for data mining (CRISP-DM) methodology (Chapman et al., 2000), which has been widely used, including in text mining efforts for customer feedback (see e.g. Villarroel-Ordenes et al., 2014).

CRISP-DM is a general framework for applying learning algorithms in the industry. It consists of six major steps or phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000). The proposed framework is an adaptation of this strategy for analyzing customer satisfaction data via NLP and text mining techniques. In other words, we elaborate on the CRISP-DM steps with the purpose of designing a framework tailored for understanding customer satisfaction.

The goal of the framework is to create a machine that classifies each open-ended customer feedback response according to the various drivers of CSAT. At the same time, we seek to extract as much knowledge as possible from the relationship between open-ended responses and ratings. This second business objective can be achieved via word clouds and other descriptive-analytics tools.

In this framework, we take advantage of both the score that classifies customers as promoters, passives, or detractors and the verbatim comments related to the open-ended question associated with the score, to gain insight into the application. Fig. 1 summarizes the four steps of the proposed framework.

3.1. Step 1: Data gathering, visualization, and data processing

The first step consists of three sub-steps associated with the CRISP-DM methodology. The first sub-step relates to business understanding

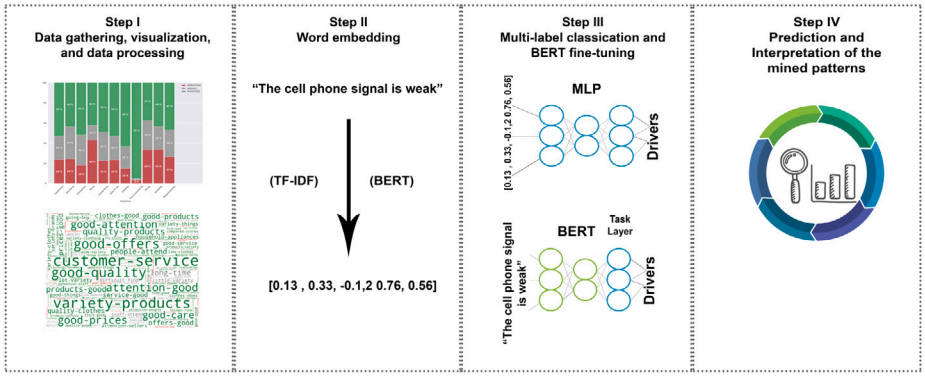

Fig. 1. Summary of the proposed four-step framework for analyzing customer feedback data.

and defining the goals of the text mining process. It identifies the major themes in the data, whereas subsequent steps relate to a deeper analysis of one or more of these themes via machine learning. The output of this step is the set of business objectives and the success criteria (Chapman et al., 2000).

The second sub-step corresponds to the data exploration and visualization step of the CRISP-DM methodology. This sub-step has two goals: (1) to understand and assess the quality of the data that provide input to the modeling process and (2) to identify initial patterns relevant to customer understanding. We consider the creation of word clouds – with words and phrases colored according to their predominant label (promoter, passive, or detractor) – particularly valuable. The output of this step is a report on the data quality and the data exploration process, which provides the necessary insights for adequate data pre-processing (Chapman et al., 2000).

It is important to notice that Step 1 identifies the drivers that have a large impact on the NPS score. This information is key for gaining managerial insights from the automatic driver identification process proposed in Step 3 since it allows the prioritization of comments according to the drivers that are included in them, leading to personalized actions in order to improve the customer experience of key customers.

The third sub-step is data preparation, including data selection, annotation, and cleansing. The output of this step is a pre-processed data set that is ready for model training.

We collected the data via a telephone survey and manually transcribed the responses. In the same process, we manually annotate the open-ended responses, thereby identifying labels for the machine learning model. To illustrate the tagging process, we present two examples of open-ended answers from the telecommunication sector (mobile phones). The statement "It has low fares and you can talk more for less money" relates to the driver "price", and the sentence "The call quality is terrible, and the cell phone's functionalities are poor" relates to the driver "reliability". With regard to pre-processing of data for text mining, we followed a standard strategy that consists of removing stopwords (frequent words such as articles and nouns) and stemming (a method that reduces derived words to their root forms) (Martínez Cámara et al., 2011).

3.2. Step II: Word embedding

Text is unstructured data that requires structuring before deriving patterns (Manning, 2015). This process, known as word embedding, traditionally has been traditionally carried out using the term frequency (TF) - inverse document frequency (IDF) approach (Taboada et al., 2011). This approach constructs document vectors using unique words from documents (Devlin et al., 2018; Manning et al., 1999).

A state-of-art approach accounts for entire text passages, embedding the text pre-processing step in the learning system (Devlin et al., 2018). Bidirectional Encoder Representations from Transformer (BERT) is a DL approach for semantic word embedding (Devlin et al., 2018); it has gained popularity in recent years because of its ability to outperform other word-embedding strategies (Amin et al., 2019; Sänger et al., 2019). The key technical aspect of BERT is its adaptation, to NLP, of the bidirectional training strategy of the popular transformer attention mechanism for machine translation. Its success relies on its ability to read an entire sequence of words at once, learning the context of a word according to its surroundings (left and right of the word). In this sense, BERT is "bidirectional", whereas previous word embedding approaches read text inputs sequentially, either from left to right or left-to-right and right-to-left combined (Devlin et al., 2018).

To achieve a bidirectional representation of sentences, BERT masks 15% of the words randomly from the input sentence to predict the masked words with the decoder (Devlin et al., 2018). The BERT system also follows a strategy to learn relationships between sentences. BERT reads pairs of sentences as input to learn whether the second sentence is the subsequent sentence in the text. To do so, 50% of inputs are pairs of sentences in the correct sequence in the text, and the other 50% is represented by a random sentence from the document that is assigned as the second sentence.

Finally, we study a hybrid word-embedding approach. We concatenate the TF-IDF vectors and the original comments in such a way that both sources are inputs for the BERT pre-trained model. The TF-IDF model is strongly related to some drivers; for example, when a customer mentions the words "price" or "cheap", the words clearly relate to the driver "price". Therefore, the idea is to achieve the best of both worlds by extracting the knowledge of single words with the TF-IDF framework while using BERT to model the relationships between words in the sentence. Our literature review did not identify any other simple strategy for combining the two strategies; therefore, our approach makes a novel development. However, there are other strategies that combine TF-IDF and BERT frameworks. For example, Schmidt (2019) proposes using weights according to the IDF function on a dense vector embedding from Word2vec or global vectors for word representation (GloVe). Despite this minor methodological development, we emphasize the fact that the novelty and contribution of our proposal is mostly applied.

3.3. Multi-label classification and BERT fine-tuning

In multi-class classification, the goal is classify instances into one of many possible classes (Herrera et al., 2016). In contrast, there are tasks in which a sample can have several labels (Herrera et al., 2016; Read et al., 2011; Wu et al., 2004). In this study, for example, a customer feedback example can have one or more drivers of customer satisfaction associated with it since the respondent may be referring to the price of a product he/she bought and the quality of the service experienced in the same comment. The main challenge is then to be able to identify the various drivers that may be present in a single input.

There are several studies on multi-label text classification. The best-known example is a classification of newswire articles, such that each

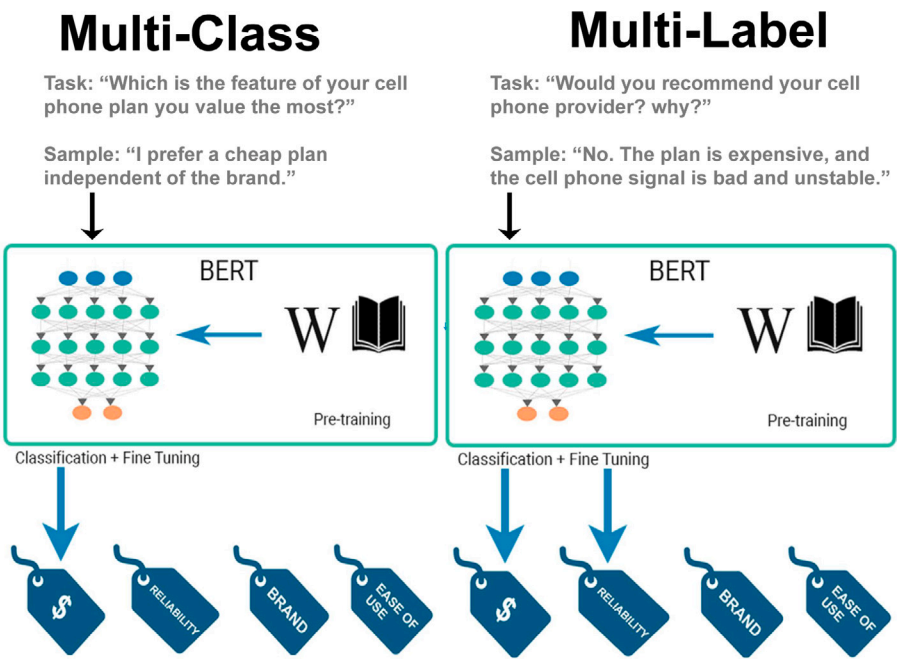

Fig. 2. Difference between multi-class and multi-label text classification using BERT.

article has one or several possible categories (McCallum, 1999). In this study, the authors achieved an 83.9% accuracy on a 10-label problem (the 10 most represented categories in the well-known Reuters-21578 corpus). Another application is the classification of biomedical texts (Du et al., 2019), in which the authors report results for three datasets. For the best predictive method, the F1 achieved was 81.5%, 73.8%, and 42.8% for tasks with 5, 8, and 70 labels, respectively. Debaere et al. (2018) reported results on seven datasets for member participation in online communities. For this task, the accuracy ranged from 64% to 83% for a three-label problem. Finally, Li et al. (2016) studies a classification model for sentiment analysis of online news. They reported F1 values for two datasets, achieving 31.4% and 50.6% for the best model on tasks with 6 and 8 labels, respectively.

Our proposed framework considers multi-label text classification via DL; it performs the word embedding process and the training of the classifier simultaneously. One advantage of language models such as BERT is "transfer learning", which fine-tunes published, transferable word vectors for specific data (Liu et al., 2019). We use a pre-trained BERT model to include published word vectors trained on a vast amount of data. This volume of data is important because survey-based customer feedback data sets usually are small, and DL models are complex artificial network architectures that require large sample sets to outperform alternative approaches (Goodfellow et al., 2016; LeCun et al., 2015).

Fig. 2 illustrates the difference between multi-class and multi-label text classification using BERT. If, for example, we want the customer to choose one alternative in the context of customer satisfaction (see left side of Fig. 2), a multi-class classification task will identify one and only one of the four available drivers. In contrast, multi-label classification considers the possibility that more than one of these alternatives can be identified in a single comment (see right side of Fig. 2).

Fig. 2 also depicts the functioning of BERT and the fine-tuning strategy. For the pre-training step, BERT is trained with large amounts of data for a general-purpose task called "language understanding". Some of the data sources considered for this step are books and Wikipedia. In this step, a deep network architecture is constructed by using the procedure described in the previous section: predicting masked words and the subsequent sentence in the text. The process of learning a new BERT architecture is computationally very demanding. However, pre-trained models are publicly available for researchers and practitioners. Using a pre-trained model as a starting point, we can adjust the weights of the model by defining an ad-hoc output layer (multi-label classification in our case) while feeding it with new data. This process is known as fine-tuning, and allows the customization of the model to a given task without losing the "language understanding" acquired during the pre-training step (Devlin et al., 2018).

We could use the generic word vectors obtained by the pre-trained approach in our documents directly. However, they would not incorporate any specific knowledge from our customer feedback data. The idea of transfer learning via BERT is to build custom word vectors for our corpus. We do so by modifying the last layer of the deep network for the multi-label text classification task, propagating the errors through the transformer. We consider using a standard cross-entropy loss function for multi-label classification based on sigmoid functions. This method constructs our classifier and improves and customizes the generic word vectors for our task, in a process known as fine-tuning (Liu et al., 2019). The fine-tuning step is key in our proposal since it allows the adaptation of models to different firms and/or industries with relatively few data requirements, assuming that pre-trained models are available.

3.4. Step IV: Prediction and interpretation of mined patterns

The final step is model validation and evaluation. For model validation, we use the standard procedure of dividing the entire data set into training, validation, and testing subsets. We construct the model in the training and validation sets to evaluate a given model and fine-tune the model's hyperparameters; our machine learning approaches, such as DL and SVMs, require this step. We then compare the models with the optimal hyperparameter configuration in the testing set, which remains unseen during the learning process. We consider 70% of the data set for training and use the remaining 30% for testing. From the training set, we use 10% of the data for validation and model selection.

To evaluate the model, we consider the F1 measure, which has been frequently used in multi-label classification tasks (Herrera et al., 2016). Given a series of documents, we let $\mathcal{T}_j$ and $\mathcal{S}_j$ be the set of true and predicted labels associated with a given class j. Precision corresponds to the true positives divided by the sum of all positives, that is, $P_j = |\mathcal{T}_j \cap \mathcal{S}_j| \,/\, |\mathcal{S}_j|$, where $|\cdot|$ represents the cardinality of a set. Recall corresponds to the true positives divided by the true positives

Table 2
Descriptive analysis for all industries.

Industry	Labels (L)	Comments (C)	Ratio L/C	Promoters	Detractors	NPS
Gas	2849	1477	1.93	71%	10%	61
Clinics	3031	1347	2.25	52%	16%	36
Gas stations	2871	1466	1.96	51%	17%	34
Retail cards	3907	2075	1.88	50%	25%	25
Banks	4875	2315	2.11	43%	25%	18
Retail	4116	1989	2.07	40%	27%	13
Mobile phone	8327	4214	1.98	43%	30%	13
Supermarkets	4095	1992	2.06	37%	29%	8
TV	3789	1983	1.91	36%	34%	2
Drug stores	2862	1490	1.92	33%	37%	−4
Health insurance	3994	1954	2.04	27%	39%	−12
AFP	4030	2121	1.9	26%	41%	−15
Internet	2233	1155	1.93	28%	44%	−16
Average	3921	1968	2.00	41.3%	28.8%	13

and false negatives: $R_j = |\mathcal{T}_j \cap S_j| \,/\, |\mathcal{T}_j|$. We then compute the F1 for this class as the harmonic mean of the precision and the recall: $F1_j = 2P_jR_j \,/\, (P_j + R_j)$. In multi-label classification, researchers often recommend using the weighted F1 measure, which is the sum of all $F1_j$ weighted by the support of each label $|\mathcal{T}_j| \,/\, sum_j|\mathcal{T}_j|$.

Finally, we discuss the predictive performance for the various economic sectors considered in our customer feedback data. This analysis allows us to interpret the results, make conclusions with regard to the success of the approach, and gain new insights into the application.

4. Experimental results

4.1. Dataset description and pre-processing

To test our methodology, we access a proprietary data set consisting of 25,943 responses from 39 Chilean companies in 13 different industries. This data set includes customer ratings based on the NPS framework presented in Section 2.1. We manually tagged all open-ended answers, identifying a minimum of 1 and a maximum of 3 of the 11 possible drivers of CSAT per response, according to the following process:

- To guarantee the quality of the data, the consultancy team we are collaborating with carefully designed and conducted the data collection strategy. This consultancy team has several years of experience in this type of survey, and they conducted phone interviews following the NPS framework with customers from a wide variety of firms and service industries.
- The original data set included the customer response and a field labeled "micro-driver". The micro-driver is a context-specific concept defined by the consulting team that provided the data to classify the responses. In the telecommunications industry, for example, one response was: "The cell phone signal is bad and unstable". In this case, the data set owner classified this response with the micro-driver "cell phone signal". There were 351 micro-drivers in the data set. The drivers discussed in Table 1 can be decomposed in these micro-drivers. In the previous example, the driver reliability can be linked to the cell signal or the expected lost traffic. Therefore, we can define the drivers associated with the methodology in terms of the micro-drivers available in our data. However, the definition of the micro-drivers is not a required step for the application of the proposed framework.
- We took a random sample from the data set to check manually whether micro-drivers were an adequate summary of the customer responses. We ran this process three times, concluding that the classifications were correct.
- We classified the micro-drivers in each of the 11 previously mentioned drivers. In this regard, we asked three persons trained in the meaning and scope of the 11 drivers to review and classify the 351 micro-drivers.
- Once the three persons issued their categorizations with total consensus, we set apart and approved the categorizations. We sent micro-drivers for which there was no consensus back to the reviewers to decrease discrepancies.
- We repeated this process three times, eventually reaching 97% consensus. Recent studies have followed a similar process (McColl-Kennedy et al., 2019).

To define our empirical setting, we opted for companies that (1) belong to different economic sectors, (2) have many customers (at least 10,000 clients), and (3) are well-known brands in Chile. We selected survey participants randomly, ensuring that they not only knew the company but also had experience(s) with companies in the given industries. Appendix A reports the various service industries included in our data set as supplementary material. Following the framework presented in the previous section, the results of the first step (i.e., data gathering, visualization, data processing) consist of descriptive information about the data set and graphical representations of the data that allow us to identify the first relevant pattern of decision-making. With regard to model implementation, we implemented all strategies in Python. We carried out the data pre-processing step (stopword removal and stemming) using the nltk library and used the sklearn library to implement the TF-IDF and SVM methods. We implemented the MLP network using the Keras library, and the DL strategies (BERT with and without fine-tuning) using the TensorFlow platform, following the original implementations of Google Research (Devlin et al., 2018).

Regarding the data pre-processing, we analyze the performance of different combinations of strategies for stopword removal and stemming, observing and slight improvement when these two techniques are considered. Notice that the preprocessing step can have a strong influence on the outcome of a text analytics model, and therefore should not be taken for granted.

Table 2 summarizes the relevant descriptive information for each industry; it presents the total number of labels (drivers) included in the comments, the number of comments (open-ended responses), the ratio between labels and comments, the fractions of promoters and detractors (in percentages), and the NPS scores, computed as the percentage of promoters (9s and 10s) minus the percentage of detractors (0s through 6s). As this summary shows, the data sets are relatively similar in size, with regard to both labels and comments. On average, and for all service industry sectors, each respondent mentions two drivers. In contrast, the distribution of the NPS scores is highly skewed, ranging from 61% (Liquid/Natural gas providers) to −16% (Cable/Fiber internet providers). This result confirms the importance of creating models tailored to each industry.

Further information on the dataset and the pre-processing steps is presented in the Appendix A provided as supplementary material. Table A.1 describes the various service industries considered in this study. Next, Figure A.1 illustrates the proportion of drivers present in each industry. From this figure we can conclude that the label distribution

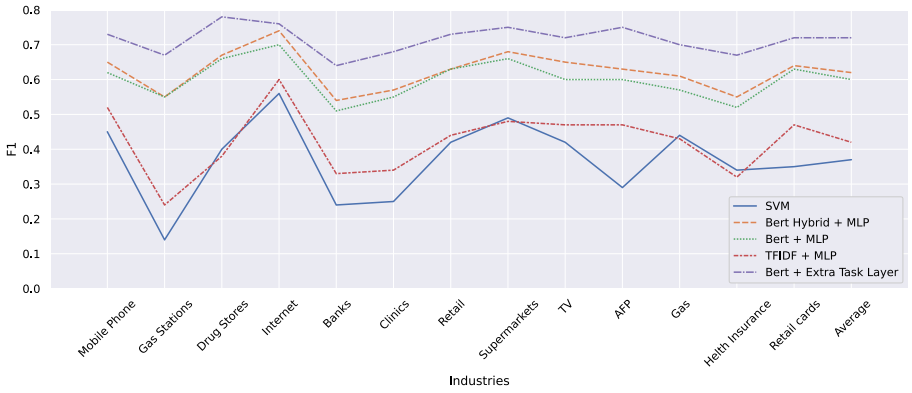

Fig. 3. F1 weighted for each industry.

clearly varies across industries, and several drivers are underrepresented in all industries. Finally, we developed word clouds to illustrate the impact of key concepts and words, declared by respondents, on NPS scores. For illustrative purposes, we present the word cloud associated with the retail sector and use bigrams to reflect the importance of a concept. Figure A.2 depicts the word cloud.

4.2. Results summary

For completeness, we compare the BERT strategy (with fine-tuning) with three alternative strategies: TF-IDF, pre-trained BERT (without fine-tuning), and the previously mentioned hybrid TF-IDF/BERT approach. For the alternative approaches, the word embedding process is independent of the learning machine. The vector representations of words are passed as inputs to traditional machine learning techniques, such as multi-layer perceptron (MLP) networks and support vector machines (SVMs). We empirically test the following multi-label classification strategies: TF-IDF embedding in combination with SVM, TF-IDF embedding in combination with an MLP network, pre-trained BERT in combination with an MLP network, the proposed hybrid word embedding approach that encompasses pre-trained BERT and TF-IDF in combination with an MLP network, and BERT with an extra layer for simultaneous classification and embedding of fine-tuning.

We acknowledge the existence of several strategies that are significantly better than the TF-IDF approach, such as BI-LSTM, GRU networks, or ELMO (Goodfellow et al., 2016; Liu et al., 2019). Although they are all worse than BERT in general due to the latter technique's ability to perform fine-tuning and the reasoning behind the transformers, they can achieve competitive results. The comparison between a state-of-the-art technique such as BERT and TF-IDF is key in this study to convince business analytics researchers and practitioners of the virtues of BERT and transformers in general over the standard approach in the field, which is still TF-IDF.

Following a standard methodology for applied research, we have explored different parameter combinations for the deep learning strategy, monitoring its performance across the different number of epochs on the validation set. We explored the following values for the mini-batch sizes: 32, 64, and 128. We trained the networks up to 20 epochs. We achieved the optimal validation performance with 10 epochs and a batch size of 32. Regarding the solver for the BERT model, we consider the stochastic gradient-based optimizer ADAM in its default configuration (initial learning rate of 0.001, and exponential decay rates for estimates of first and second moment vectors $beta_1 = 0.9$ and $beta_1 = 0.999$, respectively). This default configuration allows ADAM to adaptively find a suitable combination of the learning rate and momentum parameters without the need of a potentially time-consuming model selection process.

Regarding the alternative methods (shallow ANN and SVM), we consider the default configurations of the scikit-learn library. For the case of SVM, it consisted of a trade-off parameter $C = 1$ and a width γ associated with the Radial Basis Function (RBF) kernel of $1/n * var(X)$, with n being the dimensionality of the TF-IDF matrix X and $var(X)$ its variance.

Table 3
Performance summary for all multi-label classification approaches. General results.

Method	*F1 (weighted)*
TF-IDF + SVM	*0.37*
TF-IDF + MLP	*0.42*
BERT + MLP	*0.60*
Hybrid BERT/TF-IDF + MLP	*0.62*
BERT + fine-tuning	***0.72***

Table 3 reports the predictive performance in terms of weighted F1. It is important to notice that a good performance implies that the model is successful at identifying the drivers of customer experience automatically, allowing an efficient analysis of customer feedback data.

We can draw several conclusions from Table 3. First, the MLP network performs better than SVM when we use TF-IDF as a word embedding strategy. Therefore, we consider this classification approach for the remaining experiments. Second, using pre-trained BERT embedding improves the results significantly (from F1 = 0.42 to F1 = 0.60), thereby confirming the virtues of a semantic word embedding approach using vectors learned via AI. The hybrid BERT/TF-IDF approach provides only a slight improvement in classification performance (from F1 = 0.60 to F1 = 0.62), demonstrating the capacity of BERT to incorporate the basic patterns from words that summarize the TF-IDF framework. Third, we achieve the best performance when we introduce an extra layer of multi-label classification in the BERT architecture, such that we can fine-tune the word embeddings while training the classification model. The result of this approach is F1 = 0.72, which is an excellent performance for a multi-label classification task with 11 labels.

It is not always easy to assess whether the model is accurate or not in a multi-label setting. Furthermore, we identify up to three CSAT factors in one open-ended comment, introducing an extra complexity to the prediction task. However, note that an F1 of 0.72 can be interpreted such that each driver is correctly identified in an open-ended comment 72% of the time. As a reference, a random classifier would achieve a correct classification approximately 9% of the time (1/11). This positive result allows us to conclude that drivers can be identified automatically in survey data with very good accuracy, resulting in an important tool for decision-making in marketing and service management.

Fig. 3 reports the predictive performance of each classification method disaggregated at the industry level. In this figure, we observe that the best model (BERT with fine-tuning) shows little variability among the service industries studied in this paper, with results ranging

Table 4
Performance summary for all industries and drivers. BERT with fine-tuning as classification approach.

Industry	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Gas	0.56	0.54	0.17	0.59	0.37	0.48	0.27	0.56	0.75	0.84	**0.87**
Clinics	0.63	0.80	0.74	0.48	0.00	0.59	0.52	0.00	**0.85**	0.68	0.71
Gas stations	0.68	0.57	0.79	0.51	0.00	0.59	0.71	0.67	**0.84**	0.08	0.68
Retail cards	0.69	0.15	0.00	0.38	0.39	0.67	0.13	0.73	**0.90**	0.71	0.60
Banks	0.36	**0.76**	0.71	0.61	0.33	0.54	0.57	0.52	0.75	0.61	0.69
Retail	**0.88**	0.75	0.70	0.47	0.50	0.46	0.67	0.00	0.78	0.48	0.65
Mobile phone	0.56	0.44	0.38	0.29	0.26	0.34	0.24	0.49	0.78	**0.92**	0.77
Supermarkets	**0.89**	0.64	0.73	0.26	0.00	0.00	0.59	0.00	0.82	0.46	0.73
TV	0.36	0.00	**0.85**	0.29	0.71	0.29	0.29	0.52	0.77	0.84	0.81
Drug stores	0.85	0.67	0.73	0.61	0.00	0.00	0.72	0.00	**0.85**	0.20	0.84
Health insurance	0.78	0.44	0.48	0.45	0.42	0.58	0.52	0.26	**0.83**	0.57	0.71
AFP	0.38	0.51	0.86	0.71	0.00	0.65	0.65	0.81	0.81	**0.85**	0.74
Internet	0.44	0.00	0.00	0.33	0.00	0.38	0.00	0.46	0.67	**0.92**	0.78
Average	0.62	0.48	0.55	0.46	0.23	0.43	0.45	0.39	**0.80**	0.63	0.74

[1]Assortment [5]Convenience [9]Price
[2]Assurance [6]Ease of use [10]Reliability
[3]Atmosphere [7]Empathy [11]Responsiveness
[4]Brand [8]Past customer experience

from $F1 = 0.64$ to $F1 = 0.78$ for this best method. Accordingly, we conclude that our framework successfully identifies the drivers for any industry. Furthermore, we can conclude that the proposed framework achieves stable results across industries.

Another important analysis assesses the ability of our model to identify each driver correctly for each industry. Table 4 reports the results of this analysis, showing the F1 measures for each label and industry using BERT with fine-tuning and highlighting in bold the label with the highest $F1$ measure for each industry.

According to Table 4, the best accuracy emerges for the driver "price", followed by "responsiveness". Machine learning can identify these drivers easily and link them directly to words in open-ended answers, such as "cheap", "expensive", or "price", for the "price" driver. In contrast, drivers such as "convenience" or "past customer experience" are more abstract concepts that may require customers to use entire sentences to explain. Therefore, such drivers are more challenging for machines and affect their predictive performance negatively.

Table 4 also reveals important differences across industries for every driver. In some cases, the F1 drops to 0, even when the same drivers are classified almost perfectly in other industries. This drop is the result of data availability: In most cases, there are not enough data samples at the industry level to explain the labels accurately.

4.3. Managerial insights

The positive results achieved with the proposed framework have several implications for researchers and practitioners. From a managerial perspective, we show that a learning machine can be constructed using approximately 4000 labeled responses (the average sample size for the various industries; see Table 2). In turn, this research has several important implications for managers:

- This framework can be used to improve decision making by understanding the key drivers that affect CSAT in a given company, allocating customers as promoters or detractors, and justifying their scores according to these key drivers. By identifying the main drivers of CSAT automatically, managers can quickly implement changes to address issues that customers identify as problematic (i.e., responses rated from 0 to 6 in the NPS framework). They can also use topics that correlate with high ratings to improve their advertising messages. Although our empirical setting is based on an NPS survey, our approach can be used with any other metric of customer satisfaction that considers both customer ratings and reasons that support customer evaluations of satisfaction. The approach can be extended further to automatically identify the drivers in customer complaint data, allowing companies to provide customized service to dissatisfied customers efficiently.
- We propose a robust, accurate text classification methodology that works well in any industry (see Fig. 3). Firms can use this framework to predict drivers of CSAT in any text data repository they have, such that managers can quickly identify and summarize customer comments that might ultimately affect customer satisfaction. Customer enquiries and complaints can be prioritized according to the most relevant drivers in an automated fashion.
- As Table 4 shows, the model accurately predicts the drivers "price", "reliability", and "responsiveness" (i.e., percentage of accuracy is greater than 0.63). These drivers are critical to CSAT because they show how firms honor their value propositions (reliability), how they respond to customers, and how they charge for services or products. By applying our method, firms can automatically detect customer feedback related to these three main drivers and act accordingly.
- Our proposed method is simple and cost-effective for firms. Although this process needs a calibration step (identification of micro-drivers and 11 drivers), it then works automatically; any firm can use it. In other words, our approach is valid for any industry, although its success relies on the ability to feed the model with sufficient data from a given industry and/or company. However, the use of a pre-trained model such as BERT alleviates the requirements in terms of input data.

Our recommendation is to develop industry-level models to analyze the complexities of the customer feedback, which clearly varies across different domains in terms of the words and concepts that are relevant for customer satisfaction. For example, the (cell phone) signal is a key aspect of customer satisfaction that defines reliability in the mobile communications sector, but this concept becomes irrelevant in other industries.

The main issue in constructing industry-level models is, however, data availability. In the case of our project, we have less than 5000 comments in each industry. This may not guarantee robust results for a multi-label task with 11 classes. The use of fine-tuning certainly helps, but the fact that not all the classes are well-represented in some domains causes these differences in terms of performance across industries.

Another important recommendation is to analyze the dynamic aspects of the expert system. Marketing applications such as CX face dynamic environments that cause to changes in the data distribution, negatively affecting the predictive performance of machine learning models (Bravo & Maldonado, 2015). This issue is known as "dataset shift" and can be addressed via model monitoring or "backtesting" (Bravo & Maldonado, 2015). Although the literature in this area is scarce for textual data, we can recommend monitoring the proportion of predicted drivers in the customer responses to identify major changes in the data. Traditional backtesting metrics such as the population stability index (PSI) or the χ^2 statistic (Bravo & Maldonado, 2015) can be adapted for this task. If major changes are identified, the solution is to recalibrate the deep learning model using up-to-date data.

Our study also has important implications for researchers. We propose a novel predictive task, which can be successfully addressed via DL and NLP. We show that traditional machine learning methods such as SVM or shallow ANN are not able to perform adequately on this task, and only the recent advances in transfer learning and transformers allow an adequate experimental design. Furthermore, we present a novel multilabel classification task, which is relevant for the AI community as it defines a challenging predictive problem, especially when the number of labels per observation is not constant. The proposed task is also one of the first multilabel classification applications in business analytics.

5. Conclusions

In this study, we use a novel text-mining framework to gain insights from customer feedback data, a key source of customer information in the service industry. Our goal is twofold: to extract knowledge from ratings and open-ended surveys via descriptive analytics and word clouds, by identifying simple patterns for improving decision making, and to implement a learning machine that classifies open-ended answers according to the drivers of CSAT that marketing literature acknowledges. With regard to our second goal, we succeeded at this difficult task because of recent advances in DL, text mining, and transfer learning. Traditional text mining approaches, such as TF-IDF embedding in combination with standard artificial neural networks or SVMs, achieve an average of only about 40% classification accuracy for each driver; in contrast, BERT and transfer learning for multi-label classification achieve about 70% classification accuracy (for 11 labels), which is an extremely good result for text classification. We thus conclude that recent advances in NLP and DL open up great opportunities in the marketing world for automatic tagging of customer feedback, complaints, social media data, and surveys.

The main conclusion drawn from the experimental section is that we illustrate empirically how BERT is successful in terms of its ability to identify the drivers of customer satisfaction in open-ended survey data. This suggests that a successful implementation is possible. In contrast, traditional machine learning methods such as shallow networks or SVM in combination with the TF-IDF framework are not able to achieve good results. In other words, the novel framework for the automatic interpretation of customer feedback, which is the main contribution of this study, is only possible thanks to the recent advances in deep learning, transformers, and transfer learning.

We note, however, that our proposed approach is limited when there is insufficient data to guarantee correct classification of all labels for all industries. Multi-label classification is a very complex task, especially when large numbers of labels are considered. In our case, we aim to predict 11 drivers in industries using fewer than 1000 answers -very ambitious goal, even for the rich data set we have. Our results for each industry and driver of CSAT, reported in Table 4, reveal a key problem for deep learning and text classification: The success of a predictive task depends strongly on the number of labeled examples available for training (Du et al., 2019). This limitation implies that with regard to drivers that are not frequent in our data sets, our conclusions should be accepted with caution.

The proposed framework was constructed using a curated dataset for text analytics obtained via phone surveys and transcribed manually by the phone operator. We believe our model can be applied to other forms open-ended responses, such as those written directly by individuals or transcribed automatically via ASR (automatic speech recognition) techniques. Inevitably, the model would be less accurate because of possible orthographic mistakes by the respondents or ASR errors. Nevertheless, we strongly believe that our framework can perform well in such cases, but an empirical analysis is not possible due to data availability. Such an analysis represents an interesting avenue for future research.

Our approach contributes to the marketing literature by being a newly reported system of automatic customer feedback labeling via text classification. To date, most NLP approaches have been unsupervised, known as topic models, designed to discover abstract topics that occur in documents. Latent Dirichlet allocation (LDA) (Blei et al., 2003) is arguably the most popular topic model, and it has been previously used in opinion mining (Pietsch & Lessmann, 2019). Although supervised and unsupervised approaches are not directly comparable, the task of text classification is more effective when guided by observations that are labeled according to the goal of a study. Our study also provides a novel AI application that can be tackled only via state-of-the-art NLP techniques.

This study represents the first attempt for the automatic analysis of customer satisfaction via machine learning techniques, which proved to be successful in terms of accuracy. We believe that the rapid development of deep learning for text mining will open new research opportunities and make this technology practical for practitioners in the service industry. One example is SEER, the self-supervised learning machine proposed recently by Facebook AI for computer vision (Goyal et al., 2021). Self-supervised learning allows the development of learning machines with fewer requirements in terms of labeled data. Another expected advance is data availability. It is not far-fetched to expect large repositories of customer feedback data with the adequate information for training models such as the one presented in this paper.

CRediT authorship contribution statement

Ángeles Aldunate: Contributed data or analysis tools, Performed the analysis, Wrote the paper. **Sebastián Maldonado:** Conceived and designed the analysis, Contributed data or analysis tools, Wrote the paper. **Carla Vairetti:** Conceived and designed the analysis, Wrote the paper. **Guillermo Armelini:** Collected the data, Contributed data or analysis tools, Wrote the paper.

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 that has been used is confidential.

Acknowledgments

The authors gratefully acknowledge financial support from ANID, PIA-BASAL AFB180003 and FONDECYT-Chile, grants 1200221 and 12200007. The authors would like to thank Slodoban Ivanovic for his valuable work on this project, and Alco Consulting for providing the necessary information for this research. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions for improving the quality of the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.eswa.2022.118309.

References

Aguwa, C., Olya, M. H., & Monplaisir, L. (2017). Modeling of fuzzy-based voice of customer for business decision analytics. *Knowledge-Based Systems*, *125*, 136–145.

Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A., & Wixted, M. K. (2019). *MLT-DFKI at CLEF eHealth 2019: Multi-label classification of ICD-10 codes with BERT: CLEF (working notes)*.

Anderson, E. W., & Fornell, C. (2000). Foundations of the American customer satisfaction index. *Total Quality Management*, *11*(7), 869–882.

Baker, J., Parasuraman, A., Grewal, D., & Voss, G. B. (2002). The influence of multiple store environment cues on perceived merchandise value and patronage intentions. *Journal of Marketing*, *66*(2), 120–141.

Blei, D. M., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bravo, C., & Maldonado, S. (2015). Fieller stability measure: a novel model-dependent backtesting approach. *Journal of the Operational Research Society*, *66*(11), 1895–1905.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 step-by-step data mining guide: Technical report*, The CRISP-DM consortium.

Coussement, K., & Van den Poel, D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, *44*(4), 870–882.

De Haan, E., Verhoef, P. C., & Wiesel, T. (2015). The predictive ability of different customer feedback metrics for retention. *International Journal of Research in Marketing*, *32*(2), 195–206.

Debaere, S., Coussement, K., & De Ruyck, T. (2018). Multi-label classification of member participation in online innovation communities. *European Journal of Operational Research*, *270*(2), 761–774.

Decker, R., & Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*, *27*(4), 293–307.

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

Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). ML-NEt: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, *26*(11), 1279–1285.

Gargiulo, F., Silvestri, S., Ciampi, M., & De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, *79*, 125–138.

Gauri, D. K., Sudhir, K., & Talukdar, D. (2008). The temporal and spatial dimensions of price search: Insights from matching household survey and purchase data. *Journal of Marketing Research*, *45*(2), 226–240.

Goldberg, A. B., & Zhu, X. (2006). Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the first workshop on graph based methods for natural language processing* (pp. 45–52). Association for Computational Linguistics.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press, http://www.deeplearningbook.org.

Goyal, P., Caron, M., Lefaudeux, B., Xu, M., Wang, P., Pai, V., Singh, M., Liptchinsky, V., Misra, I., Joulin, A., & Bojanowski, P. (2021). Self-supervised pretraining of visual features in the wild.

Gregoriades, A., Pampaka, M., Herodotou, H., & Christodoulou, E. (2021). Supporting digital content marketing and messaging through topic modelling and decision trees. *Expert Systems with Applications*, *184*, Article 115546.

Grewal, D., Levy, M., & Kumar, V. (2009). Customer experience management in retailing: an organizing framework. *Journal of Retailing*, *85*(1), 1–14.

Guenther, M., & Guenther, P. (2020). The complex firm financial effects of customer satisfaction improvements. *International Journal of Research in Marketing*.

Gupta, S., & Zeithaml, V. (2006). Customer metrics and their impact on financial performance. *Marketing Science*, *25*(6), 718–739.

Herrera, F., Charte, F., Rivera, A. J., & Del Jesus, M. J. (2016). Multilabel classification. In *Multilabel classification* (pp. 17–31). Springer.

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172.

Huang, M.-H., & Rust, R. T. (2021). Engaged to a robot? The role of AI in service. *Journal of Service Research*, *24*(1), 30–41.

Huffman, C., & Kahn, B. E. (1998). Variety for sale: mass customization or mass confusion? *Journal of Retailing*, *74*(4), 491–513.

Janakiraman, N., Meyer, R. J., & Morales, A. C. (2006). Spillover effects: How consumers respond to unexpected changes in price and quality. *Journal of Consumer Research*, *33*(3), 361–369.

Jerger, C., & Wirtz, J. (2017). Service employee responses to angry customer complaints: The roles of customer status and service climate. *Journal of Service Research*, *20*(4), 362–378.

Joung, J., Jung, K., Ko, S., & Kim, K. (2019). Customer complaints analysis using text mining and outcome-driven innovation method for market-oriented product development. *Sustainability*, *11*(1), 40.

Kaltcheva, V. D., & Weitz, B. A. (2006). When should a retailer create an exciting store environment? *Journal of Marketing*, *70*(1), 107–118.

Keller, K. L. (2003). Brand synthesis: The multidimensionality of brand knowledge. *Journal of Consumer Research*, *29*(4), 595–600.

Koehn, D., Lessmann, S., & Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, *150*, Article 113342.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research*, *281*(3), 628–641.

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. *Journal of Big Data*, *6*(1), 62.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey. *Journal of Marketing*, *80*(6), 69–96.

Li, X., Xie, H., Rao, Y., Chen, Y., Liu, X., Huang, H., & Wang, F. L. (2016). Weighted multi-label classification model for sentiment analysis of online news. In *2016 international conference on big data and smart computing (BigComp)* (pp. 215–222). IEEE.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on information and knowledge management* (pp. 375–384).

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. arXiv preprint arXiv:1901.11504.

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

Manning, C. D. (2015). Computational linguistics and deep learning. *Computational Linguistics*, *41*(4), 701–707.

Manning, C. D., Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martínez Cámara, E., Martín Valdivia, M. T., Perea Ortega, J. M., & Ureña López, L. A. (2011). Opinion classification techniques applied to a spanish corpus. *Procesamiento Del Lenguaje Natural*, *47*, 163–170.

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

McCallum, A. (1999). Multi-label text classification with a mixture model trained by EM. In *AAAI workshop on text learning* (pp. 1–7).

McCauley, D. (2020). *The global AI agenda: Promise, reality, and a future of data sharing*. MIT Technology Review Insights.

McColl-Kennedy, J. R., Zaki, M., Lemon, K. N., Urmetzer, F., & Neely, A. (2019). Gaining customer experience insights that matter. *Journal of Service Research*, *22*(1), 8–26.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113.

Miguéis, V. L., & Nóvoa, H. (2017). Exploring online travel reviews using data analytics: An exploratory study. *Service Science*, *9*(4), 315–323.

Monsuwé, T. n. P. Y., Dellaert, B. G., & De Ruyter, K. (2004). What drives consumers to shop online: A literature review. International journal of service industry management. In *Information systems res*. Citeseer.

Noble, S. M., & Phillips, J. (2004). Relationship hindrance: why would consumers not want a relationship with a retailer? *Journal of Retailing*, *80*(4), 289–303.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). Servqual: A multiple-item scale for measuring consumer perc. *Journal of Retailing*, *64*(1), 12.

Pietsch, A.-S., & Lessmann, S. (2019). Topic modeling for analyzing open-ended survey responses. *Journal of Business Analytics*, *1*(2), 93–116.

Ramaswamy, S., & DeClerck, N. (2018). Customer perception analysis using deep learning and NLP. *Procedia Computer Science*, *140*, 170–178.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, *85*(3), 333.

Reichheld, F. F., & Markey, R. (2011). *The ultimate question 2.0: how net promoter companies thrive in a customer-driven world*. Harvard business Press.

Robles, J. F., Chica, M., & Cordon, O. (2020). Evolutionary multiobjective optimization to target social network influentials in viral marketing. *Expert Systems with Applications*, *147*, Article 113183.

Sänger, M., Weber, L., Kittner, M., & Leser, U. (2019). *Classifying german animal experiment summaries with multi-lingual BERT at CLEF eHealth 2019 task 1: CLEF (working notes)*.

Schmidt, C. W. (2019). Improving a TFIDF weighted document vector embedding.

Soudagar, R., Iyer, V., & Hildebrand, V. (2011). *The customer experience edge: technology and techniques for delivering an enduring, profitable and positive experience to your customers*. McGraw Hill Professional.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307.

Ur-Rahman, N., & Harding, J. A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, *39*(5), 4729–4739.

Vairetti, C., Martínez, E., Maldonado, S., Herrera, F., & Luzón, M. (2020). Enhancing the classification of social media opinions by optimizing the structural information. *Future Generation Computer Systems*, *102*, 838–846.

Van Doorn, J., & Verhoef, P. C. (2008). Critical incidents and the impact of satisfaction on customer share. *Journal of Marketing*, *72*(4), 123–142.

Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. *Journal of Retailing*, *85*(1), 31–41.

Villarroel-Ordenes, F., Theodoulidis, B., Burton, J., Gruber, T., & Zaki, M. (2014). Analyzing customer experience feedback using text mining: A linguistics-based approach. *Journal of Service Research*, *17*(3), 278–295.

Wang, Y., Lu, X., & Tan, Y. (2018). Impact of product attributes on customer satisfaction: An analysis of online reviews for washing machines. *Electronic Commerce Research and Applications*, *29*, 1–11.

Wang, Y., & Tseng, M. M. (2015). A Naïve Bayes approach to map customer requirements to product variants. *Journal of Intelligent Manufacturing*, *26*(3), 501–509.

Warren, C., Batra, R., Loureiro, S. M. C., & Bagozzi, R. P. (2019). Brand coolness. *Journal of Marketing*, *83*(5), 36–56.

Wu, T.-F., Lin, C.-J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, *5*(Aug), 975–1005.

Yang, Z., Sun, S., Lalwani, A. K., & Janakiraman, N. (2019). How does consumers' local or global identity influence price–perceived quality associations? The role of perceived quality variance. *Journal of Marketing*, *83*(3), 145–162.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, *36*(3), 6527–6535.

Yu, C. H., Jannasch-Pennell, A., & DiGangi, S. (2011). Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qualitative Report*, *16*(3), 730–744.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing*, *60*(2), 31–46.

Zeithaml, V. A., & Parasuraman, A. (2004). *Service quality*. Marketing Science Institute.