Measuring Service Quality Based on Customer Emotion: An Explainable AI Approach

Yiting Guo, Yilin Li, De Liu, Sean Xin Xu

Abstract

This paper develops an explainable artificial intelligence (AI) approach to measuring service quality in voice-based service encounters. Drawing from the psychology and computer science literature, we construct features of a customer's emotion dynamics during a service encounter. Using real-world call center data from a large insurance company, we train an ensemble model with these emotion dynamics features to predict service quality. The model has higher prediction performance than the two benchmark approaches using quality-assurance evaluation and operational indices. Our method for emotion dynamics classification outperforms a host of state-of-the-art time series classification algorithms. We further apply explainable AI methods to identify the most important features of emotion dynamics and show how they are related to service quality. For example, the location where the last emotion episode appears in a service call has a U-shaped relationship to low quality. Finally, to demonstrate utility, we design an IT artifact to automatically measure service quality after service encounters in the call center and use the measure to predict a customer's referral intention.

Keywords: service quality, emotional intelligence, explainable AI, referral, time series classification

If you can't measure, your knowledge is meager and unsatisfactory. —Lord Kelvin

1. Introduction

Firms must constantly track how customers perceive the services they receive—namely *service quality*—because service quality influences critical outcomes such as customer loyalty, word-of-mouth, firm revenue, and long-term survivability [1]. As of 2018, U.S. businesses lose more than $75 billion a year because of poor customer service.[1] A key challenge for businesses in managing service quality is measuring it accurately and effectively. This paper seeks to improve the measurement of call center service quality based on customer emotion dynamics and explainable artificial intelligence (AI) techniques.

Although service quality measures have long been an important topic in management fields including service science [2] and information systems [3], the existing approaches have notable drawbacks. Until recently, call centers primarily used three approaches to measure service quality: customer surveys, manual quality inspections, and operational indices. The first two rely on manual work and are time-consuming and unscalable. Customer surveys usually have low response rates, some customers find them intrusive [4], and there may be a gap between quality-inspection results and customers' perception of service [5]. The third approach, using operational indices (such as call time and delay), can work automatically, but most of these indices focus on call-handling efficiency instead of the underlying service quality [6]. In practice, there is a lack of real-time, efficient systems for managing the quality of call center service. At the same time, scholars have yet to fully study the prediction model that specializes in service quality.

The rapid development of digital technologies, big data, and AI technologies provides new opportunities to improve the measurement of service quality. Given their transformative potential, AI

---

[1] https://www.forbes.com/sites/shephyken/2018/05/17/businesses-lose-75-billion-due-to-poor-customer-service/

techniques can be applied to the design of intelligent artifacts to solve business problems [7]. In particular, AI has advantages in mining unstructured multimedia data (images, text, speech, etc.) recorded during the service encounter, offering much richer information and possibly obtaining more profound insights. Call centers have especially accumulated massive recorded call data that contain customer speech in time series. Moreover, AI provides an unprecedented opportunity for theory development [8]. Even though most AI techniques are seen as black boxes because of their lack of transparency, explainable AI could help extract knowledge and discover complex relationships between predictors and targets [9].

The present research proposes a theory-driven approach with the capacity for prediction and interpretation based on explainable AI. Specifically, this approach uses customer emotion during service encounters to measure service quality. The approach is grounded in the service science literature, which documents that customers' emotions during service encounters correlate with how they perceive service quality [10]. We extract features (of customer emotion) using this theoretical guidance, which helps avoid overfitting [11]. On the technical side, applications for tracking people's emotions are increasingly available. These two pillars—theoretical and technical—motivate us to leverage customer emotions to develop a measure of service quality.

Furthermore, we explore *emotion dynamics*—the dynamics of customer emotions throughout a service encounter—which refers to "changes and fluctuations in people's emotional and affective states over multiple points in time" [12]. Connecting those changes and fluctuations in customer emotions with the outcome of customer service offers a rich context for identifying new relationships between customer emotion and service quality, which, to the best of our knowledge, are under-researched. From the technological perspective, there remain some challenges in analyzing customer emotion series. The existing time series classification (TSC) methods have not been applied to emotion sequences and are limited because they are not interpretable and usually focus on merely identifying shapes or measuring global distances between time series. Therefore, this paper proposes

a framework for analyzing customer emotion dynamics including feature extraction, prediction, and explanation.

We work with a large insurance company, which provides us with data from three sources:

- a sample of recorded service calls (i.e., service encounters)

- ground truth: results of customer surveys for assessing service quality

- benchmarks: results of other approaches to service evaluation, including (a) quality-inspection-team assessments and (b) operational indices

This paper trains an ensemble model that uses six sub-categories of customer-emotion dynamic features to predict service quality. We further use explainable AI methods (i.e., permutation feature importance and accumulated local effects) to obtain the contribution of each emotion dynamics feature and to uncover its relationship to service quality. Finally, we design a real-time system based on automatic speech emotion recognition and assess the value of service-quality prediction in inferring customers' referring intentions.

Our study makes four contributions:

(1) *Measurement*. this study develops a new method to measure service quality in call centers. Based on customer emotion, this method is superior to two benchmarks (based on quality-inspection evaluation and operational indices) for assessing service quality. It offers many benefits, including reduced costs, reduced latency in handling service problems, and scalability.

(2) *Insights*. Our explainable AI approach unveils *what* dynamic characteristics of consumer emotions influence service quality, and *how*.

(3) *Technical*. When executing (1) and (2), we propose a TSC framework that uses customer emotion dynamics to predict service quality. It outperforms the state-of-the-art TSC methods in our research setting.

(4) *Practice*. this study substantiates the utility of the new measure by (a) designing an IT

artifact that can automatically track customer emotion and then measure service quality, and (b) showing that such emotional intelligence can be used to predict customers' referral intention.

2. Background

2.1. Approaches to assessing service quality

Call centers typically use three approaches to assess service encounters: operational indices, manual quality inspection, and customer feedback.

Common operational indices include the average speed to answer, abandonment rate, average talk time, calls per agent, and longest delay. These operational indices are easily quantifiable and are automatically recorded [13], but they tend toward efficiency instead of the quality of services [6]. Studies have repeatedly shown that superior operational indices do not necessarily translate to superior customer experiences [14].

Call centers also routinely employ specialists who assess service quality by listening to recorded service calls, and evaluating service employees' call-handling processes, protocol compliance, call-handling skills, and etiquette against the center's service quality standards [6]. Earlier studies identified service employees' performance as a significant factor affecting service quality [15,16]. However, quality inspections tend to emphasize standardized service processes, which may not correspond to a customer's specific needs. Moreover, quality inspections are costly and labor-intensive. Consequently, call centers typically inspect only a small fraction (e.g., 2%) of all recorded service calls.

Call centers also often follow up with customers and obtain their feedback on service quality through phone or mail surveys. The most widely used questionnaire is the service quality (SERVQUAL) scale developed by [17]. Customer feedback is seen as the ultimate way to capture service quality [17]. However, gathering customer feedback is time-consuming and resource-intensive [13]. Such surveys also suffer from low response rates and may disturb customers. The

approach described in this paper differs from the conventional approaches by automatically

predicting service quality using widely available call audio data.

2.2. Emotion dynamics and service quality

Our design of artifacts is guided by theories about the relationship between customer emotion

and service quality, which conforms to the design science paradigm [18]. Both theoretical and

empirical studies suggest that during service-encounter contexts, especially for high-contact services,

emotion is a crucial determinant of the perception and evaluation of the service experience [19–21].

If customers have a positive emotional reaction to focal service performance, their subsequent

assessment of service and satisfaction is likely to be positive [20]. Therefore, their emotions provide

clues about how they assess service quality.

Research has mostly studied emotions as a single-state consequence of a service encounter,

paying too little attention to their dynamics [22]. Yet studies show that customers' experiences and

perceptions evolve during service encounters and may affect the final perceived service quality [15].

Therefore, the present research treats customer emotion as a process instead of a single state.

Prior psychological studies have examined the relationship between people's emotion patterns

and several psychological outcomes, such as depression [23], well-being [12], emotion regulation

[24], and the results of psychotherapy [25]. This paper adds to this literature by relating emotion

dynamics to perceived service quality.

Emotion dynamics can be categorized into two general groups: trajectory-based and episode-

based [26]. Emotion trajectories are the recorded sequences of emotions over a given period,

presenting the continuous, ongoing levels of emotions [26], which reflect the global change patterns

of emotion trajectories over that period. An emotion episode presents a sub-sequence range from the

beginning to the end of an emotion, usually equal to one or more utterances expressing emotion,

revealing the characteristics of specific emotion episodes, such as an emotion's *duration*. The most

commonly applied emotion features, such as *variability*, *instability*, and *inertia*, are all trajectory-

based; for example, [12,23]. One notable omission in prior literature is time-specific features. For example, emotions at the end of an encounter may have different implications from emotions at the beginning of the encounter. The present research builds on existing emotional dynamic features and enriches the set to include time-specific features, such as the locations of negative emotions.

2.3. Time series classification (TSC) methods

The problem of predicting low-service-quality calls from a customer's emotion sequence is a special case of TSC. TSC models can be classified into three main categories: distance-based, feature-based, and model-based. Distance-based methods first measure the similarity between a whole series using elastic distances (such as dynamic time warping, DTW) and then apply k-nearest-neighbor (KNN) models to perform the classification. DTW-KNN is the model most frequently used in this category [27]. Feature-based methods map the original series to a feature space. Frequently used features in TSC tasks include statistics of time series intervals (e.g., means, standard deviations, and slopes) and series decomposition in the frequency or time domain. Representative feature-based methods include a time series forest (TSF) [28], Shapelets [29], and a bag of SFA symbols (BOSS) [30]. Model-based methods typically use generative models (e.g., hidden Markov models and autoregressive models). In addition, deep learning models have gained interest in recent years. However, perhaps because deep learning models are highly complex, studies have not found them to be superior to other methods [31]. Finally, some studies combine two or more of these approaches to achieve better classification performance, such as the elastic ensemble (EE) model [27] and HIVE-COTE [32].

We developed our own TSC framework for three main reasons. First, many TSC methods (e.g., most distance- and model-based methods) do not meet our goal of explainable AI. Second, the most widely used TSC dataset, the UCR Time Series Classification Archive [33], does not cover an emotion time series that is comparable with our emotion dataset, making the existing algorithms' performance not directly analogous to our emotion TSC task. Third, many methods focus on

identifying shapes, or measuring global distances between time series, and they disregard the timing and frequency of patterns, which make them unlikely to apply to our setting.

3. Methodology

3.1. Research context and data

Our data come from a top 10 insurance company in China. Each day, the company's call centers receive over 10,000 service calls. The company routinely samples 2 percent of all service calls in order to follow up with a customer survey. The company's representatives call the chosen customers several days after the service and ask them to rate the service. Because of the company's requirements for customer privacy protection and data security, we randomly selected 1,200 historical calls related to property insurance (between July and August 2016) with follow-up surveys as our dataset. After inspection, 57 calls were removed from the dataset either because they reflected dialogue between two employees or because they were marred by severe background noise.

3.2. Labeling

This paper labeled service quality, the target variable, using the customer survey data. The company uses a modified instrument with a rating scale of 1 to 5 (very poor to very good). SERVQUAL includes five service-quality dimensions: *tangibles*, *reliability*, *responsiveness*, *assurance*, and *empathy*. We mapped the company's survey questions into four dimensions of SERVQUAL but not *tangibles* (Table 1)—because the latter does not apply to call centers [34].

Table 1
Service quality: Dimensions and survey questions.

Dimension	Definition	Survey question	Comparable items in the literature
Reliability	Ability to perform the promised service dependably and accurately.	• Has the question you consulted / the problem you encountered been addressed? * • Please rate the solution offered by the service employee.	• When you have a problem, XYZ shows a sincere interest in solving it [35]. • XYZ is dependable [36].
Responsiveness	Willingness to help customers and provide prompt service.	• Please rate the promptness of the service provided by the service employee. • Please rate the willingness of the service employee to help you.	• Employees of XYZ give you prompt service [35]. • Employees of XYZ are always willing to help you [35].
Assurance	Knowledge and courtesy of employees and their ability to convey trust and confidence.	• Please rate the attitude of the service employee. • Please rate the tone expressed by the service employee.	• Generally, the employees are courteous, polite, and respectful [37].
Empathy	Caring, individualized attention the company provides its customers.	• Please rate the degree of the service employee's patience in listening and talking to you. • Please rate the ability of the service employee to understand your needs.	• Attention and patience of the staff [38]. • Employees of XYZ do not know what your needs are [35].

Note:* For this question, an answer of "Yes" is assigned the label *high*; otherwise, *low*.

To obtain customers' emotion dynamics, this paper first manually coded customer emotions in each call. Specifically, the labelers were trained and experienced service employees. Every call was segmented into natural utterances (i.e., a continuous piece of speech beginning and ending with a clear pause) by machine. Then, the labelers listened to each utterance in the calls and assigned it an emotion tag. According to the annotation scheme, the labelers identify the speakers' role, understand the context, and recognize the customers' emotions based on both dialogue and utterance information. The given tag is *negative* or *non-negative*,[2] with the negative class encompassing

---

[2] Some research has labeled six emotions (anger, disgust, fear, happiness, sadness, and surprise) in studies of acted speech. This finer emotion recognition is not practical in the present context, for a few reasons. First, the distribution of emotions in the service context is highly imbalanced: some basic emotions (e.g., surprise, disgust, and fear) occur sparsely [40]. Moreover, in natural speech, people tend to use euphemistic and subtle emotional expressions, resulting in fewer frequent and less dramatic emotions than in acted speech. Consequently, most emotion studies based on natural speech can only detect coarse emotional states [40].

several basic negative emotions (anger, anxiety, fear, and sadness) and the non-negative class encompassing positive emotions (rare in this dataset) and neutral emotions. Each speech was tagged by three labelers, and the final emotion tags were decided by a majority vote.

Tagging all utterances yields a sequence $\{e_1, e_2, \ldots, e_n\}$ for each call, where e_i equals 1 for a negative emotion and 0 for a non-negative emotion. This paper followed a common practice of transforming the utterance-based time series into an equally spaced time series using a sliding window. Specifically, we applied a 20-second sliding window with a 10-second sliding interval and defined the new time series as $\{x_1, x_2, \ldots, x_N\}$, where x_i is the duration (in seconds) of negative emotion within the i-th window $[t_i, t_{i+1}]$, and N is the number of periods (windows) in the call. The overlap of adjacent windows was chosen to ensure the smoothness of the time series. Both representations of the customer emotion sequence were used in the analyses to extract features.

3.3. Extracting emotion dynamics features

This paper extracted interpretable characteristics of the emotion sequence as input features for the models. Such features are easy to understand by humans, so the model can remain simple while achieving high descriptive and predictive accuracy [39]. Following the literature, this paper organized the emotion dynamics features into two broad categories: *trajectory-based* and *episode-based*.

3.3.1. Trajectory-based features

Variability represents the range of fluctuations in emotions [23]. In our context, emotional variability reflects the amount of variation in a customer's emotions during a service encounter. Higher variability indicates a higher likelihood of genuine emotional fluctuations. Following the literature [12], the *emotional variability*, denoted as *Var*, can be calculated as

$$Var = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1} \tag{1}$$

where $\bar{x}$ denotes the average value of $\{x_1, x_2, \ldots, x_N\}$.

Table 2
Features of emotion dynamics.

Category	Variable	Mean	Std	Min	25%	50%	75%	Max
			Trajectory-based features					
Variability: amplitude of emotion fluctuations	Var	3.62	8.23	0	0	0.66	3.65	86.66
Instability: magnitude of consecutive emotional changes	$MSSD$	2.74	4.75	0	0	0.62	3.46	37.58
Inertia: degree of emotion carried over from one moment to another	ACF_2	-0.03	0.16	-0.76	-0.10	0	0	0.71
	ACF_3	-0.06	0.14	-0.56	-0.14	-0.02	0	0.55
	ACF_4	-0.06	0.13	-0.57	-0.14	-0.02	0	0.49
	ACF_5	-0.06	0.12	-0.55	-0.13	-0.02	0	0.43
	ACF_6	-0.05	0.11	-0.56	-0.11	-0.02	0	0.43
	ACF_7	-0.04	0.11	-0.55	-0.10	-0.01	0	0.54
	ACF_8	-0.04	0.11	-0.45	-0.08	0	0	0.45
	ACF_9	-0.03	0.10	-0.47	-0.08	0	0	0.52
	ACF_{10}	-0.03	0.10	-0.40	-0.08	0	0	0.50
	$PACF_1$	0.33	0.27	-0.05	0.00	0.43	0.54	0.94
	$PACF_2$	-0.31	0.27	-1.18	-0.51	-0.40	0	0.23
	$PACF_3$	0.19	0.38	-1.52	0.00	0.14	0.29	5.53
	$PACF_4$	-0.26	0.76	-10.92	-0.44	-0.26	0	11.22
	$PACF_5$	0.24	14.50	-281.62	0.00	0.01	0.27	375.21
	$PACF_6$	0.02	4.51	-19.30	-0.46	-0.07	0	130.18
	$PACF_7$	0.18	5.47	-75.81	0.00	0	0.27	140.91
	$PACF_8$	-0.53	15.86	-515.73	-0.48	0	0	102.08
	$PACF_9$	0.45	23.87	-182.95	-0.49	0	0	724.53
	$PACF_{10}$	-0.76	13.74	-246.59	-0.10	0	0.30	25.31
Trend: overall tendency of emotional changes	$Slope$	9.99	35.20	-1.99	-0.03	0	0.01	1.38
			Episode-based features					
Duration: elapsed time between the start and end of (an) emotion episode(s)	Dur_{Total}	2.76	4.71	0	0	3.19	10.01	931.95
	Dur_{Avg}	4.12	8.08	0	0	1.95	3.86	66.58
	Dur_{Max}	1.84	3.72	0	0	2.36	5.51	146.62
	Dur_{Min}	49.71	67.69	0	0	1.07	2.33	66.58
Location: absolute and relative location of emotion episodes	Loc_{First}	0.33	0.30	0	0	16.60	78.35	408.29
	Loc_{Last}	106.08	117.87	0	0	83.43	179.36	1529.44
	Loc_{RelAvg}	-0.02	0.16	0	0	0.35	0.56	0.97

Instability refers to the amplitude of emotional changes across adjacent periods [12] and includes both variability and temporal dependency [40]. Higher emotional instability reflects a more drastic change in emotional valence from one moment to the next. This paper adopted the most common instability measure, the mean square of successive differences (MSSD) [23], defined as

$$MSSD = \sqrt{\frac{\sum_{i=1}^{N-1}(x_i - x_{i+1})^2}{N-1}} \tag{2}$$

Inertia is the degree to which an emotion carries over from one moment to another, indicating

resistance to change [24]. Following psychological studies (e.g., [12]) and TSC research (e.g., [32]), this paper used the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to measure emotional inertia. Given a time series $\{x_1, x_2, \ldots, x_N\}$, the ACF of lag k is the correlation between x_t and x_{t+k}. The PACF of lag k is the conditional correlation between x_t and x_{t+k} with the linear dependence of x_t on x_{t+1} through x_{t+k-1} removed. A positive value of ACF (PACF) indicates a positive correlation (partial correlation), or a stronger tendency for emotions to linger. [41] suggested that the lag should not exceed a quarter of a sequence's length. Applying this guideline, we calculated ACFs and PACFs with lags from 1 to 10 periods. Formally, the definition of ACF and PACF are

$$ACF_k = corr(x_t, x_{t+k}) = \sum_{i=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x}) \bigg/ \sum_{t=1}^{N}(x_t - x)^2 \tag{3}$$

$$PACF_k = corr(x_t - P_{t,k}(x_t), x_{t+k} - P_{t,k}(x_{t+k})) \tag{4}$$

where $P_{t,k}(x)$ is a surjective operator of an orthogonal projection of x onto the linear subspace of the Hilbert space spanned by $x_{t+1}, x_{t+2}, \ldots, x_{t+k-1}$.

This paper follows the literature to measure trends using slope (e.g., [25,28]), defined as *slope* = $(x_N - x_1) / (N - 1)$, to capture the overall direction of the emotion trajectory without regard to local fluctuations. In our context, a rising trend indicates that a customer's emotion is increasingly negative.

3.3.2. Episode-based features

Duration refers to the length of an emotion episode [22]. A longer duration of negative emotion may indicate a longer exposure to undesirable services. We computed the total, average, longest, and shortest duration of all episodes in a service call (Dur_{Total}, Dur_{Avg}, Dur_{Max}, and Dur_{Min}).

Location of an emotion episode has never been used as a predictive feature but matters in the present context. For example, a negative emotion episode at the beginning of a call may be the result of a previous service problem, whereas one at the end may indicate poor service and unresolved problems during the focal call. We computed the absolute location as the time elapsed since the

beginning of the call and the relative location as the quantile of the emotion episode. We calculated the absolute and relative locations of the first and last emotion episodes (Loc_{First}, Loc_{RelFirst}, Loc_{Last}, and Loc_{RelLast}) and the average absolute and relative locations of all episodes (Loc_{Avg} and Loc_{RelAvg}).

In total, this study has 33 emotion dynamics features: 23 *trajectory-based* and 10 *episode-based*. After checking collinearity and removing highly redundant features, 29 features remain (See Table 2).[3]

3.4. Prediction method

After obtaining emotion dynamics features as the predictors, we chose the approach to predicting service quality. To effectively combine the heterogeneous features, we built an ensemble of component models, one for each feature set [42]. In addition, some of the features (e.g., Dur_{Total}, Dur_{Avg}) have severely skewed distributions due to the sparsity of negative emotions (as shown in Table 2), and there are inevitable outliers in the dataset. The objective nature of the dataset implies that the outliers cannot be simply treated as data errors and discarded, because these extreme values may be strongly related to service quality. Ideally, this paper wanted to adopt a predictive modeling approach that is robust to skewness and noise. Meanwhile, we hoped to obtain both good predictive performance and interpretable findings; thus, this paper chose a *post hoc* interpretation approach to explain complex underlying relationships while obtaining higher predictive accuracy [43].

This paper first constructed separate component models for trajectory- and episode-based features and then created an ensemble of the two using the stacking method for ensemble learning [44], which takes all the predictions of component models as the input of a combiner/classifier to make a final prediction. A decision tree served as a combiner to automatically determine the contribution weight of each input. For each component model, we chose XGBoost, which is widely

---

[3] Specifically, for each pair of features whose correlation coefficient is larger than 0.9, we kept the more predictive one. We determined predictiveness using four metrics: mutual information, ANOVA, and the prediction performances (F1-score and AUC) of a univariate decision tree. We ranked the features by each metric and used the average rank to determine which features to keep.

used and is known to handle skewness and noise well and permits the computation of feature importance, and set the number of trees to 50 after experimentation. 70% of the sample is used for training, while the remaining 30% is used for testing.

3.5. Explanation method

Our explainability design has two parts: (1) obtaining the predictive ability of each emotion dynamics feature, and (2) extracting the relationship between each feature and service quality. For the first part, we adopted permutation feature importance [45]. When we permute that feature while keeping others unchanged, the more important a feature is, the more likely the prediction error will change. For the second part, we used the accumulated local effects (ALE) method [46] to describe and visualize the effect of features on prediction targets. We first divided a feature into several windows and then measured how the model predictions change when replacing the feature value with boundary values for data instances in that window, and then accumulated the average effects across all windows. Compared with similar methods, such as the partial dependence plot, ALE is more trustworthy, effective, and unbiased when predictors are correlated [39] and ALE plots can capture different types of linear and nonlinear relationships.

4. Prediction performance

This section reports the predictive performance of our approach (using customer emotion to predict service quality) against the benchmarks. Because the company's priority is to find low-quality service calls, the ideal performance metric should focus on the low-quality class. Following literature predicting customer satisfaction [47,48], we use the F1-score – the harmonic mean of precision and recall – of the low-quality class as our performance metric.

Based on common practices for gauging service quality, this study constructed two models as benchmarks: a quality-inspection-based model and an operational-indices-based model. In our context, the quality inspection covers multiple dimensions, including attitude, behavior, and

expertise, with a total of 22 items.[4] The operational-indices-based benchmark model includes basic

information on customers (e.g., gender, city, and inbound call history) and call details (e.g., call

duration, silence duration, and number of turns). There are 12 items in this feature set (see Appendix

A for details). We tried several commonly used classifiers (XGBoost, logistic regression, decision

tree, neural network, and Adaboost). As seen in Table 3, we obtained the best result using XGBoost

and our approach obtained a higher F1-score (of 0.529) than the two benchmark models using the

quality-inspection features and operational indices (their best F1-scores are 0.261 and 0.458,

respectively). While our model's F1-score is not particularly high,[5] it is a significant improvement over

the two benchmark approaches. In other words, if the company uses our approach instead of using quality

inspection results or calls' operational indices to predict service quality, they would obtain a notably better

performance. This confirms the benefit of using customer emotion to predict service quality,

answering our first research question.

Table 3
Prediction performance: Comparison with benchmarks (F1-score for negative class).

	Using	XGBoost	Logistic	Decision	Neural	Adaboost

[4] Quality-inspection items include the following: service scripts; language expression; familiar with the process; proficient in business knowledge; work order record information is accurate; work order flow; work order content is smooth and typo-free; solution provision; protect customer privacy; service attitude (friendly, active, patient); tone (friendly); speed of speech matching; tone (smooth, not stiff); accurate understanding of customer's needs; patience; express actively not passively; tactful rejection; try different solutions until the customer is satisfied; agent quality-inspection result; process quality-inspection result; overall quality-inspection result.

[5] There are a few reasons the F1-score is not very high. Firstly, our sample dataset exhibits an imbalance. with 38% of the calls categorized as low-quality. Previous studies using imbalanced phone call data have routinely reported less than 0.5 F1-scores (Cong et al, 2016, Luque et al 2017). Despite this imbalance, our model achieves an F1-score of 0.529, which is competitive relative to the literature and significantly better than random predictions which have an F1-score of 0.216. Secondly, we only employed the customer emotion dynamic features in our model because of our focus on illustrating the relationship between customer emotion and service quality. We anticipate an improved F-1 score if we are to incorporate other non-emotion features. Lastly, we are limited to a small dataset due to the sensitivity of the insurance data. We anticipate enhanced performance when applying our approach to a larger dataset.

			regression	tree	network	
Our method	Customer emotion dynamics	**0.529**	0.499	0.497	0.410	0.468
Benchmark 1	Quality inspection	0.201	0.261	0.212	0.205	0.215
Benchmark 2	Operational indices	0.458	0.417	0.426	0.259	0.423

We further evaluated the performance of our method for emotion-sequence classification by comparing it with a host of widely used, high-performance TSC methods, including DTW-KNN, EE, TSF, BOSS, random interval spectral ensemble (RISE) [32], Shapelets, and HIVE-COTE. This paper excluded deep learning models because our sample size is too small to meet the requirements of these models, and, in any case, studies have shown that they do not outperform the aforementioned models (e.g., [31]). We implemented the aforementioned models using sktime, a Python package with a comprehensive collection of advanced TSC algorithms proposed by recent research [49]. Table 4 describes the features and classifiers of each model. Table 4 shows that our approach obtained a higher F1-score than the other TSC methods (Rows (2) through (8) in Table 4). We thus conclude that the TSC method we develop for customer emotion classification is superior to the state-of-the-art methods as documented in the literature. It is also worth noting that most benchmark methods are not explainable and only focus on a subset of features (See Section 2.3 for details).

Table 4
Prediction performance: Comparison with other TSC methods (F1-score for negative class).

	Method	Features	Classifier	F1-score
(1)	Our method	Emotion dynamics features listed in Table 2	XGBoost	**0.529**
(2)	DTW-KNN	DTW distance	KNN	0.336
(3)	EE	Eight distance metrics	KNN	0.314
(4)	TSF	Standard deviation, average, and slope of random intervals	Ensemble of decision trees	0.442
(5)	BOSS	Word and Fourier transformation	KNN	0.446
(6)	RISE	*ACF*, *PACF*, and powerspectrum of random intervals	Ensemble of decision trees	**0.448**
(7)	Shapelets	Discriminative subseries	Ensemble of 50 decision trees	0.176
(8)	HIVE-COTE		Weighted ensemble of (2), (3), (4), (5), (6), and (7)	0.346

Finally, this paper drilled down on emotion's effect on specific dimensions of service quality. In practice, companies require such knowledge to better diagnose which dimension of service needs

improvement and to provide recovery strategies. We constructed models to predict each of the four

dimensions of service quality (reliability, responsiveness, assurance, and empathy) based on emotion

dynamics features. Table 5 shows that our method outperforms the benchmarks based on quality

inspection and operational indices in three service quality dimensions, i.e., responsiveness, assurance,

and empathy. The quality-inspection-based model (benchmark 1) obtain a higher F1-score than our

model on the reliability dimension. One explanation is that most quality assurance metrics are closely

related to service reliability, giving the quality-inspection-based model an advantage in predicting

consumers' reliability ratings.

Table 5
Prediction results of four dimensions of service quality (F1-score for negative class).

	Using	Reliability	Responsiveness	Assurance	Empathy
Our method	Customer emotion dynamics	0.187	**0.466**	**0.308**	**0.352**
Benchmark 1	Quality inspection	**0.269**	0.227	0.178	0.159
Benchmark 2	Operational indices	0.143	0.332	0.130	0.319

5. Explanation

From the prediction model, we proceed to the explanation element of our method by

investigating the relationship between specific features of emotion dynamics and service quality.

We first examined feature importance. Table 6 shows the permutation feature importance score

and rank of each feature. We observe that the trajectory-based features, which capture global

sequence patterns, have greater predictive power than the local episode-based features.

Table 6
Feature importance.

Category	Measure	Importance	Rank	Category	Measure	Importance	Rank
Variability	Var	0.037	8	Inertia	$PACF_5$	0.015	23
Instability	$MSSD$	0.097	3	(cont'd)	$PACF_6$	0.023	15
Inertia	ACF_2	0.024	13		$PACF_7$	0.036	9
	ACF_3	0.01	27		$PACF_8$	0.016	21
	ACF_4	0.014	24		$PACF_9$	0.04	7
	ACF_5	0.072	5		$PACF_{10}$	0.134	1
	ACF_6	0.006	28	Trend	$Slope$	0.026	10
	ACF_7	0.023	14	Duration	Dur_{Total}	0.019	17
	ACF_8	0.054	6		Dur_{Avg}	0.019	18
	ACF_9	0.016	19		Dur_{Max}	0.015	22
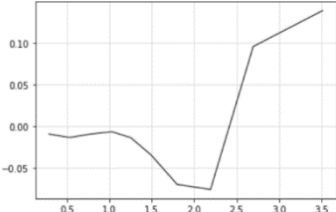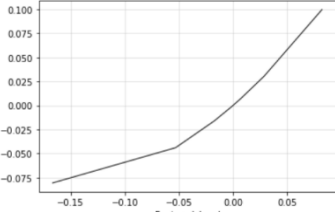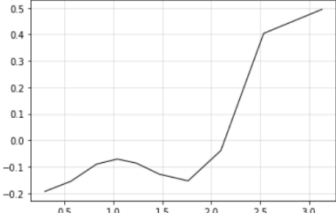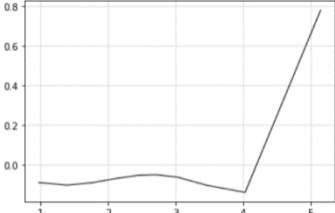
|						
ACF_{10}	0.111	2		Dur_{Min}	0.016	20
$PACF_1$	0.081	4	Location	Loc_{First}	0.013	25
$PACF_2$	0.011	26		Loc_{Last}	0.024	12
$PACF_3$	0.004	29		Loc_{RelAvg}	0.019	16
$PACF_4$	0.026	11				

Of the trajectory-based features (including the categories of variability, instability, inertia, and trend), $PACF_{10}$ and ACF_{10} have the highest predictive ability. Long-lag (6–10) inertia is more predictive than short-lag (1–5) inertia, with average feature importance of 0.046 and 0.022, respectively. The result indicates that whether a consumer's negative emotion is likely to persist for a long time is an important signal of service quality. Among the remaining features, instability ($MSSD$) is more important than variability (Var) and trend ($Slope$).

Of the episode-based features (including duration and location), the top feature is the location of the last emotion episode (Loc_{Last}). The average importance of location-related features (0.0187) is larger than that of the duration-based features (0.0173).

To understand how the change of emotion impacts service quality, we visualized the marginal effect of features using ALE plots (Table 7).

Table 7
Accumulated Local Effects (ALE) of important features.

Category	ALE Plot	Category	ALE Plot
Variability (a) Feature: Var (feature importance = 0.037)		Trend (d) Feature: $Slope$ (feature importance = 0.026)	
Instability (b) Feature: $MSSD$ (feature importance = 0.097)		Duration (e) Feature: Dur_{Total} (feature importance = 0.019)	

Inertia (c) Feature: $PACF_{10}$ (feature importance = 0.081)		Location (f) Feature: Loc_{Last} (feature importance = 0.024)	
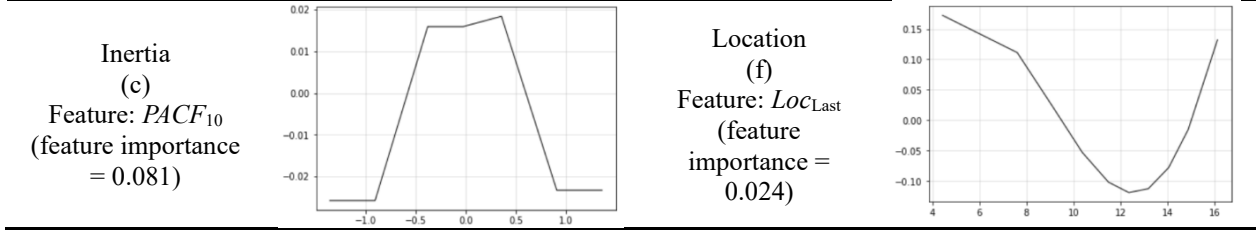
Notes: The x-axis of the ALE plots is the scaled value of the features. For better visualization, we transform skewed distributions with a square-root function or quantile transformation to make them closer to normal distribution. Then we keep the value between the 5% to 95% quantiles and use polynomial functions to fit the curve. The y-axis is the ALE value, representing the centered probability of low service quality. The larger the ALE value, the more likely the service encounter is of low quality. A zero ALE value represents the mean effect of the feature on service quality.

Because of space limitations, this paper presents ALE plots for the most important feature in each category. We observe the following from the ALE plots.

First, when the customer's negative emotions in the service encounter have a high fluctuation (more variable and unstable), positive trend, or long duration, the perceived quality is likely to be low. The ALE plots of variability and instability (Table 7 (a) (b)) show that a large fluctuation in negative emotions during service calls indicates low service quality. For the trend feature, the ALE curve (Table 7 (d)) is almost linear, indicating the covariate relationship between customers' negative emotions and low service quality. Moreover, the longer the duration of emotion episodes (Dur_{Total}, Dur_{Max}, Dur_{Min}, and Dur_{Avg}), the more likely the encounter is of low quality. Table 7 (e) shows that when the total duration of negative emotions is larger than a threshold (16 seconds), it is positively correlated with the likelihood of low quality.

Second, when a customer's emotion status has a lower intention to linger, the service quality is more likely to be low. The ALE plot of the most important inertia feature ($PACF_{10}$, Table 7 (c)) shows that a small (large) absolute value indicates low (high) quality. In other words, when the partial correlation between emotion at t and $t+10$ is small, the service quality is likely to be low.

Third, the location where the last emotion episode appears in a call has a U-shaped relationship with low service quality. The plot in Table 7 (f) indicates that if the last negative emotion episode appears more than 170 seconds after the service begins, the service is more likely to be low-quality.

After a long service procedure, it is certainly the employee's responsibility that he or she failed to soothe the customers' feelings or even aroused anger. If the last emotion episode occurs shortly after the call begins, the service encounter might also be of low quality. [6]

6. Application

This section instantiates the application of our method by developing an IT artifact that automatically measures service quality based on automatic emotion-recognition techniques. We also apply the measure for a specific business purpose, which, as suggested by our research site (the insurance company), is to identify customers who have referral intentions.

Customers' referral intentions are the best predictor of revenue growth because consumers tend to trust recommendations from friends [50]. It is crucial for companies to carefully manage referral programs [51]. In the present context, the insurance company conducts post-service surveys, directly asking sampled customers whether they would recommend the company. Again, the surveys are time-consuming and labor-intensive and may disturb the customer. In theory, customers' referral intention covaries with the quality of the service they have received [37], so we developed an IT artifact using the service quality measure to predict referral intention by the following three steps.

In Step 1, the artifact automatically detects customer emotions from service calls. As described in Section 3, in developing our method, we manually labeled customers' emotions. To be fully automated, the IT artifact must be able to automatically tag a customer's emotion during a service call. Based on a review of state-of-the-art automatic emotion recognition techniques, we designed a system for real-time emotion recognition. The system leveraged three AI techniques (CNN, RNN, and SVM) and two types of features (linguistic and acoustic features) to build six independent

---

[6] Examining the data, we find two primary explanations for this situation: (a) the encounter is short and the expected service is not performed (e.g., the service employee announces that he/she cannot offer any help), leading to low service quality, or (b) the customer had high expectations before the service but soon received only perfunctory service, which does not lead to the expression of additional negative emotions but is not satisfying.

models. The acoustic features include Mel frequency cepstrum coefficient (MFCC) and spectrogram, and the linguistic features include n-gram and word2vec. We then integrated the six models using the ensemble method. The ensemble model achieved an accuracy of 87%, an average recall of 65.05%, and an average F-score of 0.684, which outperforms most models based on natural speech (see model details and evaluation results in Appendix B). It provided an emotion tag for each speech utterance. And all the identified emotion tags of the speech utterances in one service call made up the customer's emotion sequence.

In Step 2, we used the emotion sequence generated from the above process to predict the service quality of a phone call. We retrained the method developed in Section 3 using automatically tagged emotions and applied it. The company provided a set of 2,140 service calls that include complete records of quality inspections and post-service customer surveys. We divided the set into two parts, with 1,500 calls for training and the rest for prediction. Our method of using customer emotion (automatically tagged this time) to predict service quality outperformed the benchmark model based on operational indices. These operational indices can be obtained automatically, whereas the quality inspection is executed manually and therefore does not serve as a reasonable benchmark for an automatic IT artifact.

In Step 3, we used the automatic measure of service quality to predict referral intention. The company provided another set of 1,500 service calls, in which 40 percent of customers indicated, in the post-service survey, willingness to recommend the company to friends. For each of the phone calls, we automatically generated service-quality measures based on the above two steps. Then, we used 70 percent of the data to train a referral prediction model and reserved the rest for testing. To be specific, we used the probabilities of four dimensions of service quality as features and XGBoost as the classifier. We have two benchmarks in this analysis. The first uses a set of operational indices of a service call to predict referral intention. For the second, we trained a model to predict referral intention by using actual service quality (in the four dimensions) collected via customer surveys. This

provides the best possible prediction result based on service quality, in that customer surveys provide the "ground truth" of service quality. Table 8 presents the prediction model's performance (recall and precision) on the recommendation probability. The automatic IT artifact significantly outperforms the benchmark of using operational indices. The performance of the automatic IT artifact is slightly inferior to the benchmark of using actual service quality. It is deemed acceptable by the insurance company because the automatic IT artifact can scan all service encounters, far more than the original level of 2 percent, potentially enlarging the pool of sales prospects.

Table 8
Referral intention prediction.

	Predict referral intention using	Recall	Precision
(1)	Operational indices of the service call	52.74%	51.73%
(2)	Actual service quality (ground truth, collected via customer survey)	91.56%	58.33%
(3)	Service-quality measure based on customer emotion (our method)	86.50%	54.25%

7. Discussion

7.1. Contributions

This study makes four contributions. First, we develop an effective method for measuring service quality, a core variable in service science, based on customer emotion. Our method leverages AI techniques (e.g., TSC, automatic emotion recognition, and explainable AI methods) and is automatic and scalable. Using real-world data, we verify that it is superior to manual quality inspection and operational indices for identifying low service quality. Compared with widely used approaches such as customer surveys and quality inspection, our method is automatic, cost-effective, real-time, and better positioned to generate a large-scale or longitudinal sample of service quality. This enables better theory testing with greater statistical power and the development of theory interested in temporal changes for domains such as e-service, e-government, and e-commerce. Using real-world data, we show that the measure is useful for predicting customers' referral intentions, which is a further testament to the measure's nomological validity.

Second, our findings based on explainable AI contribute to the service literature. The

interpretation of our model highlights the importance of features of customer emotion dynamics and shows which dynamic characteristics of consumer emotions may correlate with low service quality and these characteristics' particular effects. Prior research has mainly examined the customer's emotion at a single-point state, usually at the end of the service encounter [20]. Our approach sheds light on the relationships between characteristics of customer emotional dynamics during the service encounter and perceived service quality. The results show that when negative emotions are highly fluctuating, showing a positive trend, long-lasting, or lingering, perceived service quality is likely to be low. The results also show that the location where the last emotion episode appears in a service call has a U-shaped relationship to low quality. These findings deepen our understanding of the relationship between customer emotion and service quality.

Third, on the technical side, we propose a classification framework for using customer emotion dynamics. Although the literature provides a rich pool of TSC methods, it lacks research on which methods are suitable for emotion sequence analysis. Unlike existing methods, we propose a TSC framework that is explainable, sensitive to the time location of emotional events, and combines trajectory- and episode-based features. We use unstructured multi-media data containing extremely rich information that aids in decision-making. Our approach outperforms the state-of-the-art TSC methods in our research context. This approach for emotion-sequence analysis is also promising for other contexts, such as recommendations, online reviews, and online education.

Fourth, we demonstrate how to integrate emotion recognition technologies and emotion sequence analysis into decision support systems. Our prototype system enables real-time prediction of service quality using information contained in customer emotions. While previous studies have extensively studied emotion recognition techniques, they seldom study how they can be used in real-world applications. We provide one of the first evidence of their value in predicting service quality and referral intention.

Finally, we showcase how to harness the potential of customer speech data for generating

valuable business insights, thereby making a significant contribution to the decision support systems literature. Customer speech, particularly in call centers, represents a valuable yet underutilized data source. Our system offers a comprehensive integration of real-time voice input analysis, encompassing noise reduction, voice activity detection, voice feature extraction, and classification. By leveraging this system, we can gain valuable insights that can inform the design and implementation of future speech-based decision support systems.

7.2. Managerial implications

Our study offers significant practical value to managers and customer service professionals. First, we designed an effective approach that is highly scalable, automated, and of low cost, addressing a crucial bottleneck in service-quality management. The entire service-quality prediction process using real-time speech data can be completed in seconds, which opens up opportunities for in-time interventions. Companies can also integrate the service-quality prediction model into other applications, such as referral reward programs, to enhance business performance.

Our explainable AI approach has the key advantage of allowing practitioners to gain insights into why a particular call receives a particular service-quality score. For example, managers can use our ALE plots in coaching service employees and pinpoint whether a low-quality call suffers from large instability of negative emotion. Similarly, our analysis of feature importance can help managers prioritize their monitoring and intervention efforts. Overall, our system enables a better understanding of customer emotions and more prompt managerial support.

7.3. Limitations and future research

This work is subject to several limitations that should be addressed in future research. First, our analysis is based on a limited-size dataset. We expect that with increased data size, one can obtain higher predictive performance (Phillips-Wren et al. 2021). Second, our findings on how emotional dynamics are related to service quality are data-driven and exploratory. Future research should complement these findings by further validating such relationships, perhaps using experiments and

surveys. Third, because we lack information about a customer's service history, we cannot personalize the prediction by leveraging past information. With repeat customers, this could be a useful direction for future work. With more customer information, prediction performance can be improved. Finally, this study focuses on voice data only. Future research could combine our method with other data sources (e.g., video and texts) to further improve predictive performance.

Appendix A. Operational indices

Table A.1

Operational indices.

	Item	Reference
Customer information	• Customer gender • Customer city classification (1–5 classes) • Number of historical inbound calls before this encounter	[52]
Call information	• Call time (morning, afternoon, evening)	[53]
	• Silence duration	[54]
	• Number of turns taken by an agent • Number of turns taken by a customer	[55]
	• Total utterance duration of customers • Average utterance duration of customers • Average utterance duration of agent • Duration of dialogs	[55]

Appendix B. Automatic customer emotion recognition system

The system for real-time automatic customer emotion recognition consists of three modules: preprocessing, feature extraction, and emotion recognition.

Preprocessing

Noise reduction. The input audio stream may be mixed with background noises such as wind, TV, vehicles, and other people. We apply a classic method, the adaptive Weiner filter, to reduce the noise and increase the signal-to-noise ratio [56].

Voice activity detection and speaker recognition. After de-noising, we divide speech into segments, each of which is an utterance or complete sentence spoken by one person, using a popular algorithm proposed by [57]. Segments that are not suitable for customer emotion recognition (e.g.,

overlapping utterances, without human speech) are excluded from further analyses. We then classify the speech segments by speakers using the Gaussian Mixture Model with Bayesian Information Criterion [58]. Only customers' speech segments are used in the following analyses.

Speech-to-text conversion. We convert the speech into text using the automatic speech recognition service of Iflytek, a leading vendor for Mandarin speech recognition.

Text preprocessing. First, we perform tokenization and expand abbreviations to their full forms. We remove words that provide little information about emotion (e.g., special entities' names) and retain stop words because they may contain important emotional information (e.g., "not" in "A delay of two weeks does not please me").

Feature extraction

After preprocessing, the raw data has been converted into audio segments and the corresponding texts. We extract two kinds of acoustic features from the audio: Mel frequency cepstrum coefficient (MFCC) and spectrogram. MFCC is one of the most well-known acoustic features, widely used in speech analysis [59]. Spectrogram is a visual representation of audio signals in the time-frequency domain and is usually used as input into deep learning models [60]. Each speech segment is transformed into a spectrogram using the Short-Time Fourier Transform algorithm. The size of the spectrogram is $129 \times n$, where 129 is the dimension of the frequency domain and n is the number of frames in the segment.

The linguistic features are extracted from speech texts, including n-gram and word2vec. N-gram refers to a contiguous sequence of n words in the given sentence [61]. N-gram has the merit of simplicity and has been applied in text categorization, machine translation, and emotion recognition. We keep unigrams (n = 1) and bigrams (n = 2) because spoken dialogues are usually short, incomplete, and disjointed. We select 70% of the most informative n-gram features using Chi-square tests.

Word2vec is a widely used word embedding that maps all the words and phrases to vectors of real numbers (Phan et al. 2023). In this study, the word vector's dimension is 50, and the context window size is 8. Words with a frequency of less than 5 are removed. Finally, we obtain $50 \times n$ word vectors, where n is the number of words in the segment.

Emotion recognition

The emotion-recognition module classifies customers' speech segments into negative and non-negative emotions using extracted acoustic and linguistic features as inputs. After exploring a host of classifier-feature combinations,[7] the current study applies three classifiers with better performance—CNN, RNN, and SVM—on different acoustic/linguistic inputs and obtains six classification models. We then integrate the six models using the ensemble method (as shown in Figure B.1).
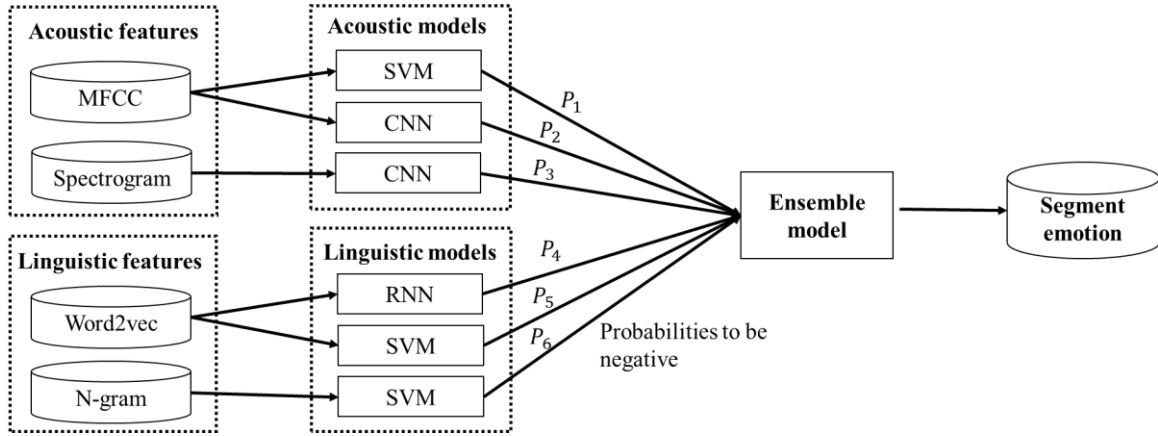

Fig. B.1. Speech emotion recognition models.

We construct two CNN models, one using MFCC features (called CNN-MFCC hereafter) and the other using spectrograms (CNN-Spectrogram). CNN has advantages in extracting high-level abstract features and reducing dimensionality, which is also important in classifying high-dimension audio data. To model the temporal dynamic linguistic feature, we run an RNN model. Specifically,

---

[7] We also tried the HMM, Deep Believe Network, and Variational Autoencoder models on acoustic inputs and the Decision Tree, Random Forest, and Multinomial Naive Bayes models on linguistic inputs.

we choose Gated Recurrent Unit (GRU) for its low computational costs.

Support Vector Machine (SVM) is a supervised machine learning algorithm, especially suitable for nonlinear classification problems. Compared with neural networks, SVM shows superiority in small datasets. We create three SVM models—SVM-MFCC, SVM-Ngram, and SVM-Word2vec—for MFCC (acoustic), n-gram (linguistic), and word2vec (linguistic) features, respectively.

We then create an ensemble of these different models. Specifically, we assign each classifier a weight, W_k, to represent its importance and use the weighted average ensemble method. The ensemble output y_{ij} is given by $y_{ij} = \sum_{k=1}^{K} W_k \cdot P_{ijk}$, where W_k ranges from 0 to 1, with a constraint $\sum_{k=1}^{K} W_k = 1$ and P_{ijk} is the predicted probability of the negative emotion class by classifier k for the j-th segment of the i-th call. We train three ensemble models: the linguistic ensemble model (including RNN-Word2vec, SVM-Ngram, and SVM-Word2vec), the acoustic ensemble model (including CNN-MFCC, CNN-Spectrogram, and SVM-MFCC), and the acoustic-linguistic ensemble model (which integrates all six models).

Evaluation

In management practice, we need to identify customers' negative emotions as precisely as possible. We use three evaluation metrics (precision, recall, and F-score) in the *negative* class and set β in the F-score at 0.5 to assign more weight to precision. Table B.1 reports the performance of the six independent models and three ensemble models. The results show that ensemble models perform better overall. And integrating both acoustic and linguistic information of speech is more beneficial than using uni-channel information.

The results show that all models perform well beyond the chance level. All three ensemble models achieve higher precision than the independent models, and the acoustic-linguistic-ensemble model achieves the best precision of the nine models. The acoustic-linguistic-ensemble model has an average recall of 65.05% and an average F-score of 0.684, higher than the literature using real-world

speech data. And this model achieves an accuracy of 87%, which is among the best models based on

natural speech in the literature.

Table B.1

Performance of individual and ensemble models.

Feature type	Models	Recall	Precision	F-score
Acoustic	(1) CNN-MFCC	21.12%	40.96%	0.345
	(2) CNN-Spectrogram	63.98%	44.02%	0.470
	(3) SVM-MFCC	38.82%	38.82%	0.388
	(4) Acoustic Ensemble: 1+2+3	36.96%	57.77%	0.519
Linguistic	(5) SVM-Ngram	11.49%	55.22%	0.314
	(6) SVM-Word2vec	35.40%	40.43%	0.393
	(7) RNN-Word2vec	34.16%	48.46%	0.447
	(8) Linguistic Ensemble: 5+6+7	24.53%	56.43%	0.448
Acoustic + Linguistic	(9) Acoustic-linguistic Ensemble: 1+2+3+5+6+7	33.23%	66.05%	0.552

Notes: In order to maintain stable predictive performance under realistic conditions, we construct the test data set with the ratio of positive: negative =5.5:1, so the chance level of precision is 15.38%.

References

[1]   R. Ladhari, A review of twenty years of SERVQUAL research, Int. J. Qual. Serv. Sci. 1 (2009) 172–198. https://doi.org/10.1108/17566690910971445.

[2]   M. Blut, N. Chowdhry, V. Mittal, C. Brock, E-service quality: A meta-analytic review, J. Retail. 91 (2015) 679–700.

[3]   R. Nishant, S.C. Srivastava, T.S. Teo, Using polynomial modeling to understand service quality in e-government websites, MIS Q. 43 (2019) 807–826.

[4]   M. Saberi, O. Khadeer Hussain, E. Chang, Past, present and future of contact centers: a literature review, Bus. Process Manag. J. 23 (2017) 574–597. https://doi.org/10.1108/BPMJ-02-2015-0018.

[5]   Zeithaml, Berry, Parasuraman, Communication and Control Processes in the Delivery of Service Quality, 52 (1988) 35–48.

[6]   N. Gans, G. Koole, A. Mandelbaum, Telephone Call Centers: Tutorial, Review, and Research Prospects, Manuf. Serv. Oper. Manag. 5 (2003) 79–141. https://doi.org/10.1287/msom.5.2.79.16071.

[7]   N. Berente, B. Gu, J. Recker, R. Santhanam, Managing AI, Call Pap. MIS Q. (2019).

[8]   E. Brynjolfsson, C. Wang, X. Zhang, The economics of IT and digitization: Eight questions for research, MIS Q. 45 (2021) 473–477.

[9]   A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access. 6 (2018) 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

[10]   C.J. White, The impact of emotions on service quality, satisfaction, and positive word-of-mouth intentions over time, J. Mark. Manag. 26 (2010) 381–394. https://doi.org/10.1080/02672571003633610.

[11]   I. Guyon, A. Elisseeff, An introduction to feature extraction, in: Featur. Extr., Springer, Berlin, Heidelberg, 2006: pp. 1–25.

[12]   M. Houben, W. Van Den Noortgate, P. Kuppens, The Relation Between Short-Term Emotion Dynamics and Psychological Well-Being: A Meta-Analysis, Psychol. Bull. 141 (2015) 901–930.

[13]   A.K. Jaiswal, Customer satisfaction and service quality measurement in Indian call centres, Manag. Serv. Qual. 18 (2008) 405–416. https://doi.org/10.1108/09604520810885635.

[14]   A. Miciak, M. Desmarais, Benchmarking service quality performance at business-to-business and business-to-consumer call centers, J. Bus. \& Ind. Mark. (2001).

[15]   W. Boulding, A. Kalra, R. Staelin, V.A. Zeithaml, A Dynamic Process Model of Service Quality: From Expectations to Behavioral Intentions, J. Mark. Res. 30 (1993) 7. https://doi.org/10.2307/3172510.

[16]   H. Liao, A. Chuang, A multilevel investigation of factors influencing employee service performance and customer outcomes, Acad. Manag. J. 47 (2004) 41–58. https://doi.org/10.2307/20159559.

[17]   A. Parasuraman, V.A. Zeithaml, L.L. Berry, A Conceptual Model of Service Quality and Its Implications for Future Research, J. Mark. 49 (1985) 41. https://doi.org/10.2307/1251430.

[18]   J. Walls, G. Widmeyer, O. El-Sawy, Building an Information System Design Theory for Vigilant EIS, Inf. Syst. Res. 3 (1992) 36–59. https://doi.org/10.1287/isre.3.1.36.

[19]   M.K. Brady, J.J. Cronin Jr, Some new thoughts on conceptualizing perceived service quality: A hierarchical approach, J. Mark. 65 (2001) 34–49. https://doi.org/10.1509/jmkg.65.3.34.18334.

[20]   A.S. Mattila, C.A. Enz, The Role of Emotions in Service Encounters, J. Serv. Res. (2002).

[21]   S. Gregor, A.C.H. Lin, T. Gedeon, A. Riaz, D. Zhu, Neuroscience and a nomological network for the understanding and assessment of emotions in information systems research, J. Manag. Inf. Syst. 30 (2014) 13–48.

[22]   P. Verduyn, P. Delaveau, J.Y. Rotgé, P. Fossati, I. Van Mechelen, Determinants of Emotion Duration and Underlying Psychological and Neural Mechanisms, Emot. Rev. 7 (2015) 330–335. https://doi.org/10.1177/1754073915590618.

[23]   E.M. Seabrook, M.L. Kern, B.D. Fulcher, N.S. Rickard, Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates, J. Med. Internet Res. 20 (2018). https://doi.org/10.2196/jmir.9267.

[24]   P. Kuppens, N.B. Allen, L.B. Sheeber, Emotional inertia and psychological maladjustment, Psychol. Sci. 21 (2010) 984–991. https://doi.org/10.1177/0956797610372634.

[25]   A. Pascual-Leone, How clients "change emotion with emotion": A programme of research on emotional processing, Psychother. Res. 28 (2018) 165–182. https://doi.org/10.1080/10503307.2017.1349350.

[26]   P. Kuppens, P. Verduyn, Looking at Emotion Regulation Through the Window of Emotion Dynamics, Psychol. Inq. 26 (2015) 72–79. https://doi.org/10.1080/1047840X.2015.960505.

[27]   J. Lines, A. Bagnall, Time series classification with ensembles of elastic distance measures, Data Min. Knowl. Discov. 29 (2015) 565–592. https://doi.org/10.1007/s10618-014-0361-2.

[28]   H. Deng, G. Runger, E. Tuv, M. Vladimir, A time series forest for classification and feature extraction, Inf. Sci. (Ny). 239 (2013) 142–153. https://doi.org/10.1016/j.ins.2013.02.030.

[29]   L. Ye, E. Keogh, Time series shapelets: a new primitive for data mining, in: Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2009: pp. 947–956.

[30]   P. Schäfer, The BOSS is concerned with time series classification in the presence of noise, Data Min. Knowl. Discov. 29 (2015) 1505–1530. https://doi.org/10.1007/s10618-014-0377-7.

[31]   H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. 33 (2019) 917–963. https://doi.org/10.1007/s10618-019-00619-1.

[32]   J. Lines, S. Taylor, A. Bagnall, Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles, ACM Trans. Knowl. Discov. Data. 12 (2018). https://doi.org/10.1145/3182382.

[33]   H.A. Dau, A. Bagnall, K. Kamgar, C.-C.M. Yeh, Y. Zhu, S. Gharghabi, C.A. Ratanamahatana, E. Keogh, The UCR time series archive, IEEE/CAA J. Autom. Sin. 6 (2019) 1293–1305.

[34]   J.J.P.-A. Hsieh, A. Rai, S.X. Xu, Extracting business value from IT: A sensemaking perspective of post-adoptive use, Manage. Sci. 57 (2011) 2018–2039. https://doi.org/10.1287/mnsc.1110.1398.

[35]   A. Parasuraman, L.L. Berry, V.A. Zeithaml, Refinement and Reassessment of the SERVQUAL Scale, J. Retail. 67 (1991) 420. https://doi.org/10.1111/j.1438-8677.2010.00335.x.

[36]   J.J. Cronin, S.A. Taylor, Measuring Service Quality: A Reexamination and Extension, J. Mark. 56 (1992) 55. https://doi.org/10.2307/1252296.

[37]   M.K. Brady, J.J. Cronin, R.R. Brand, Performance - Only Measurement of Service Quality: A Replication and Extension, J. Bus. Res. 55 (2002) 17–31. https://doi.org/10.1016/S0148-2963(00)00171-5.

[38]   Naik, Service Quality (Servqual) and its Effect on Customer Satisfaction in Retailing, Eur. J. Soc. Sci. (2010).

[39]   C. Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, (2020). https://christophm.github.io/interpretable-ml-book.

[40]   T.J. Trull, S.P. Lane, P. Koval, U.W. Ebner-Priemer, Affective Dynamics in Psychopathology, Emot. Rev. 7 (2015) 355–361. https://doi.org/10.1177/1754073915590617.

[41]   A. Bagnall, J. Lines, J. Hills, A. Bostrom, Time-Series Classification with COTE : The Collective of Transformation-Based Ensembles, IEEE Trans. Knowl. Data Eng. 27 (2015) 2522–2535. https://doi.org/10.1109/TKDE.2015.2416723.

[42]   R. Xia, C. Zong, S. Li, Ensemble of feature sets and classification algorithms for sentiment classification, Inf. Sci. (Ny). 181 (2011) 1138–1152. https://doi.org/10.1016/j.ins.2010.11.023.

[43]   W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Interpretable machine learning: definitions, methods, and applications, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 22071–22080. https://doi.org/10.1073/pnas.1900654116.

[44]   S. Džeroski, B. Ženko, Is combining classifiers with stacking better than selecting the best one?, Mach. Learn. 54 (2004) 255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e.

[45]   A. Fisher, C. Rudin, F. Dominici, All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance, ArXiv Prepr. ArXiv1801.01489. (2018) 237–246.

[46]   D.W. Apley, J. Zhu, Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, (2016). http://arxiv.org/abs/1612.08468.

[47]   A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, Y. Aono, T. Toda, Customer Satisfaction Estimation in Contact Center Calls Based on a Hierarchical Multi-Task Model, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2020) 715–728. https://doi.org/10.1109/TASLP.2020.2966857.

[48] J. Luque, C. Segura, A. Sanchez, M. Umbert, L.A. Galindo, The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls, Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH. 2017-Augus (2017) 2346–2350. https://doi.org/10.21437/Interspeech.2017-424.

[49] M. Löning, A. Bagnall, S. Ganesh, V. Kazakov, J. Lines, F.J. Király, sktime: A unified interface for machine learning with time series, ArXiv Prepr. ArXiv1909.07872. (2019).

[50] F.F. Reichheld, The one number you need to grow, Harv. Bus. Rev. 81 (2003) 46–55.

[51] J. Jung, R. Bapna, A. Gupta, S. Sen, Impact of incentive mechanism in online referral programs: evidence from randomized field experiments, J. Manag. Inf. Syst. 38 (2021) 59–81.

[52] C.-P. Wei, I.-T. Chiu, Turning telecommunications call details to churn prediction: a data mining approach, Expert Syst. Appl. 23 (2002) 103–112.

[53] J. Zhou, Data Mining for Individual Consumer Credit Default Prediction under E-commence Context : A Comparative Study, Icis. (2017) 1–18.

[54] G. Zweig, O. Siohan, G. Saon, B. Ramabhadran, D. Povey, L. Mangu, B. Kingsbury, Automated quality monitoring for call centers using speech and NLP technologies, (2006) 292–295. https://doi.org/10.3115/1225785.1225796.

[55] S. Stoyanchev, S. Maiti, S. Bangalore, Predicting interaction quality in customer service dialogs, in: Adv. Soc. Interact. with Agents, 2019: pp. 149–159. https://doi.org/10.1007/978-3-319-92108-2_16.

[56] M.A. Abd El-Fattah, M.I. Dessouky, A.M. Abbas, S.M. Diab, E.-S.M. El-Rabaie, W. Al-Nuaimy, S.A. Alshebeili, F.E. Abd El-samie, Speech enhancement with an adaptive Wiener filter, Int. J. Speech Technol. 17 (2014) 53–64.

[57] J.A. Haigh, J.S. Mason, Robust voice activity detection using cepstral features, in: Proc. TENCon'93. IEEE Reg. 10 Int. Conf. Comput. Commun. Autom., 1993: pp. 321–324.

[58] A. Mehrjou, R. Hosseini, B.N. Araabi, Improved Bayesian information criterion for mixture model selection, Pattern Recognit. Lett. 69 (2016) 22–27.

[59] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, H. Fan, Heart sound classification based on improved MFCC features and convolutional recurrent neural networks, Neural Networks. 130 (2020) 22–32.

[60] O. Abdel-hamid, L. Deng, D. Yu, Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition, 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH 2013). (2013) 3366–3370. https://doi.org/10.1.1.703.648.

[61] S. Banerjee, T. Pedersen, The design, implementation, and use of the ngram statistics package, in: Int. Conf. Intell. Text Process. Comput. Linguist., 2003: pp. 370–381.