Annals of Operations Research (2024) 333:939–970
https://doi.org/10.1007/s10479-022-04874-2

ORIGINAL RESEARCH

Data-driven subjective performance evaluation: An attentive deep neural networks model based on a call centre case

Abdelrahman Ahmed[1] · **Uthayasankar Sivarajah**[1] · **Zahir Irani**[1] · **Kamran Mahroof**[1] · **Vincent Charles**[1]

Accepted: 15 July 2022 / Published online: 26 October 2022

Abstract
Every contact centre engages in some form of Call Quality Monitoring in order to improve agent performance and customer satisfaction. Call centres have traditionally used a manual process to sort, select, and analyse a representative sample of interactions for evaluation purposes. Unfortunately, such a process is marked by subjectivity, which in turn results in a distorted picture of agent performance. To address the challenge of identifying and removing subjectivity, empirical research is required. In this paper, we introduce an evidence-based, machine learning-driven framework for the automatic detection of subjective calls. We analyse a corpus of seven hours of recorded calls from a real-estate call centre using Deep Neural Network (DNN) for a multi-classification problem. The study establishes the first baseline for subjectivity detection, with an accuracy of 75%, which is comparable to relevant speech studies in emotional recognition and performance classification. We conclude, among other things, that in order to achieve the best performance evaluation, subjective calls should be removed from the evaluation process or subjective scores deducted from the overall results.

Keywords Subjective evaluation · Agent Performance · Customer Behaviour · Deep neural network · Call Centre

✉ Abdelrahman Ahmed
A.ahmed110@bradford.ac.uk; Abdolrahman.salem@gmail.com

✉ Uthayasankar Sivarajah
U.Sivarajah@bradford.ac.uk

Zahir Irani
Z.Irani@bradford.ac.uk

Kamran Mahroof
K.Mahroof@bradford.ac.uk

Vincent Charles
C.Vincent3@bradford.ac.uk

[1] School of Management, University of Bradford, BD7 1DP Bradford, UK

1 Introduction

Effective and efficient call centre operations are key ingredients to success and a good reputation. Every call centre draws a baseline that measures overall performance according to the ultimate call centre objectives and strategies based on quantitative and qualitative methods (Abbott, 2004; Rubingh, 2013). The quantitative method considers the first call resolution, average handling time of the call, wrap-up time, and adherence time (Reynolds, 2010). The qualitative method is performed by the quality team, which listens to the recorded calls and evaluates the delivered services (Judkins et al., 2003). The quantitative and qualitative methods aim to fulfil the Key Performance Indicators (KPIs) widely used to measure agent performance and, in turn, overall call centre productivity (Reynolds, 2010).

However, objectively evaluating the call centre agents' performance is difficult. The quantitative method and relevant KPIs measure the efficiency of the agent rather than the quality of service, so customer satisfaction is not taken into account (Willis & Bendixen, 2007). On the other hand, the qualitative method is dependent on the prior experience of the quality team, which is based on finesse standards such as communication and language skills. As a result, the qualitative method allows for style and individuality while also leaving room for interpretation (Cleveland, 2012). There is a common stereotype in call centres that the job is basic and simple, which deceives management when seeking to fairly evaluate agents by measuring their productivity (Bain & Taylor, 2000). The second challenge is that call centres are dynamic and frantic, with customer interactions and massive calls coming in through various communication channels (Cleveland, 2012). The manual sorting of recorded calls over a given period is critical, impractical, and unaffordable. The third challenge is the diversity of the evaluators' perceptions, as different people can evaluate the same agent's performance differently. When the baseline is overlooked, the absence of a unified evaluation system can have a significant and negative impact on a call centre.

From a theoretical perspective, the resource-based theory (RBT) asserts that firms can achieve a sustained competitive advantage by collecting and integrating rare, valuable, inimitable, and nonsubstitutable resources (Barney, 1991). Sirmon (2007) discussed this subject and stated that acquiring rare, inimitable, and nonsubstitutable resources is necessary but insufficient; firms have to manage the resources efficiently to achieve competitive advantage. The resources portfolio structuring (tangible/intangible), bundling of equipment, technology, and human capital to build capabilities and leverage capabilities all contribute to the creation of value (Sirmon et al., 2011). Value creation in operations management is a process that involves developing or implementing operational strategies, ensuring efficient performance and high-quality products, and increasing customer satisfaction through the use of a diverse range of resources and skills (Gunasekaran & Ngai, 2012). The firm maintains a sustained competitive advantage as long as its resources are immobile and heterogeneous. Heterogeneity exists when the resource mix of one organisation differs greatly from that of another. Immobile means that the resources cannot be moved from the organisation to a competitor in order for rivals to imitate the value.

The RBT proposes a relationship between resources and competitive advantage. However, it does not propose a mechanism to ensure regular and objective choices for keeping the competitive edge, i.e., management perception of a valuable resource that may not be valuable (Bromiley & Rau, 2016; Wegge et al., 2006). Secondly, subjective evaluation occurs when management is unable to objectively judge or assess the resources because of

the ambiguity effect. Causal ambiguity (characteristic/lexical) is a critical factor causing subjective decisions due to the effect of uncertainty (ambiguity) (Powell et al., 2006; Mosakowski, 1997) argued that ambiguity does not reside in the resources but in the people themselves, so that the managers perceive the performance of resources and causes differently. The disadvantage of RBT is that it creates ambiguity in the management's understanding of the resources' competency, which is likely to be the same as it is with competitors but with a difference in form and substance: (1) The difference in form is because the management has the advantage of full access to the process flow, documents, and reports to reveal the ambiguity much more than competitors. (2) The difference in substance is the management's differing perspectives from rivals. For example, regardless of the actual evaluation, management assesses resources above the average for self-serving purposes, i.e., management's self-interest (Powell et al., 2006). When it comes to management perceptions of rivals' competencies, they are subject to self-judgement bias by ignoring competitors' competencies even when there is a significant difference (Klar & Giladi, 1999).

The previous challenges broaden the evaluation aspects by conducting a thorough analysis of customer calls. As a core communication channel for the call centre business, customer recorded or live calls represent an important aspect of agent performance. The analysis should focus on the most salient aspects of the agents' conversations that are relevant to performance evaluation. The customer call includes numerous temporal features such as the call script, appropriate responses, communication skills, and emotional control (Cleveland, 2012). Defining the salient features of the calls and determining their relationships using legacy quantitative methods, such as regression, is difficult due to the exclusion of subjective factors. Assuming, for the sake of argument, that it is possible to construct a comprehensive regression analysis between the performance features and relevant coefficients, there is still a gap in dealing with unstructured data such as text and speech, which is questionable. Furthermore, when using inferential instruments such as regression analysis in classification studies as machine learning, the research paradigm is misplaced (Li & Tong, 2020). According to Li & Tong (2020), regression analysis is concerned with hypothesis testing and determining the factors' relationships, among others. The classification problem, on the other hand, is dedicated to data-driven analysis by focusing on similarities and anomalies in their structures. It implies that the role of the machine learning paradigm is to deal with massive amounts of recorded calls for classification while excluding subjectivity.

Machine learning is capable of performing a variety of tasks in a sophisticated manner, including self-learning and adaptation for performance enhancement (Bengio et al., 2003). The tasks performed by machine learning either reduce human intervention for repetitive and daily tasks or go beyond human capabilities for large and complex data sets that cannot be articulated by humans (Alzubi et al., 2018). According to Alzubi et al., (2018), machine learning can perform a variety of complex tasks such as problem classification, anomaly detection, regression, clustering, and reinforcement learning. As a result, a deep learning classification algorithm will be used in the study to model and classify unstructured data and extract the subjectivity factors embedded in speech features.

Data modelling and analytics have become the core aspects of performance evaluation and enhancement (Akter et al., 2016). Machine learning has achieved significant contributions in operations management for performance evaluation and data modelling (Choi et al., 2018; Wamba et al., 2017). Data-driven machine learning approaches dominate the studies concerned with innovations, especially in rapid operations like call centres, where big

data agglomerates (Akter et al., 2020). Furthermore, call centres have become an effective communication channel that provides rich data sets and creates outstanding data-driven informational opportunities in various industries, like telecom, banking, healthcare, and the public sector (Zillner et al., 2016). Much innovation and research seek to take advantage of the data collected in call centres, with significant contributions to performance measurement (Echchakoui & Baakil, 2019; Helper, 2019a; Hudson et al., 2017; Ibrahim et al., 2019; Karakus & Aydin, 2016; Shire et al., 2017).

Machine learning data modelling is booming in emotional recognition (Busso et al., 2008; Hifny and Ali, 2019), complaint analysis (Bae et al., 2005), performance applications (Ahmed et al. 2016; Ahmed et al., 2018; Perera et al. 2019), and quality of service (Karakus & Aydin, 2016; Sudarsan & Kumar, 2019). As a result, several attempts have been made to apply machine learning algorithms to the objective performance evaluation of call centre agents using data mining, predefined factors, speech recognition for transcription, word analysis, language modelling, and customer feedback analysis (Ahmed et al. 2016; Ahmed et al., 2018; Ahmed et al., 2020; Ahmed et al., 2021; Helper, 2019b; Ibrahim et al., 2019; Paprzycki et al., 2004; Wöllmer, 2013). Unfortunately, previous studies did not take subjective factors into account during the evaluation process, making them less useful because the resulting evaluations are most likely collections of individual evaluators' subjective perceptions. Avoiding subjectivity based on automated detection is essential to ensuring that the resulting calls can be considered fair evaluation. Furthermore, these studies focused on specific predefined evaluation factors, i.e., slang words out of context for text processing. Also, previous studies ignored contextual speech features and the vocal tract in favour of focusing on limited aspects such as energy, rhythm, and emotion detection. Considering the entire call or long speech segments to train the machine for a comprehensive view of the conversation rather than spotting minor factors is essential. Consequently, giving the machine the role of detecting eminent utterances and shedding light on significant parts of the call is critical for better revealing the aspects of performance and subjectivity embedded in the call. As a result, the study aims to close the gap between subjective performance evaluation and agents' performance in a more objective manner. It is possible to achieve this by detecting and eliminating the calls that most likely comprise the subjective factors and focusing solely on the objective calls for evaluation.

The study is built on the concept that eliminating subjective calls results in better performance evaluation. To accurately model and classify subjective factors, data modelling using machine learning is required. Long-short term memory (LSTM) is used in this study to consider the temporal propagation of speech with the corresponding feature relations (training weights). The CNN achieved a significant improvement in speech processing, where it was proposed in the experiment to extract the best representation of the speech features. In addition, an attention layer is added in front of the CNN/LSTM combination to allow the machine to distinguish the most informative segments of the call that are most likely classified as subjective. All of these methods will be discussed in detail in the following sections.

This empirical study aims to automate the detection of subjective calls from a call centre. We seek to answer the following research questions:

- RQ1: What are the subjective factors embedded in the recorded calls that distort or bias the performance measurement of the call centre agent?

- RQ2: How can data-driven modelling be used to determine the subjective factors using machine learning?

In line with the above research questions, the first study objective is to develop a machine learning model that can detect subjective calls. The second is to determine the study baseline for the best classification accuracy of the subjective factors. Finally, the study sets out to investigate how the deep learning model can detect subjectivity. Furthermore, it proposes a graphical representation of the paralinguistic features of speech segments, which contribute to subjective factors not considered in data modelling. The study should be able to detect and eliminate the subjective segments, which is the main contribution of the research. Subjectivity detection can be extended to several counter-subjective studies, including discourse analysis and public speech. The study concludes with several recommendations for call centre supervisors and managers regarding the deployment and use of the evaluation system. The study also aims to contribute to the data-driven innovations in operations management and competitive advantage based on resource-based theory (Sultana et al., 2021, 2022). It tries to close the gap in performance measurement that affects the efficiency of operations and leads to biased strategic decisions. Moreover, the study proposes several bias detection and elimination techniques that are a common problem in the age of AI (Akter et al., 2021).

The next sections cover (1) a literature review about subjective evaluation, relevant factors, and an overview of machine learning performance measurement, (2) the proposed machine learning framework for subjectivity classification using the attention layer and paralinguistic attribute analysis, and (3) discussions and conclusions.

2 Literature review

2.1 Search procedure

A systematic review is proposed in this study to review the literature on performance measurement in the call centre domain. The systematic literature review is carried out in three stages (Tranfield et al., 2003): planning the review, conducting the review, and reporting and dissemination. The stages are illustrated in Fig. 1.

The next three sections are organised according to the stages, with each stage containing several phases. The first stage discusses subjective performance evaluation as well as relevant studies. The second stage focuses on subjective performance evaluation factors and their associated outcomes. The final stage identifies gaps in previous studies and explores the best methods for performance modelling to be used in the experiment.

2.2 Stage-1: Exploring subjective performance evaluation in call centres

Subjective evaluation is a broad study area in business that deals with different paradigms. Firstly, it is essential to clarify the definition of performance evaluation between human resources management and operations management. Performance evaluation is a significant part of human resource management (HRM) and centres around motivation, well-being, emotional labour, and performance appraisal (Flamholtz & Lacey, 1981; Hackman & Oldham, 1975, 1980; McKelvey & Aldrich, 1983). HRM theories try to interpret the work-

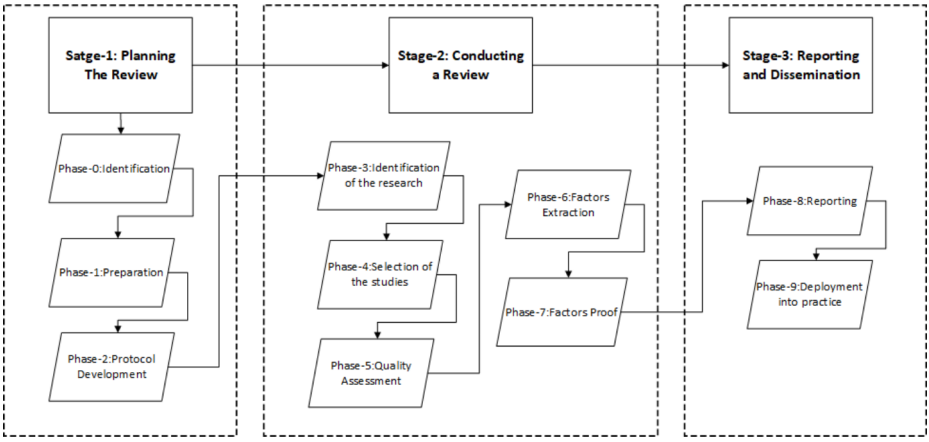


Fig. 1 Stages of a systematic literature review (Tranfield et al., 2003)

place environment and determine the factors that reflect people's performance (Grégoire & Lachance, 2015; Taylor, 1998). Operations management (OM) is concerned with performance from a different perspective of productivity, efficiency, quality of service, and resources' capabilities (Aksin et al., 2007; Gunasekaran & Ngai, 2012). Productivity is commonly understood as the ratio of outputs produced to resources consumed (Card, 2006). Therefore, the current study is concerned with performance measurement from an operations management (OM) perspective.

The subjective evaluation process has many drawbacks because of the evaluators' perceptions (Judkins et al., 2003). Subjective performance evaluation is limited by collusion (Tirole, 1986), influence costs (Milgrom, 1988), bias (Prendergast & Topel, 1993), leniency in rating efficiency (Kane et al., 1995) and favouritism (Breuer et al., 2013; Prendergast & Topel, 1996). Research has shown that subjective performance evaluation is sometimes biased toward recent performance, which is not comprehensive; or towards future performance, which has not happened yet (Frederiksen et al., 2017). Because of common stereotyping, the evaluator's perception has a severe and prominent effect on biasing the agent's evaluation. The management, including evaluators, has a stereotype that 'call centres are neither complicated nor demanding and most of the interactions are basic, simple and scripted' (Wegge et al., 2006, p. 61). On the other hand, the agents perceive their job nature as 'demanding and almost needing high attention through simultaneous subtasks', such as listening and asking questions, operating the keyboard for data input, reading data on the screen, and answering the customer (Wegge et al., 2006, p. 61).

Subjective evaluation in call centres is different according to the resource type. For human resources, subjective evaluation is the essence of the qualitative method, which is performed by monitoring and evaluating the interactions between the agent and the customer according to the evaluator's perception (Frederiksen et al., 2017). It is performed by listening to the agent's recorded call, taping a live call or a test call, i.e., a mystery shopper (Rubingh, 2013). The quality team listens to the agents' recorded calls and uses predefined evaluation forms (evaluation checklist) (Reynolds, 2010). It is concerned with the communication style, like if the agent listens carefully to the customer or if the agent's tone of voice is clear, and with the agent's knowledge and their competency in performing the task (Marr & Neely, 2004).

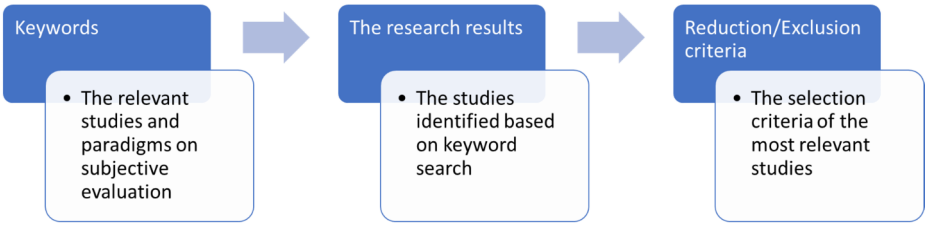


Fig. 2 The Literature Review Strategies

Those standards are subjective because they allow for style and individuality and provide/require room for interpretation (Cleveland, 2012, p. 343).

The subjective evaluation is not limited to human resources (agents) but is extended to include the process structure (Angelovski et al., 2016), the applied technology (technologies) (Shachaf, 2008), the management perception (Powell et al., 2006), and customer feedback (Abbott, 2004). Management may overestimate or underestimate the technological capability for achieving strategic goals. For customer satisfaction, many call centres have fallen into the trap of believing that the call duration and the average time to answer are objective measures of customer satisfaction (Marr & Neely, 2004). On the contrary, those factors measure efficiency and financial performance rather than actual performance. It is obvious that most of the common practices in call centres are subjected to individual evaluation, which leads to a biased evaluation process. It implies that subjective evaluation has a high impact on strategic decisions and may cause substantial inefficiencies, reflecting the competitive advantage.

A typical challenge in performance evaluation studies is that true performance is not observable to the researcher. Hence, it is hard to assess the gap and detect evaluation distortions (Breuer et al., 2013). The subjective evaluation may underestimate the agent when his performance could be better. On the other hand, the agent may be overestimated in the evaluation because of other factors that may not be relevant to the true performance or the quality of service.

The literature review planning in this section has been conducted based on the three main strategies illustrated in Fig. 2.

The keywords are a combination of several synonyms and words used in subjective call centre evaluation. Table 1 lists the keywords and their corresponding search results.

The reduction criteria were applied by locating intersections between the aforementioned topics where the most relevant journals contribute the greatest number of keywords.

Table 2 presents relevant previous studies on subjective evaluation, the domain, and the paradigm.

Table 2 presents three categories of studies that are concerned with performance measurement. The first category is about operations management theories like Resource-Based Theory (RBT) (Bromiley & Rau, 2016; Hitt et al., 2016; Ketokivi, 2016) and the subjective factors that reside therein due to, i.e., causal ambiguity (Brito & Sauan, 2016; Powell et al., 2006). The second category is about the studies of performance measurement in call centres and its influence on workers' performance and well-being (Ahmed et al., 2018; Deery et

Table 1 Keywords and search results

#	Keywords	The Initial Search Results Count (papers)
1	Call centres' agents	1900
2	Performance evaluation	1200
3	Subjective evaluation	350
4	Resource-based theory (or 'resource-based view')	220
5	Productivity measurement	2000+
6	Management perception	320

al., 2002; Wegge et al., 2006). The third category discusses subjective performance evaluation and related factors in different perspectives (Breuer et al., 2013; Frederiksen et al., 2017). The previous studies formulate the cornerstone of subjectivity performance evaluation, especially in call centres. Accordingly, the problem statement can be formulated under the following aspects; firstly, it is essential to determine the factors that impact the subjective performance evaluation in call centres. Secondly, eliminating subjective factors from the performance evaluation may lead to a better performance evaluation in the call centres. Thirdly, extracting the subjective factors from the call may lead to automating the subjective detection to better evaluate the agents' performance.

2.3 Stage-2: Conducting the review for the subjective factors

Referring to Stanton's categorisation of performance monitoring, he mentioned three types of monitoring: the source (i.e., supervisory, peer, or self-monitoring), the frequency of monitoring, and the target of monitoring (Stanton, 2000). The source and frequency of monitoring are relevant to management perception. The target of monitoring is concerned with three types of evaluations: the task itself, the context, and the situations behind the task (Sonnentag & Frese, 2003). The job-specific proficiency factor can be considered objective because the agent's answers can be "yes" or "no" or itemised. The evaluator gives a score for each item fulfilled out of the required items. For example, when the customer asks about the balance in a bank checking account, assuming the answer is a set of five main items (verifying caller identity, mentioning the account currency, the date of final balance update, pending transactions, and the amount). When the agent misses one of the answer items, the score is deducted by one point. The second type of task performance is non-job-specific task proficiency, which includes written and oral communication proficiency (Campbell, 1990). The non-job-specific task proficiency is irrelevant to the call centre's technical core. The technical core is related to the cognitive ability of the agent to grasp knowledge and perform the task correctly (Sonnentag & Frese, 2003). A non-specific-job task is an individual skill, such as communication, oral, or listening skills. Non-specific-job proficiency creates room for individual evaluation because behaviour and attitude cannot be standardised like technical tasks (Deery et al., 2002). For example, listening skills are those parts of communication skills concerned with conversation switching points (Wooffitt, 2005). The agent knows when he/she can start talking without interruption to the customer. As the agent can be trained to respect and carefully handle the conversation, repeated interruptions by the customer may confuse the evaluator, who might consider it a sign of a lack of listening

Table 2 Relevant subjective evaluation studies

#	The Study	Domain	Paradigm	Reference
1	The effect of subjective evaluation on the employee's career.	Economic Behaviour & Organisations – Economic Reviews	HRM	(Frederiksen et al., 2017)
2	The study discusses how bias or discrimination affects how agents are paid and how their performance is measured.			(MacLeod, 2003), (MacLeod & Tan, 2016)
3	The social ties between the hiring managers and the employees.			(Angelovski et al., 2016)
4	The study is about the bias in paying employees' compensation.			(Prendergast & Topel, 1993)
5	The social ties between managers and employees in the workplace.	Management Science		(Breuer et al., 2013)
6	The subjective evaluation of organisational performance and its reflection on marketing orientation.	Marketing management		(González-Benito & González-Benito, 2005)
7	The effect of performance evaluation on people's performance and work experience.	Human Resources		(Deery & Kinnie, 2002)
8	People's motivation and well-being in call centres. Other studies are focusing on emotional exhaustion and the reflection on performance and evaluation, as well.		HRM	(Wegge et al., 2006); (Taylor et al., 2002) (Bain & Taylor, 2000) (Taylor & Bain, 1999) (Taylor, 1998) (Grebner et al., 2003)
9	The studies are trying to determine performance through language features and machine processing.		IS	(Friginal, 2013); (Ahmed et al., 2018); (Carmel, 2005); (Anton et al., 1999) (Ahmed et al., 2020; Ahmed et al., 2021)
10	The study examined service providers' behavioural discretion regarding the length of service time and the variables that affect their discretion over 225 call centre employees.		Operations Management	(Gil et al., 2015)
11	The management system architecture and reliability in call centres.			(Andrade et al., 2018)
12	Examining the usefulness of RBV/RBT.	Operations management		(Ketokivi, 2016)
13	The level of management practices and their relationship with the three significant dimensions of firms' performance (profitability, growth, and productivity).	General Business		(Brito & Sauan, 2016)
14	The effect of subjective evaluation and causal ambiguity on performance.	Operations Management		(Powell et al., 2006)
15	The RBT in operations management.	Operations Management		(Hitt et al., 2016), (Bromiley & Rau, 2016)

skills. All the aforementioned skills are equivocal and evaluated according to antecedent experience and perceptions.

The second factor is the customer's behaviour during the call or series of calls. Management may perceive a call as successful when an angry customer becomes calm or funny, and vice versa. Controlling customer behaviour is an outstanding effort by the agent. Nevertheless, considering it as an evaluation factor is unadvisable because the causes are not tightly relevant to agent performance. Even in cases of 'phone rage' (Deery et al., 2002), it should ultimately be considered a subjective case to be excluded from the evaluation.

Wilson (2009) discusses handling anger over the phone and suggests four steps to treating an angry customer. The steps are: (1) listening, (2) asking appropriate questions, (3) proposing a solution, and (4) checking agreement with the customer. When offering a solution, the anger may soften or erupt again due to the proposed solution: 'The problem is that [agents] are not the person delivering the solution nor have they the responsibility of system design' (Wilson, 2009: 144). The customer behaviour fallacy proposes that the evaluator links customer behaviour to agent performance. For instance, when the caller's behaviour changes from being angry to being calm, the evaluator assumes that the agent has delivered excellent service and vice versa. Sometimes, the call is adequately performed with smooth and calm communication between the customer and the agent. However, the customer is left anxious, disappointed, or unhappy as a result rather than the manner of communication itself (Rychalski & Palmer, 2017).

Many studies highlighted subjective factors like favouritism (Breuer et al., 2013), stereotyping (Angelovski et al., 2016; Breuer et al., 2013; Stangor & Walinga, 2010; Taylor et al., 2002), self-evaluation (Suls & Wheeler, 2013), and self-serving (Bertini et al., 2019; Powell et al., 2006), as well as the customer's behaviour and non-specific job tasks. Table 3 summarises the studies discussing the subjective factors in call centres and similar domains like service disks. The factors contribute to subjective evaluation from agents' evaluations, customer behaviours, technology effects, and management perception.

The next step is to select the factors relevant to the subjective performance detection that can be applied to the selected data set, the call centre recorded calls. The next section discusses the review results and the literature review outcomes.

2.4 Stage-3: Reporting the selected factors and the proposed modelling approach

Many research studies are still motivated to objectively evaluate call centres' overall performance using machine learning technology. ML modelling can be divided into generative, discriminative, and deep learning approaches based on the features of text, speech, or both. The generative approach has been developed by Ahmed et al. (2016a), who transcribed the recorded calls into text using a speech recognition engine, then binary classified them into productive/nonproductive. The productivity modelling was based on a Naïve Bayes generative model, determining the posterior probability of the text given a productivity class with an accuracy of 67% (Ahmed et al., 2016b). A similar study was conducted based on a discriminative approach (Ahmed et al., 2018), employing Logistic Regression and Linear Support Vector Machines (LSVM). The discriminative approach could improve text classification accuracy to 83%. Another application has been developed to automatically handle call centre agent performance (Perera et al., 2019a). It is built on predefined factors like speech rate, voice intensity level, and emotional state and uses them to evaluate the performance of contact centre agents via a Support Vector Machine (SVM). However, the features

Table 3 The subjective factors concluded from the performance evaluation literature and call centres

Category	Factors(s)	Definition	References
Agents	- Non-specific job task - Traits	- Individual skills such as communication skills, oral proficiency, or listening skills. - Altruism, conscientiousness, civic virtue, courtesy, and sportsmanship - Helping co-workers and protecting the organisation	- (Sonnentag & Frese, 2003) - (Viswesvaran & Ones, 2000)
Process Structure	- The job designs - Contradictory Standards - Intersubjectivity	- Standards formulation, orientation, script generation, performance [on calls], monitoring, and feedback. - The contradiction between the standards for an action - Evaluator feedback for the agent's performance and feedback about his/her evaluation.	- Frese and Zapf (1994: 288) - (Wilson, 2009) - Cleveland 2013
Customer Experience	- Customer Behaviour - Customer Sovereignty	- Behaviour changes through the call - The customer relationship with management and power	- (Wilson, 2009; Rychalski & Palmer, 2017) - (Frenkel, 1999)
Technology Development	- Channels Development - Technology Acceptance	- The channels complexity and evaluation challenges (Dynamic Capability) - Performance expectancy	- (Sirmon et al., 2007) - (Venkatesh et al., 2003; Foroudi et al., 2018)
Management Perception	- Stereotyping - Cognitive bias - Self-evaluation - Self-serving - Evaluation exhaustion	- Management perception toward workers - Availability, Representativeness, Anchoring Bias - Evaluation based on the evaluator's skills - Evaluation based on self-interest or advantage - Evaluator exhaustion	- (Breuer et al., 2013) - (Ehrlinger & Kim, 2016) (Suls & Wheeler, 2012, 2013) (Bertini et al., 2019) (Deery et al., 2002)
Gender Differences	- Gender differences	- The non-specific task, contextual performance, job design, intersubjectivity, customer behaviour, channels development, and stereotyping	- (Belt, 2002; Mirchandani, 2005; Connerley & Wu, 2016; Wang et al., 2016)

were limited to these predefined factors, which creates a need for more investigation into other hidden factors.

A similar study has been developed to evaluate call centre representatives (Sudarsan & Kumar, 2019). It uses several transcription system APIs (google, wit, sphinx) to analyse performance based on emotions, banned words, greeting words, and competitors' names. Also, data analytics research has been conducted on recorded calls to detect the quality of service delivered to the customer. This was based on the Hadoop Map Reduce framework and utilised text similarity algorithms such as Cosine and n-gram (Karakus & Aydin, 2016). It also integrated slang word lists into the monitoring system. Previous studies relied on Natural Language Processing (NLP) as a modelling strategy, based on legacy machine learning approaches (generative/discriminative) and transcribed text, to measure agent performance. Speech processing studies have been conducted using Deep Neural Networks (DNNs) (Neumann & Vu, 2017; Hifny and Ali, 2019; Ahmed et al., 2020). DNN is required for

Table 4 The statistical data modelling for performance evaluation in call centres

SN	The Study Approach	Features	Reference
1	Statistical Analysis	Structured data	Rychalski and A. Palmer, 'Customer Satisfaction and Emotion in the Call Centre Context,' in *The Customer is NOT Always Right?* 2017 D. Chicu, M. del Mar Pàmies, G. Ryan, and C. Cross, 'Exploring the influence of the human factor on customer satisfaction in call centres,' 2019
2	Data Mining approach	Data mining for structured data	M. Paprzycki, A. Abraham, R. Guo, and S. Mukkamala, 'Data mining approach for analysing call centre performance,' 2004
3	Dis-criminative approach	Unstructured data- speech processing	K. Perera, Y. Priyadarshana, K. Gunathunga, L. Ranathunga, P. Karunarathne, and T. J. I. J. S. R. P. Thanthriwatta, 'Automatic evaluation software for contact centre agents' voice handling performance,' 2019 V. Sudarsan and G. Kumar, 'Voice call analytics using natural language processing,' 2019
		Unstructured Data – language processing	A. Ahmed, Y. Hifny, S. Toral, and K. Shaalan, 'A call center agent productivity modelling using discriminative approaches,' 2018
4	Deep learning and generative approach	Speech recognition, n-gram	B. Karakus and G. Aydin, 'Call centre performance evaluation using big data analytics,' 2016
	Generative approach	Unstructured data – generative approaches – language processing	A. Ahmed, S. Toral, and K. Shaalan, 'Agent productivity measurement in a call centre using machine learning,' 2016 G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, 'Automatic analysis of call-centre conversations,' 2005 M. A. Valle, S. Varas, and G. A. J. E. S. w. A. Ruz, 'Job performance prediction in a call centre using a naive Bayes classifier,' 2012
5	Deep learning	Unstructured data – speech processing Multimodal for text and speech processing	A Multimodal Approach to improve Performance Evaluation of Call Center Agent Agent Productivity Modelling in a Call Center Domain Using Attentive Convolutional Neural Networks

unstructured data modelling conducting multilayer training for model parameters. Speech modelling replaces speech recognition and detects productivity directly from the voice features. It requires 13 features extracted in the frequency domain to determine Mel Frequency Cepstral Coefficient (MFCC) features, which outperform the NLP/legacy approaches. Furthermore, a multimodal approach has been developed to combine text and speech models for performance measurement (Ahmed et al., 2021). That study extended the speech features from 13 MFCC to 65 Low-Level Descriptors (LLD) and the text from a bag of words to embedded words, outperforming the previous baseline with an accuracy of 93%. Table 4 lists the statistical data modelling for performance measurement in call centres.

The drawback of the previous studies is that they did not consider subjective factors through the evaluation process. Hence, the resulting evaluations most likely comprise the evaluators' perceptions, making them less useful. Avoiding subjectivity is critical to ensure that the resulting calls can be considered fair evaluations. The next section proposes the

Table 5 DSRM applied to the current study

DSRM activities	Activity description	Knowledge base
Problem identification and motivation	Failure to evaluate agent performance correctly due to the existence of subjectivity in the Call Quality Monitoring evaluation process.	Literature review. Understanding the shortcomings of the traditional manual process of sorting, selecting, and analysing calls for evaluation purposes. Real-world problem.
Define the objectives of a solution	Design an approach that can automatically detect subjective calls.	Literature review. Knowledge of existing tools.
Design and development	Design an evidence-based, machine learning-driven framework for subjectivity classification using Deep Neural Networks.	Convolutional Neural Network, Long-Short Term Memory Network, Attention Layer
Demonstration	Case study demonstration. The proposed approach is used to detect subjective calls in a corpus of seven hours of recorded calls from a real-estate call centre in Egypt.	Applying the proposed approach to a real-world case study.
Evaluation	Comparative analysis.	Understanding of the current solution and its advantages.

Note. The DSRM activities described here follow the arrangement that can be found in the works by Peffers et al. (2008), Charles et al. (2019), and Tsolas et al. (2020)

study framework for subjectivity detection. It discusses the experimental procedures for data preparation, modelling, and classification.

3 Method

We approach the problem as one of design and position it in view of Design Science Research Methodology (DSRM). We develop a simple artefact, an evidence-based machine learning-driven framework. The artefact possesses two essential characteristics: relevance and novelty (Geerts, 2011; Hevner et al., 2004). On the one hand, it is relevant. It solves the ongoing problem posed by subjectivity in the traditional Call Quality Monitoring evaluation process through a machine-learning framework. On the other hand, it is novel in the sense that although there have been previous speech processing studies using a machine-learning framework, they did not consider the subjective factors present in the evaluation process.

In line with the above, the design problem (see, for example, Wieringa, 2014) can be formulated as follows: *Improve the Call Quality Monitoring evaluation process by designing an evidence-based machine learning-driven framework that can automatically detect subjective calls in order to measure agent performance more accurately*. In Table 5, we discuss the nominal sequence of activities (Peffers et al., 2008) relevant to the present study's context of creating the artefact. The first column lists the DSRM activities. The second describes each of these activities. The third indicates the materials or resources (i.e., models, methods, theories, instruments, and frameworks) from and through which the activities are executed (Hevner et al., 2004).

Proceeding further with a description of our approach, the study aims to classify the recorded calls as subjective or nonsubjective. Then, the nonsubjective calls are forwarded to productivity models to classify them as productive or nonproductive. The proposed framework is shown in Fig. 3.

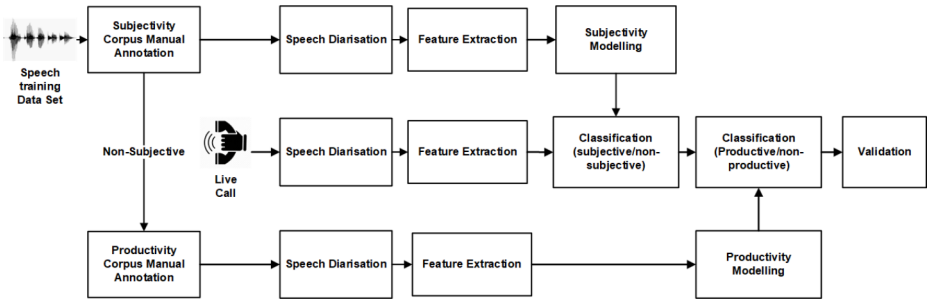


Fig. 3 The proposed framework for productivity measurement

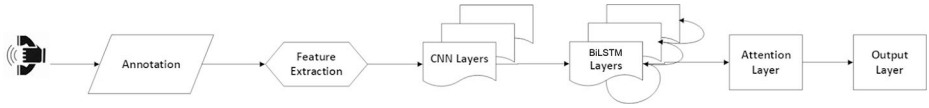


Fig. 4 The study framework

Figure 3 illustrates the complete framework for productivity measurement. First, as illustrated later, the calls go through cascaded preparation stages of annotation, diarisation, and feature extraction. Then, data modelling is conducted for the annotated calls to classify them as subjective/nonsubjective. Finally, the nonsubjective calls are forwarded to productivity models for evaluation. The experiment focuses on subjectivity classification, for which productivity evaluation, elaborated in (Ahmed et al., 2016b, 2018; Ahmed et al., 2020; Ahmed et al., 2021), has been excluded. The study proposes a DNN using cascaded Convolutional Neural Networks (CNNs), Long-Short Term Memory Networks (LSTMs) and an attention layer to determine the best accuracy for subjectivity classification. The study framework is illustrated in Fig. 4. The next section briefly discusses the CNN, LSTM, and attention layers.

The cascaded methods are proposed to provide the ultimate training weights to achieve the highest accuracy. The CNN is proposed to highlight the best prominent weights through several iterations, similar to feature extraction. To consider the context over a long stream of training weights, the DNN weights are propagated to the cascaded BLSTM. Then finally, the attention layer prioritises the most informative and distinguished weights in the voice frame over the remaining vector values. Each layer is eager to highlight the best part of the subjective aspects and pass them on to the classifier for the final and highest classification accuracy. The next sections go over each DNN layer and its mathematical representation.

3.1 Long short-term memory networks (LSTMs)

The LSTM is a type of neural network based on cell processes, where the state of the previous cell is used to produce the new state $\mathrm{C_t}$, as shown in Fig. 5. The output state $\mathrm{C_t}$ is controlled by three gates, each of which serves to control the amount of information in the cell: the *input gat*e determines which information to add to the current memory, the *forget gate* determines which to delete, and the *output gate* determines which to output from the current memory. When the sigmoid function $\mathrm{f_t}$is small and multiplied by the previous state, $\mathrm{C_{t-1}}$ is eliminated and marked as a repeated state, thus making room for a newer state. A

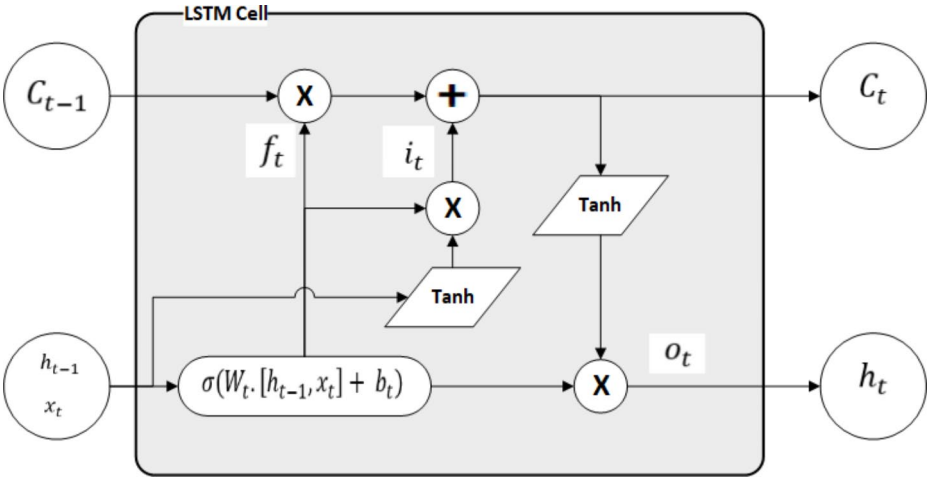


Fig. 5 The LSTM cell

high value for the sigmoid function means it accumulates previous states, meaning that this is a newer state that has never been repeated.

The state equations at each time step t are as follows:

$$i_t = \sigma\left(w_i h_{t-1} + U_i x_t\right) \tag{1}$$

$$f_t = \sigma\left(w_f h_{t-1} + U_f x_t\right) \tag{2}$$

$$o_t = \sigma\left(w_o h_{t-1} + U_o x_t\right) \tag{3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh\left(w_c h_{t-1} + U_c x_t\right) \tag{4}$$

$$h_t = o_t \odot tanh\left(C_t\right) \tag{5}$$

Where i_t is the input gate, f_t is the forget gate, o_t is the output gate, C_t is the cell state and h_t is the hidden state. $\odot$ denotes element-wise multiplication. Wi, Ui, Wf, Uf, Wo, Uo, Wc, and Uc are the weight matrices (parameters) of the LSTM network. A variant of LSTM known as bidirectional BiLSTM (Graves & Schmidhuber, 2005) allows the integration of both past and future information. It combines LSTMs operating in two directions: one forward and the other backward. Hence, each input word at time t is aware of the past and future contexts, which improves accuracy. The LSTM replaces RNN for temporal dependencies longer than simple RNNs and helps overcome the gradient vanishing problem, which occurs when the gradient becomes very small for long stretches, preventing the weights from being updated (Hochreiter, 1998).

3.2 Convolutional neural networks (CNNs)

CNN is a modified version of DNN for cases of a massive volume of data, i.e., image processing, which requires a big network, vast parameters, many resources, and a lot of

time. The CNN structure forms the layers into processing parts, i.e., image dimension, and generates the corresponding parameters in terms of filters. A filter is a vector of trainable parameters concerned with a smaller part of the whole image, for faster and more accurate processing.

Data training based on CNNs is split into three stages. The first stage is feature extraction using the filters (called the kernel) multiplied with a part of the data features, i.e., voice frames, before summing them up to produce a more concentrated and better data definition than the original (Albawi et al., 2017; Teow, 2017). The second stage sees the weights produced from the extracted features propagated to the neural networks for training to determine the weights that best match the input data. This is performed by forward or backward propagation of the data. Forward propagation is concerned with randomly initiated weights multiplied by input data and passed to the activation function. Backward propagation uses gradient descent to determine the optimum weights (Murugan, 2017). The third stage is the classifier, using the sigmoid function for binary classification or the Softmax function for multiple classes.

3.2.1 Attention layer

The attention layer focuses on the critical parts of the hidden weights. It converts the sequence of frames into a context vector for the final layer (the classifier).

Figure 6 illustrates the role of the attention layer in the modelling process:

- The modelling layers are shown in solid grey.
- The attention layer is the dotted box, including the circles representing the calculation of the attention weights. The Softmax function calculates the attention weights and generates the context vector C.
- The context vector is fed into a dense layer with a $tanh$ activation function.
- The output layer is another Softmax function for the three-class classification outlined below in Sect. 4.2.

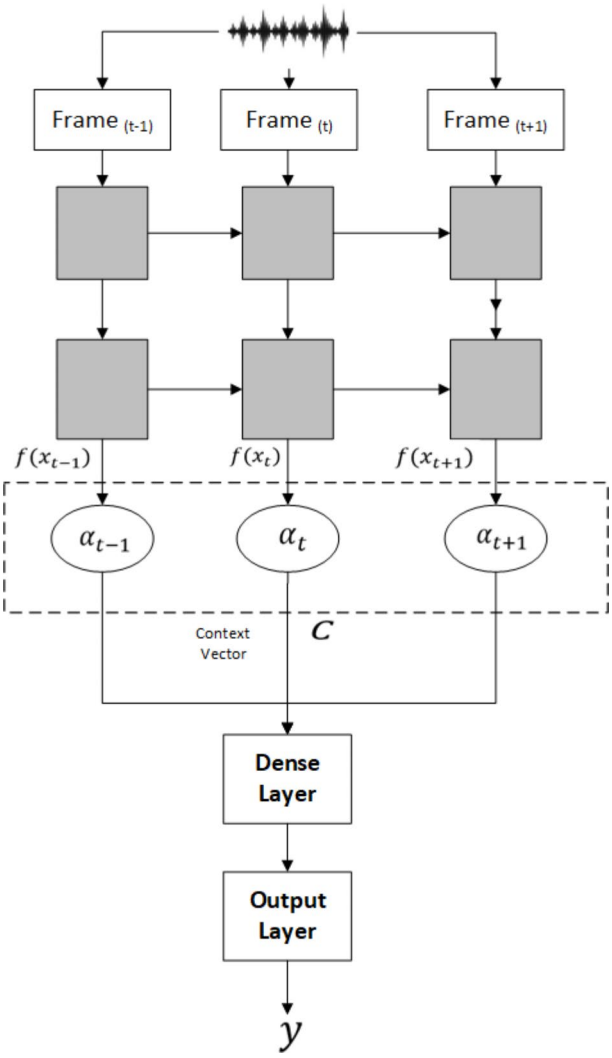
The context vector is the weighted average of the weights of the hidden layer and the input data. The Softmax function is used in the attention layer to determine the strength of frame occurrence for the remaining frames at time t. It is similar to the probability determination of a vector occurrence among the remaining vectors in a stream when the total counts are equal to 1. For each frame vector x_t in a sequence of inputs $x_1, x_2, \ldots, x_T$, given activation function f, the attention weight α_t is given by Eq. (6):

$$\alpha_t = \frac{e^{f(x_t)}}{\sum_{j=1}^{T} e^{f(x_j)}} \tag{6}$$

The context vector generated is calculated as in Eq. (7):

$$C = \sum_{t=1}^{T} \alpha_t x_t \tag{7}$$

Fig. 6 The attention layer



4 Data analytics and inferences

4.1 The data

The availability and accessibility of the data set are challenges that require exposing a contact centre's internal and confidential data for modelling. Hence, recorded calls were collected by Luminous Technologies for research purposes from a real estate call centre in Egypt. A VoIP call centre with a built-in call recording system was deployed between 2014 and 2015 to collect real calls over landline phones with a sampling rate of 8 kHz. The selected random calls consisted of seven hours and over 30 calls (14 min per call on average), which was considered adequate compared to similar studies (Hifny & Ali, 2019; Ahmed et al., 2020). The corpus comprised six different agents, between 25–35 years old; two females and four

males. The calls were diarised previously in Ahmed et al. (2020), and the talking time was 40% for females and 60% for males. The naming convention for the recorded calls was built from the metadata: Date, Time, Agent ID, Speaker ID (by the diariser), call direction (Inbound, Outbound). The modelling process goes through sequential stages: data preparation, data modelling, and classification, as elaborated upon in the next sections.

4.2 Stage 1: Data preparation

Data preparation is an essential process for data modelling (Richert et al., 2013). It first requires focusing on the targeted data to be processed. This is the raw data fed to the machine for processing. It usually excludes irrelevant, noisy, and corrupted content. Arranging, preparing, and cleaning the data is essential to getting optimal outcomes. The machine learning experiment will be applied to the two study variables: non-specific-job and customer behaviour reflected in two classes. An additional class is required for the nonsubjective cases, classified as 'nonsubjective', for a total of three classes.

The first step, as per Fig. 4, is data annotation for the seven hours of recorded calls. The annotation is performed using 2–3 participants based on 'the technical core', which is considered 'nonsubjective'. If the participant's response is not part of the technical core, such as oral proficiency or customer frustration, it will then be considered 'subjective' and categorised under the corresponding variable. The annotation procedure is performed on recorder files (calls) located in the system folder. The participant is requested to copy the file to the destination folder (the annotated class folder, i.e., agent-subjective). The annotation process is vulnerable to subjective evaluation, impacting the practical application via possible bias in the results. Hence, the annotation should be verified by Cohen's Kappa (for two raters) or Krippendorff's Alpha (for more than two raters).

Two raters are selected as a minimum for Cohen's Kappa, with five and eight years of experience in call centres, respectively. They are full-time supervisors with experience in the real estate domain. First, they had an orientation session on subjectivity and how subjective factors can be defined by listening to recorded calls. Then, they started rating the recorded calls as nonsubjective, non-specific-job tasks for the agent (subjective), or issues of customer behaviour (subjective). Finally, they were asked to move each audio file to a folder belonging to the designated class (agent folder, customer folder, and nonsubjective folder). The Cohen's Kappa method is used to verify the raters' agreement using IBM SPSS for data analysis.

For the nonsubjective class, rater 1 and rater 2 agreed on 580 audio files and disagreed on 33 files for the customer class and 37 files for the agent class. They agreed on 52 files for customer behaviour, 140 for agent differences and 580 for nonsubjective calls. They disagreed on 87 files. The Kappa agreement was $\alpha = 0.767$, and the total number of files $N = 859$. The agreement between the raters should be more than 80% ($\alpha > 0.8$) to consider the annotation valid (Hayes & Krippendorff, 2007). However, the value of $\alpha = 0.767$ is deemed acceptable because of the small difference (less than 3.35%). A Krippendorff's alpha can be calculated using the Python Natural Language Toolkit (NLTK) and its library 'Agreement' for two raters or more.

The second step in data preparation is feature extraction, which is the most challenging part, not only in this study but in machine learning as a whole (Dave, 2013). The features may include word-based language models; letters, or lexicon models; or speech signals

Table 6 Machine Learning Parameters Summary

Layer	CNN-LSTM-Attention
Input	13
CNN-1	256
Max Pooling-1	256
CNN-2	64
Max Pooling-2	64
BiLSTM-1	128
BiLSTM-2	128
Attention	128
Dense	64
Output (Classifier)	3

for acoustic modelling. MFCC conversion is a way to convert audio signals into numbers (frequencies) to identify the salient features (coefficients) from the audio file and ignore unimportant features (noise). MFCC extraction goes through multiple stages: windowing, discrete Fourier transform, and extraction of the voice coefficients. The Essentia toolkit is used to extract 13 MFCCs (Bogdanov et al., 2013). MFCC feature extraction is performed by going through the audio files with a 25 millisecond (ms) window size and 10ms steps for each window (overlapped) (Zheng et al., 2001). Each frame is represented by 13 MFCC features, forwarded to the model input layer for training.

4.3 Stage 2: Modelling and data analytics

In the experiment, the deep neural network (DNN) is constructed with the following layers: the input layer, the convolutional neural network (CNN) layer, the long short-term memory network (LSTM) layer, the attention layer, and the output layer for classification. The CNN layer is required to squeeze the features and expedite the data processing. The second layer comprises the bidirectional LSTMs, two networks, one in each direction (forward-backward), to improve the training parameters. The LSTMs can efficiently process a long sequence of dependencies as compared to the legacy recurrent neural network (RNN) (Graves & Schmidhuber, 2005). LSTMs also overcome the gradient vanishing problem, which occurs when the updates to the training weights are so tiny that the network is stuck in an endless loop or prolonged training rate (Δw) (Hochreiter, 1998). The next layer is the attention layer, for highlighting the most probable frames.

The neural network structure and hyper-parameters are set following the configuration defined in the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) (Busso et al., 2008), which is followed by various previous studies in emotional recognition and performance measurement (Hifny and Ali, 2019; Ahmed et al., 2020; Ahmed et al., 2021). The model summary presents the layers and the size per layer (Table 6).

Table 6 illustrates the layer type and the size of the corresponding parameters/units for each layer. The model is read from top to bottom, indicating the data flow sequence through the training process. The model starts with an input layer that should be adjusted to the training feature size (13 features). As mentioned in the previous section, file duration is unknown, so frames are taken sequentially, with overlapping between the sequences. The feature size is known (13 features), but the model can accept an unknown number of sequential frames. The frames are forwarded to the first layer of the CNN. The CNN size is 256 filters for processing the frames of one-dimensional CNNs, denoted by (1D-CNNs).

The CNN architecture consists of 256 filters for layer 1, 64 filters for layer 2, five kernel sizes each, and a ReLU activation function. The CNN neuron output is the dot product between randomly selected weights (initial weights for the training) and part of the frame. ReLU is a linear activation function (max(0,x)), so the output excludes the negatives and continues to the next layer.

The bidirectional LSTMs (BiLSTMs) comprise 128 neurons each. A max-pooling layer is required to downsample the CNN's output and reduce the dimension to match the size of the next layer. The dense layer is 64 units to convert the dimensions of the vectors and forward them to the output layer. The data modelling starts by splitting the data randomly into training (90%) and test (10%) sets, then determining the optimum parameters. The test set is smaller because it is required only for model verification. Every iteration (epoch) is a one-time training on a batch size of 32 files from the data set to determine the test set's accuracy. The training keeps iterating until it reaches the optimum weights of the model. The model created is called 'Fold'. At this point, part of the data had not been trained (test data), so the data was shuffled and re-split into training and test sets. This process is called cross-validation, to recompose the training and test sets into the corresponding folds. The models created are denoted by 'Fold_Number_iterations_accuracy'.

The training is performed on a dual-processor server supported by 24 cores, with 2.1 GHz and 97 GB RAM. Three Nvidia cards (GPUs) have been used—TESLA k80, Quadro M4000, and M5000—to process the data on an Ubuntu 18.04 LTS operating system. The code was developed using Python 3.7, based on the Conda virtual environment, KERAS library 2.2, and TensorFlow backend 1.14.

4.4 Stage 3: Classification and validation

The Softmax function is used in the final layer for classification. The Softmax function uses Eq. (6)'s attention weights but on the class level. The output will be classified into three classes: nonsubjective, agent (non-specific-job tasks), and customer (customer behaviour). The five cross-validations have been applied and forwarded to the F1 score to verify the model accuracy. The annotation process may lead to an imbalanced data set when the annotated class sizes are quite different. An imbalanced data set may bias the accuracy, favouring the largest class (Guo et al., 2008). Imbalanced data should therefore be re-adjusted with a bigger corpus and re-annotated. Because the experiment is limited to only seven hours, the F1 score is used to verify the model's accuracy on the data set. It is based on the average precision and recall of the resulting label as compared to the annotated class (Wardhani et al., 2019).

4.5 The experiment findings

The model accuracy is calculated using the average of the accuracies of the folds generated. The accuracy of a subjective evaluation is 82.53%, so the error rate is around 17.46%. Table 7 states the accuracy for each fold in the experiment. The total number of folds (models) is five.

When reaching the ultimate accuracy, an early stopping configuration is used to stop modelling the current fold and move to the next fold. Several optimisation methods, such as

Table 7 Subjectivity models and corresponding accuracies

#	The Fold name	Iterations	Accuracy
Fold-1	Fold0-030-0.82558	30	0.82558
Fold-2	Fold1-002-0.83721	2	0.83721
Fold-3	Fold2-020-0.81977	20	0.81977
Fold-4	Fold3-053-0.84884	53	0.84884
Fold-5	Fold4-058-0.79532	58	0.79532
Total		**163**	**Average=0.8253**

ADAM and SGD (Kingma & Ba, 2014), have been used. So, the iterations start at 2 iterations and progress to 58.

The main concern in the previous experiment was the imbalanced data set for each class. The file segments for each class are as follows, out of a total of 859 files: non-subjective=650 files (75%), agent=156 files (18%), and customer=53 files (7%). The Scikit-learn Python library has been used to measure this imbalanced data set (Raschka, 2015). By applying the F1 score to each of the five models' folds, the accuracy of the model is determined based on the average score over the five folds – see Tables 8 and 9.

The accuracy is less (75%) than the cross-validation accuracy (82.5%), with a difference of 7.48%. This is an effect of the imbalanced data set. By extending the corpus size in call centres, the data set may become more balanced with fewer classification errors. As there is no other study in the same context (subjectivity) using machine learning, the resulting accuracy is the baseline for subjective performance, considering the agent and customer factors. Nevertheless, a semi-comparison can be applied to other studies, like emotional recognition using speech processing and productivity measurement using text and speech classification. The accuracy of subjectivity modelling is comparable (75%) to that of emotional recognition (82.5%), performance measurement–speech-based (83%) and productivity measurement–text-based (82.69%). Table 10 summarises the subjective classification accuracy as compared to other studies.

The attention layer generates the context vector of the frames computed by the input data and attention weights. By applying Eq. (6) to the attention weights, a graphical presentation is generated of the attention weights versus the frames for all training segments. The attention weights may help indicate the linguistic and paralinguistic features relevant to the subjective factors. The four sample graphs selected are illustrated in Fig. 7.

The attention weight graphs, plotted using the '*Plotly'* python library, have peaks for specific frames, which indicate high attention for those frames. Revising the peaks over the calls/segments prompts the following observations:

- The attention weights are high for some irrelevant parts of the call, like telephone line noise, crowd background, silence, etc. Therefore, this part is dropped from the analysis.
- The attention weights are high for high tones, especially for angry customers, which indicates a high level of subjectivity.
- The attention weights are high in the cross-talk parts, indicating poor listening skills or an issue in handling the switching points of the conversation between the agent and the caller.
- The nonsubjective folder, the customer folder, and the agent folder have almost the same call content. The differences are only relevant at the tone level, which indicates

Table 8 F1-score for the Five Models' Folds

	Fold-1			*Fold-2*			*Fold-3*			*Fold-4*			*Fold-5*		
Class	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***	***Precision***	***Recall***	***F1-score***
Agent	0.667	0.25	0.364	0.8	0.5	0.615	0.667	0.5	0.571	0.444	0.5	0.471	0.625	0.625	0.625
Customer	0.667	0.857	0.75	0.667	0.857	0.75	0.6	0.857	0.706	0.75	0.857	0.8	0.778	0.99	0.875
Non-subjective	0.778	0.875	0.824	0.84	0.875	0.857	0.826	0.792	0.809	0.818	0.75	0.783	0.864	0.792	0.826
Accuracy			0.744			0.795			0.744			0.718			0.795
Weighted Average			0.716			0.782			0.741			0.722			0.794

that tone is a subjective factor in the agent and customer parts, but that the content remains the same.

5 Discussion

The study is inspired by the resource mix of RBT to define call centre resources that contribute to subjective performance evaluation. It has been concluded that two variables (agent non-specific task and customer behaviour) have a direct influence on subjective performance. Accordingly, the theoretical contribution to the RBT theory is providing a mechanism to prevent causal ambiguity from influencing performance evaluation and, in turn, the management strategies for sustained competitive advantage (Barney, 1991, 2001; Barney et al., 2001). It is recommended that RBT states that "the valuable, rare, inimitable, nonsubstitutable resources that are objectively evaluated and positioned lead to an organisation's sustained competitive advantage" or that "the sustained competitive advantage is created by the contribution and fair judgement of valuable, rare, inimitable, and nonsubstitutable resources".

The study supports Powell's (2006) study about how management perceptions about performance are influenced by the organisation's resources (agents). Furthermore, it expands on Breuer's (Breuer et al., 2013) findings about the subjective evaluation due to stereotyping and social attachment to include other external factors like customer behaviour. The study outcomes help mitigate common issues in call centres of the finer standards like communication skills, tone, and language proficiency that allow for style and individuality and provide room for interpretation (Cleveland, 2012, p. 343). The study highlighted significant customer behaviour issues in call centres (Rychalski & Palmer, 2017; Wilson, 2009). It proposes excluding those variables from the evaluation, which are known to have a long debate in the literature on call centres. Furthermore, the study helps measure performance with less subjectivity based on the conceptual model.

Machine learning is a cutting-edge technology currently being used to train models for specific human behaviour. Based on the theoretical model variables (factors), it could automatically classify subjectivity in evaluation. Subjectivity detection is not the same as emotion recognition, which served as a baseline for this study. The data modelling process is divided into three steps: data preparation, data modelling, and data classification. The machine requires manual annotation to train each class and create the final model. In cases of detected subjectivity, the data-driven model specifies the classification classes as either agent class or customer behaviour class. It can also combine data from various communication channels, such as evaluation forms, manager meetings, chats, emails, and social media, in a manner similar to recorded calls. The model can be extended to include additional annotation classes for classifications based on the extraction of relevant features. It can explore additional information about subjectivity that the literature review may have missed. To study the results, the classification process can be exposed to graphical presentations or charts. The attention layer was used in the current study to draw attention weights to focus on the most informative features in the recorded calls. These subjective features, which were divided into linguistic and para-linguistic features, were relevant to a specific frame in the call. In a subjective manner, the study concluded with para-linguistic features from

Table 9 F1-Score average for the five folds

#	The Fold name	Weighted scoring
Fold-1	Fold0-030-0.82558	0.716
Fold-2	Fold1-002-0.83721	0.782
Fold-3	Fold2-020-0.81977	0.741
Fold-4	Fold3-053-0.84884	0.722
Fold-5	Fold4-058-0.79532	0.794
Total		**Average = 0.75**

Table 10 Relevant studies accuracy comparison

#	Study	Classification method	Type	Accuracy	References
1	Productivity Measurement	Naïve Bayes	Text	67.3%	(Ahmed et al., 2016b)
2	Productivity Measurement	Logistic Regression	Text	80.76%	(Ahmed et al., 2018)
3	Productivity Measurement	Linear Support Vector Machine (LSVM)	Text	82.69%	(Ahmed et al., 2018)
4	Emotional Recognition	CNN-LSTM	Speech	82.5%	(Hifny and Ali, 2019)
5	Performance Measurement	LSVM	Speech	83%	(Perera et al., 2019b)
5	Subjectivity Modelling	CNN-LSTM-Att	Speech	75%	The Current Study

the conceptual model, such as cross-talk, the agent's tone, and the customer's loudness of voice. To contribute more knowledge, the modelling process can be re-tested using different hypotheses (factors) and other statistical instruments.

The experiment proves that subjectivity can be automatically detected and classified under each study construct. The baseline generated is close to relevant studies in emotional recognition and performance classifications, which encourages further investigation in this area. The next recommended study for objective performance evaluation is to annotate and model the nonsubjective recorded call and conclude the productivity model/classifier as mentioned in previous studies (Ahmed et al., 2020; Ahmed et al., 2021).

Machine learning's proposed structure is eligible to be extended for a sequence of complicated events, actions, streams, or repetitive activity patterns. The application can classify those patterns under the level of conformity (accuracy), extending the application continuity for many different domains in the social sciences. Therefore, the application is not limited to subjectivity detection but can be extended to many other applications to assist performance management. Machine learning opens the door to examining human behaviour and attitudes in the workplace for action prediction. The application may give robust findings as long as the training data set is plentiful (features extraction from photos, videos, structured data, etc.) and the annotation is performed according to rigorous studies and hypotheses tests.

Where unsupervised learning can be applied, machine learning can be used for induction or abduction approaches. Similar to statistical correlation, the uncategorised data set can be linked to different categories. The researcher can investigate and validate the relationships by listing hypotheses based on the classified classes. Thousands of phone calls, for exam-

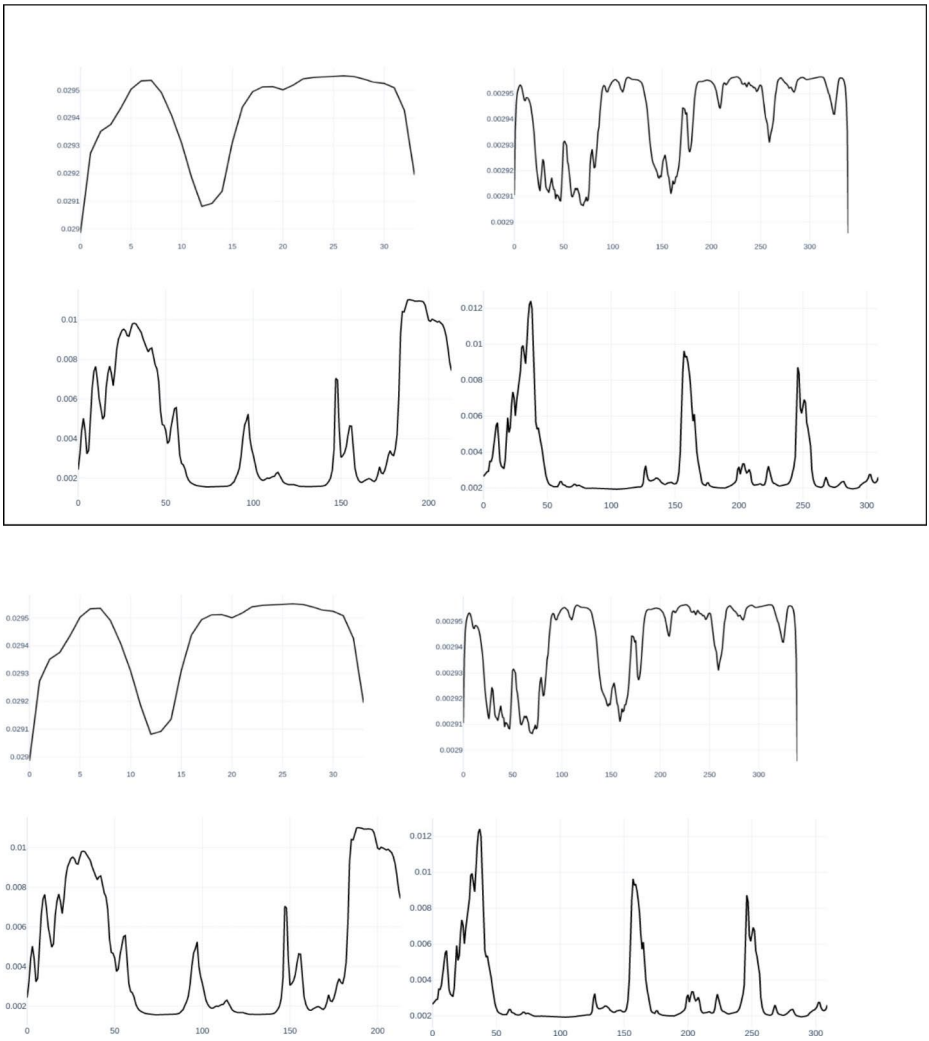


Fig. 7 Attention weight graphs (x-axis is the time in milliseconds, y-axis is the attention weights)

ple, linked to demographic data sets, locations, regions, and seasons may help researchers develop hypotheses based on the resulting classes. As a result, the research' application is based on more robust and justified constructs.

6 Conclusions

Data-driven subjectivity classification is an automatic technique to detect the subjectivity of recorded calls from call centres using machine learning. Subjectivity modelling comprises three main stages: data preparation, data modelling, and classification. As a part of data preparation, data annotation was performed based on a quantitative study to determine

the subjective factors. The model used a cascaded CNN-LSTM-attention layer for the best classification accuracy. The experiment was based on seven hours of recorded calls from a real estate call centre. The data was annotated manually and verified using Cohen's Kappa method. To the best of our knowledge, there are no previous studies regarding subjectivity modelling; so, the study contributes a baseline for subjectivity with 82.5% classification accuracy. Due to an imbalanced data set, F1 scoring was applied to determine that the model's accuracy is 75%. The accuracy suggests a preliminary baseline for subjectivity, as there is no comparable antecedent baseline based on machine learning. However, a semi-comparison with several speech studies in emotional recognition and performance measurement is useful. The attention layer aims to focus on the highest frame probabilities resulting from the SoftMax function. Furthermore, the attention weights provide an outlook on the paralinguistic features embedded in the calls, relevant to subjective and nonsubjective performance. The paralinguistics can be summarised as a high tone, indicating, as compared to a low tone, wakefulness and enthusiasm on the part of the agent; cross-talk, a subjective factor for the agent; and a loud customer voice, a subjective factor on the customer side.

The study contributes in several ways in the context of data-driven innovations. First, there is the mixed approach of combining the work of the human participants through the annotation process to achieve a more accurate classification. The raters' agreement should be significant to ensure a consistent data set for modelling. Then, the initial annotation can be extended and automated using the K-means method for distance measurement with the centroids of the annotated data. Second, the study examines the theoretical variables and data modelling using machine learning for detection and classification. Moreover, the study proposes a graphical presentation that pays attention to the significant attributes of data 'subjectivity'.

Supervised machine learning is limited to human subjective evaluation through the annotation process. This limitation is rectified by using several raters with a good reliability check (Krippendorff's alpha). Feature extraction using MFCC is commonly used to detect the vocal tract, ignoring the contextual and prosodic information. Therefore, it is highly recommended to extend the speech features by considering prosodic features for better classification and detection. Furthermore, it is recommended for future research to include classifications over other channels (chat, SMS, and social media) and to combine classification techniques (speech-text) through multimodal approaches to detect subjectivity overall across call centre channels. Cross-validation can be tested alongside other techniques such as zero-shot learning to validate classification accuracy due to an imbalanced data set where the test is set in the same domain but in a different context.

6.1 Study implications

The study resolves critical problems in call centre operations arising from unfair judgment about agent performance. Underestimating agent performance, frivolous orientations, and narrow annual raises lead to emotional exhaustion, burnout, and high turnover, which impact the business and the quality of customer service. On the other hand, less subjective evaluation keeps workers confident and loyal. It helps maintain their well-being as every piece of work is monitored and evaluated based on a unified and robust baseline. The baseline assists in uncovering customer complaints by considering the attention weights and corresponding performance attributes on the customer side. Spotting a specific part of the

call with the corresponding subjectivity level instantly reduces the time and effort necessary to identify weak points for corrective action.

Nevertheless, closely monitoring performance has several drawbacks. The panopticon concept is applied to agents through performance monitoring and evaluation, similar to how prisoners are monitored by guards. Unfortunately, this can also result in exhaustion, burnout, and turnover. It is highly recommended that agents participate in the evaluation process by sharing evaluation results with them in order to provide feedback and justify the results. This helps agents overcome their fears, improve their performance, and brilliantly enhance the ML modelling based on their experience.

Based on the resource-based theory, the study adds another chapter to the data-driven innovations in operations management and competitive advantage (Sultana et al., 2021, 2022). It draws attention to the grey area of performance measurement, which leads to evaluation bias and misguided strategic decisions. The study proposes several techniques for detecting and eliminating bias, which is a common problem in the age of AI (Akter et al., 2021).

Acknowledgements The authors are grateful to the Editor-in-Chief, the Special Issue Editors, and the anonymous reviewers for their valuable feedback on the previous version of this manuscript.

Authors' contributions: Conceptualization, A.A., U.S., and V.C.; methodology, A.A., U.S., and V.C.; software, A.A.; validation, A.A., and V.C.; writing—original draft preparation, A.A.; visualization, A.A.; supervision, U.S., and Z.I.; project administration, K.M., and U.S. All authors have read and agreed to the published version of the manuscript.

Funding: No funding to declare.

Data Availability Private data set for internal use.

Code Availability The code is custom developed using Python.

Declarations

Conflicts of interest/competing interests: The authors declare that there is no conflict of interest.

References

Abbott, J. C. (2004). *The executive guide to call center metrics*. Robert Houston Smith Publishers

Ahmed, A., Hifny, Y., Shaalan, K., & Toral, S. (2016a). Lexicon free Arabic speech recognition recipe. *International Conference on Advanced Intelligent Systems and Informatics.* Springer

Ahmed, A., Hifny, Y., Toral, S., & Shaalan, K. (2018). A Call Center Agent Productivity Modeling Using Discriminative Approaches. *Intelligent Natural Language Processing: Trends and Applications* (pp. 501–520). Springer

Ahmed, A., Shaalan, K., Toral, S., & Hifny, Y. J. S. (2021). A Multimodal Approach to improve Performance Evaluation of Call Center Agent. *Sensors, 21*(8), 2720

Ahmed, A., Toral, S., & Shaalan, K. (2016b). Agent productivity measurement in call center using machine learning. *International Conference on Advanced Intelligent Systems and Informatics.* Springer

Ahmed, A. T., Shaalan, S., & Hifny, K., Y (2020). Agent Productivity Modeling in a Call Center Domain Using Attentive Convolutional Neural Networks. *Sensors (Basel, Switzerland)*, *20*(19), 11

Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, *16*(6), 665–688

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. (2021). *Algorithmic bias in data-driven innovation in the age of AI*. Elsevier

Akter, S., Michael, K., Uddin, M. R., McCarthy, G., & Rahman, M. (2020). Transforming business using digital innovations: The application of AI, blockchain, cloud and data analytics. *Annals of Operations Research,* 1–33

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). J. I. J. o. P. E. How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Production Economics, 182*, 113–131

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET),* Antalya, Turkey

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: an overview. *Journal of Physics: Conference Series, 1142*(1), 012012

Andrade, R., Moazeni, S., & Ramirez-Marquez, J. (2018). Contact Center Operations Management Systems Architecture and Reliability. Available at SSRN: https://ssrn.com/abstract=3320821

Angelovski, A., Brandts, J., & Sola, C. (2016). Hiring and escalation bias in subjective performance evaluations: A laboratory experiment. *Journal of Economic Behavior & Organization*, *121*, 114–129

Anton, J., Bapat, V., & Hall, B. (1999). *Call center performance enhancement using simulation and modeling*. Purdue University Press

Bae, S. M., Ha, S. H., & Park, S. C. (2005). A web-based system for analyzing the voices of call center customers in the service industry. *Expert Systems with Applications, 28*(1), 29–41

Bain, P., & Taylor, P. (2000). Entrapped by the 'electronic panopticon'? Worker resistance in the call centre. *New technology work and employment*, *15*(1), 2–18

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99–120

Barney, J., Wright, M., & Ketchen Jr, D. J. (2001). The resource-based view of the firm: Ten years after 1991. *Journal of Management*, *27*(6), 625–641

Barney, J. B. (2001). Is the resource-based "view" a useful perspective for strategic management research? Yes. *Academy of Management Review*, *26*(1), 41–56

Belt, V. (2002). A female ghetto? Women's careers in call centres. *Human Resource Management Journal*, *12*(4), 51–66

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155

Bertini, M., Halbheer, D., & Koenigsberg, O. (2019). Price and quality decisions by self-serving managers. *International Journal of Research in Marketing, 37*(2), 236–257.

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., & Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4–8; Curitiba, Brazil.[place unknown]: ISMIR; 2013. p. 493-8.* International Society for Music Information Retrieval (ISMIR)

Breuer, K., Nieken, P., & Sliwka, D. (2013). Social ties and subjective performance evaluations: an empirical investigation. *Review of Managerial Science*, *7*(2), 141–157

Brito, L. A. L., & Sauan, P. K. (2016). Management practices as capabilities leading to superior performance. *BAR-Brazilian Administration Review, 13*(3), e160004

Bromiley, P., & Rau, D. (2016). Operations management and the resource based view: Another view. *Journal of Operations Management*, *41*, 95–106

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359

Campbell, J. (1990). Modeling the performance prediction problem in industrial and the impact of HR practices on the performance of business units organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (pp. 687–732). Palo Alto, CA: Consulting Psychologists Press

Card, D. N. (2006). The challenge of productivity measurement. *Proceedings of the Pacific Northwest Software Quality Conference*

Carmel, D. (2005). *Automatic analysis of call-center conversations*. Proceedings of the 14th ACM CIKM International Conference on Information and Knowledge Management, pp. 453–459. Bremen, Germany

Charles, V., Aparicio, J., & Zhu, J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *European Journal of Operational Research*, *279*(3), 929–940

Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. *Production and Operations Management, 27*(10),1868–1883

Cleveland, B. (2012). *Call Center Management on Fast Forward: Succeeding in the New Era of Customer Relationships*. ICMI Press

Connerley, M. L., & Wu, J. (2016). *Handbook on Well-Being of Working Women*. Springer

Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology*, *1*(6), 1–4

Deery, S., Iverson, R., & Walsh, J. (2002). Work relationships in telephone call centres: Understanding emotional exhaustion and employee withdrawal. *Journal of Management Studies*, *39*(4), 471–496

Deery, S., & Kinnie, N. (2002). Call centres and beyond: a thematic evaluation. *Human Resource Management Journal*, *12*(4), 3–13

Echchakoui, S., & Baakil, D. (2019). Emotional Exhaustion in Offshore Call Centers: A Comparative Study. *Journal of Global Marketing*, *32*(1), 17–36

Ehrlinger, J., Readinger, W. O., & Kim, B. (2016). Decision-Making and Cognitive Biases. Encyclopedia of Mental Health (2nd ed.), 5-12.

Flamholtz, E., & Lacey, J. M. (1981). *Personnel management, human capital theory, and human resource accounting*. Institute of Industrial Relations, University of California, Los Angeles.

Foroudi, P., Gupta, S., Sivarajah, U., & Broderick, A. (2018). Investigating the effects of smart technology on customer dynamics and customer experience. *Computers in Human Behavior*, *80*, 271–282

Frederiksen, A., Lange, F., & Kriechel, B. (2017). Subjective performance evaluations and employee careers. *Journal of Economic Behavior & Organization*, *134*, 408–429

Frenkel, S. (1999). *On the front line: Organization of work in the information economy*. Cornell University Press

Frese, M., & Zapf, D. (1994). Action as the core of work psychology: A German approach. In H. C. Triandis, M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (pp. 271–340). Consulting Psychologists Press.

Friginal, E. (2013). Evaluation of oral performance in outsourced call centres: An exploratory case study. *English for Specific Purposes*, *32*(1), 25–35

Gil, L., Iddo, G., & Dana, Y. (2015). Spending more time with the customer: service-providers' behavioral discretion and call-center operations. *Service Business*, *9*(3), 427–443

González-Benito, Ó., & González-Benito, J. (2005). Cultural vs. operational market orientation and objective vs. subjective performance: Perspective of production and operations. *Industrial Marketing Management*, *34*(8), 797–829

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, *18*(5–6), 602–610

Grebner, S., Semmer, N., Faso, L. L., Gut, S., Kälin, W., & Elfering, A. (2003). Working conditions, well-being, and job-related attitudes among call centre agents. *European Journal of Work and Organizational Psychology*, *12*(4), 341–365

Grégoire, S., & Lachance, L. (2015). Evaluation of a brief mindfulness-based intervention to reduce psychological distress in the workplace. *Mindfulness*, *6*(4), 836–847

Gunasekaran, A., & Ngai, E. W. (2012). The future of operations management: an outlook and analysis. *International Journal of Production Economics*, *135*(2), 687–701

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the class imbalance problem. *2008 Fourth international conference on natural computation.* Vol. 4. IEEE

Hackman, J. R., & Oldham, G. R. (1975). Development of the job diagnostic survey. *Journal of Applied Psychology*, *60*(2), 159-170

Hackman, J. R., & Oldham, G. R. (1980). *Work redesign.* Reading, MA: Addison-Wesley

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, *1*(1), 77–89

Helper, C. C. (2019a). *White Paper: 2019 Contact Centre Trends You Need to Know*

Helper, C. C. (2019b). *White Paper: Quality Management Automation – ROI Calculation Guide*

Hifny, Y., & Ali, A. (2019). Efficient Arabic Emotion Recognition Using Deep Neural Networks. *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and* Signal Processing *(ICASSP).* IEEE

Hitt, M. A., Xu, K., & Carnes, C. M. (2016). Resource based theory in operations management research. *Journal of Operations Management*, *41*, 77–94

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116

Hudson, S., González-Gómez, H. V., & Rychalski, A. (2017). Call centers: is there an upside to the dissatisfied customer experience? *Journal of Business Strategy*, *38*(1), 39–46

Ibrahim, S. N. H., Suan, C. L., & Karatepe, O. M. (2019). The effects of supervisor support and self-efficacy on call center employees' work engagement and quitting intentions. *International Journal of Manpower, 40*(4), 688–703

Judkins, J. A., Shelton, M., & Peterson, D. (2003). *System and method for evaluating agents in call center*. Google Patents

Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J. (1995). Stability of rater leniency: Three studies. *Academy of Management Journal*, *38*(4), 1036–1051

Karakus, B., & Aydin, G. (2016). Call center performance evaluation using big data analytics. *2016 International Symposium on Networks, Computers and Communications (ISNCC).* IEEE

Ketokivi, M. (2016). Point–counterpoint: Resource heterogeneity, performance, and competitive advantage. *Journal of Operations Management*, *41*, 75–76

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Klar, Y., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy? *Personality and Social Psychology Bulletin*, *25*(5), 586–595

Li, J. J., & Tong, X. J. P. (2020). Statistical Hypothesis Testing versus Machine Learning Binary Classification: Distinctions and Guidelines. *Patterns, 1*(7), 100115

MacLeod, W. B. (2003). Optimal contracting with subjective evaluation. *American Economic Review*, *93*(1), 216–240

MacLeod, W. B., & Tan, T. Y. (2016). *Optimal Contracting with Subjective Evaluation: The Effects of Timing, Malfeasance and Guile*. National Bureau of Economic Research

Marr, B., & Neely, A. (2004). *Managing and measuring for value: the case of call centre performance*. Cranfield School of Management

McKelvey, B., & Aldrich, H. (1983). Populations, natural selection, and applied organizational science. *Administrative Science Quarterly, 28*(1), 101–128

Milgrom, P. R. (1988). Employment contracts, influence activities, and efficient organization design. *Journal of political economy*, *96*(1), 42–60

Mirchandani, K. (2005). Gender eclipsed? Racial hierarchies in transnational call center work. *Social Justice*, *32*(102), 105–119

Mosakowski, E. (1997). Strategy making under causal ambiguity: Conceptual issues and empirical evidence. *Organization Science*, *8*(4), 414–442

Murugan, P. J. a. p. a. (2017) Feed forward and backward run in deep convolution neural network. Available at https://arxiv.org/abs/1711.03278

Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*

Paprzycki, M., Abraham, A., Guo, R., & Mukkamala, S. (2004). Data mining approach for analyzing call center performance. In B. Orchard, C. Yang, & M. Ali (Eds.), Innovations in Applied Artificial Intelligence (pp. 1092–1101). IEA/AIE 2004. Lecture Notes in Computer Science, vol 3029. Springer, Berlin

Perera, K. K. A. N. N., Priyadarshana, Y. H. P. P., Gunathunga, K. I. H., Ranathunga, L., Karunarathne, P. M., & Thanthriwatta, T. M. (2019a). Automatic Evaluation Software for Contact Centre Agents' voice Handling Performance. *International Journal of Scientific and Research Publications, 5*(1), 1–8

Perera, K. K. A. N. N., Priyadarshana, Y. H. P. P., Gunathunga, K. I. H., Ranathunga, L., Karunarathne, P. M., & Thanthriwatta, T. M. (2019b). Automatic Evaluation Software for Contact Centre Agents' voice Handling Performance

Powell, T. C., Lovallo, D., & Caringal, C. (2006). Causal ambiguity, management perception, and firm performance. *Academy of Management Review*, *31*(1), 175–196

Prendergast, C., & Topel, R. (1993). Discretion and bias in performance evaluation. *European Economic Review*, *37*(2–3), 355–365

Prendergast, C., & Topel, R. H. (1996). Favoritism in organizations. *Journal of Political Economy*, *104*(5), 958–978

Raschka, S. (2015). *Python Machine Learning*. Packt Publishing

Reynolds, P. (2010). *Call center metrics: Best practices in performance measurement and management to maximize quitline efficiency and quality*. North American Quitline Consortium

Richert, W., Chaffer, J., Swedberg, K., & Coelho, L. (2013). *Building Machine Learning Systems with Python* (1 vol.). GB: Packt Publishing

Rubingh, R. (2013). *Call Center Rocket Science: 110 Tips to Creating a World Class Customer Service Organization*. CreateSpace Independent Publishing Platform

Rychalski, A., & Palmer, A. (2017). Customer Satisfaction and Emotion in the Call Centre Context. *The Customer is NOT Always Right? Marketing Orientationsin a Dynamic Business World* (pp. 67–70). Springer

Shachaf, P. (2008). Cultural diversity and information and communication technology impacts on global virtual teams: An exploratory study. *Information & Management*, *45*(2), 131–142

Shire, K., Holtgrewe, U., & Kerst, C. (2017). Re-organising customer service work: an introduction. *Reorganising Service Work: Call Centres in Germany and Britain: Call Centres in Germany and Britain*, 1

Sirmon, D. G., Hitt, M. A., & Ireland, R. D. (2007). Managing firm resources in dynamic environments to create value: Looking inside the black box. *Academy of Management Review*, *32*(1), 273–292

Sirmon, D. G., Hitt, M. A., Ireland, R. D., & Gilbert, B. A. (2011). Resource orchestration to create competitive advantage: Breadth, depth, and life cycle effects. *Journal of Management*, *37*(5), 1390–1412

Sonnentag, S., & Frese, M. (2003). Performance concepts and performance theory. In S. Sonnentag (Ed.), *Psychological Management of Individual Performance* (pp. 1-25). John Wiley & Sons

Stangor, C., & Walinga, J. (2010). *Introduction to psychology.* Flatworld Knowledge

Stanton, J. M. (2000). Reactions to employee performance monitoring: Framework, review, and research directions. *Human Performance*, *13*(1), 85–113

Sudarsan, V., & Kumar, G. (2019). Voice call analytics using natural language processing. *International Journal of Statistics and Applied Mathematics, 4*(6), 133–136

Suls, J., & Wheeler, L. (2012). Social comparison theory. *Handbook of theories of social psychology*, *1*, 460–482

Suls, J., & Wheeler, L. (2013). *Handbook of social comparison: Theory and research*. Springer Science & Business Media

Sultana, S., Akter, S., & Kyriazis, E. (2022). How data-driven innovation capability is shaping the future of market agility and competitive performance? *Technological Forecasting and Social Change*, *174*, 121260

Sultana, S., Akter, S., Kyriazis, E., & Wamba, S. F. (2021). Architecting and developing big data-driven innovation (DDI) in the digital economy. *Journal of Global Information Management (JGIM)*, *29*(3), 165–187

Taylor, P., & Bain, P. (1999). 'An assembly line in the head': work and employee relations in the call centre. *Industrial Relations Journal*, *30*(2), 101–117

Taylor, P., Mulvey, G., Hyman, J., & Bain, P. (2002). Work organization, control and the experience of work in call centres. *Work Employment & Society*, *16*(1), 133–150

Taylor, S. (1998). Emotional labour and the new workplace. *Workplaces of the Future.* Springer. 84–103

Teow, M. Y. (2017). Understanding convolutional neural networks using a minimal model for handwritten digit recognition. *2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS).* IEEE

Tirole, J. (1986). Hierarchies and bureaucracies: On the role of collusion in organizations. *Journal of Law Economics & Organization*, *2*(2), 181–214

Tranfield, D., Denyer, D., & Smart, P. J. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207–222

Tsolas, I. E., Charles, V., & Gherman, T. (2020). Supporting better practice benchmarking: A DEA-ANN approach to bank branch performance assessment. *Expert Systems with Applications*, *160*, 113599

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). A Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, *27*(3), 425–478

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment*, *8*(4), 216–226

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356–365

Wang, P., Wagner, T. A., Boyar, S. L., Corman, S. A., & McKinley, R. B. (2016). The Relationship Between Organizational Family Support and Burnout Among Women in the Healthcare Industry: Core Self-Evaluation as Moderator. *Handbook on Well-Being of Working Women* (pp. 283–296). Springer

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D., & Lestantyo, P. (2019). Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data. *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA).* IEEE

Wegge, J., Van Dick, R., Fisher, G. K., Wecking, C., & Moltzen, K. (2006). Work motivation, organisational identification, and well-being in call centre work. *Work & Stress*, *20*(1), 60–83

Willis, S. J., & Bendixen, M. (2007). A Review of Call Center Measurements. Production and Operations Management Society. Available at https://www.poms.org/conferences/cso2007/talks/30.pdf

Wilson, J. P. (2009). *The Call Centre Training Handbook: A Complete Guide to Learning & Development in Contact Centres*. Kogan Page

Wöllmer, M. (2013). *Context-Sensitive Machine Learning for Intelligent Human Behavior Analysis*. München: Universitätsbibliothek der TU

Wooffitt, R. (2005). *Conversation analysis and discourse analysis: A comparative and critical introduction*. Sage

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer science and Technology*, *16*(6), 582–589

Zillner, S., Becker, T., Munné, R., Hussain, K., Rusitschka, S., Lippell, H., Curry, E., & Ojo, A. (2016). Big data-driven innovation in industrial sectors. *New Horizons for a Data-Driven Economy* (pp. 169–178). Cham: Springer