applied sciences

MDPI

Article

Emotion Recognition Method for Call/Contact Centre Systems

Mirosław Płaza [1,*], Robert Kazała [1], Zbigniew Koruba [2], Marcin Kozłowski [1], Małgorzata Lucińska [3], Kamil Sitek [1] and Jarosław Spyrka [1]

[1] Faculty of Electrical Engineering, Automatics Control and Computer Science, Kielce University of Technology, 25-314 Kielce, Poland
[2] Faculty of Mechatronics and Mechanical Engineering, Kielce University of Technology, 25-314 Kielce, Poland
[3] Faculty of Management and Computer Modelling, Kielce University of Technology, 25-314 Kielce, Poland
* Correspondence: m.plaza@tu.kielce.pl; Tel.: +48-41-342-4167

Abstract: Nowadays, one of the important aspects of research on call/contact centre (CC) systems is how to automate their operations. Process automation is influenced by the continuous development in the implementation of virtual assistants. The effectiveness of virtual assistants depends on numerous factors. One of the most important is correctly recognizing the intent of clients conversing with the machine. Recognizing intentions is not an easy process, as often the client's actual intentions can only be correctly identified after considering the client's emotional state. When it comes to human–machine communication, the ability of a virtual assistant to recognize the client's emotional state would greatly improve its effectiveness. This paper proposes a new method for recognizing interlocutors' emotions dedicated directly to contact centre systems. The developed method provides opportunities to determine emotional states in text and voice channels. It provides opportunities to explore both the client's and the agent's emotional states. Information about agents' emotions can be used to build their behavioural profiles, which is also applicable in contact centres. In addition, the paper explored the possibility of emotion assessment based on automatic transcriptions of recordings, which also positively affected emotion recognition performance in the voice channel. The research used actual conversations that took place during the operation of a large, commercial contact centre. The proposed solution makes it possible to recognize the emotions of customers contacting the hotline and agents handling these calls. Using this information in practical applications can increase the efficiency of agents' work, efficiency of bots used in CC and increase customer satisfaction.

Keywords: call centre; contact centre; emotion recognition; chatbot; voicebot; AI

Citation: Płaza, M.; Kazała, R.; Koruba, Z.; Kozłowski, M.; Lucińska, M.; Sitek, K.; Spyrka, J. Emotion Recognition Method for Call/Contact Centre Systems. *Appl. Sci.* **2022**, *12*, 10951. https://doi.org/10.3390/app122110951

Academic Editors: Kuo-Ching Ying, Shih-Wei Lin and Chen-Yang Cheng

Received: 4 October 2022
Accepted: 27 October 2022
Published: 28 October 2022

1. Introduction

In recent years, we have observed a very high interest in call/contact centre (CC) systems [1–3]. There are strong trends in the development of CC systems related to the development of applications of virtual assistants implemented in the form of voicebots and chatbots [4]. The solutions currently available on the market in this area are usually based only on the recognized intent of the client. However, the emotional states occurring during the conversation are also important in terms of being able to correctly identify intentions. It is often only possible to correctly assess intentions after taking into account the emotional context of an utterance. Therefore, an important component of intention recognition methods should be emotion recognition methods. The authors are not aware of solutions for intention recognition methods that incorporate specific types of emotions occurring during conversation in their functionalities. What is known is a sentiment (positive, negative, and neutral affect) analysis solution [5]. In this solution, sentiment determination is possible only for English and does not take into account other, more complex language corpuses, e.g., Polish [6]. In addition, the sentiment is recognized here based only on textual data and does not consider the parameters of the audio signal.

From the point of view of building intelligent bots that automatically conduct CC hotline conversations, the sentiment-sensitive approach is not fully sufficient.

Looking at practical applications from another perspective, knowledge about the emotions of hotline users is also important for supervisors who evaluate both the work of individual agents and the effectiveness of conducted campaigns. For example, when an agent expresses negative emotions, it could result in a direct review and possibly in the call being transferred to another person. In current environments, there is no such functionality and call verification is usually performed randomly, which is neither sufficient nor optimal. Another aspect of using the knowledge of emotions surrounding CC conversations is the ability to build behavioural profiles of both the agent and the client. Identifying behavioural profiles would provide new opportunities for matching interlocutors. Currently, CC solutions that match CC interlocutors are typically based on knowledge of the agent's skills. This approach is called skills-based routing [7]. Knowledge of the behavioural profile that takes into account the emotional characteristics of interlocutors could help improve many of the KPIs (key performance indicators) used as metrics in CC systems.

Considering the above, it should be noted that an important problem from the point of view of the functionality of CC systems is the lack of a target emotion recognition method dedicated directly to these systems, which can be easily integrated with their various functionalities to improve performance. The identified deficiencies in this area were the motivation for the authors to develop an intelligent method for recognizing the emotions of clients and agents communicating in both text and CC voice channels.

The contribution of this paper to the body of knowledge is:

- development of a design methodology for a new emotion recognition method dedicated directly to the contact centre industry, providing the ability to study emotional states in both text channels and voice channels of CC systems;
- analysis of the effectiveness of classification depending on the types of classifiers used and the mechanisms implementing the key tasks of the proposed methodology;
- the proposed hybrid approach is the first comprehensive solution to the problem of automatically recognizing emotions simultaneously in both CC channels, while assuming additional possibilities for using transcriptions of recordings in assessing emotional states.

The developed approach involves the integration of data balancing techniques, vectorization methods, word embedding techniques, a dedicated transcription method, selected speech signal descriptors, a dictionary of emotional expressions and various classification methods—thus creating a new emotion recognition system for CC applications. The necessity of using the above mentioned methods is dictated by the nature of the processed data, which for the real problem under consideration is usually unstructured. Experimental results showed that the approach proposed in the paper provides the quality of classification of emotional states in text and audio channels at a level satisfactory from the point of view of practical applications.

The subsequent part of the paper is organized as follows. Section 2 analyses the current state of the art in emotion recognition in text and voice data. Section 3 describes the methodology of the research work conducted. Section 4 presents methods and techniques used in the context of improving correct emotion recognition. Section 5 describes a new emotion recognition method proposed by the authors intended specifically for CC applications. The proposed method determines the emotional states of clients and agents in text and voice channels. Section 6 describes the results obtained, while Section 7 presents conclusions and directions for further research.

2. Related Work

With the steady popularization of conversation-based services, automatic recognition of human emotions using computer systems has become an important research trend. This is evidenced by the large number of review papers describing the most interesting systems for automatic recognition of emotions in voice data [8–10] and automatic recognition of

emotions in text (usually comments left by users in various web applications) [11–14]. These solutions are often systematized according to the emotion model adopted, the feature extraction method, and the classification algorithm used. To implement the emotion recognition system, emotion models must be defined and specified. So far, however, there have been difficulties in defining the term emotion, which is still an open problem in psychology. An attempt to solve it, in terms of the problem of CC systems, was made through research [15]. Two types of emotion models—categorical and dimensional—are the most commonly used. The categorical model classifies emotions as a fixed number of discrete categories. In contrast, the dimensional model describes an emotion as a point in a continuous multidimensional space. Many works use the Ekman model [16] or its derivatives.

Recognition of emotions in text is often used in systems that allow analysis of comments left by users. These systems mostly boil down to sentiment analysis, which, as mentioned, involves assigning one of three categories to a given entry: positive affect, negative affect, and neutral attitude. The study of sentiment allows for a simplified classification that does not assume the presence of different types of emotions, but only a greatly simplified model. For many applications focused on assessing user satisfaction or evaluating user preferences, it is perfectly adequate [17]. There are known applications where a further reduction of the problem into positive and negative categories only is made without separating the neutral opinion [18]. Sentiment analysis has been repeatedly performed on data extracted from platforms such as Twitter [19,20] and Facebook [21]. Well known methods used in sentiment analysis include deep learning, in particular using convolutional neural networks (CNN) [22] and networks using long short term memory (LSTM) [23]. Popular solutions also include works using both CNN and LSTM [24]. The paper [19] uses team learning methods to assess sentiment. The paper [14] uses a fuzzy system to assess sentiment. The combination of fuzzy system along with neural network-based models is presented in [12]. More specific methods include Spider Monkey hybrid optimization [25] or models based on rule systems [26]. An approach based on unsupervised learning can be found in paper [27]. Yet all of the above solutions apply only to sentiment assessment. However, in many applications requiring assessment of the user's emotional state, the use of sentiment analysis is not sufficient. The paper [19] proposes to extend the sentiment model to five states (strong positive, positive, neutral, negative, strong negative). The paper [26] distinguishes seven emotional states (fun, happiness, love, neutral, anger, hate, sadness), while the paper [28] distinguishes 27 different emotional states. In order to classify sentiment or emotional states, it is necessary to transform text into features suitable for use in machine learning models. Among the most popular methods are word embedding techniques such as BERT (bidirectional encoder representations from transformers.) [29] or GloVe (global vectors for word representation) [19]. The use of POS (part-of-speech tagging) has been proposed in papers [18,19,21]. The paper [22] uses the bag-of-words technique. Term frequency and inverse document frequency techniques were used in [19,30]. The first part of Table 1 summarizes selected solutions for emotion recognition in text.

In designing an emotion recognition system based on speech signal, one of the more challenging tasks is to identify and extract various emotion-related speech features. Due to the complexity of representing a speech signal in the time and frequency domains, parameters are sought to reflect a quantitative description of the signal. The most commonly used group of descriptors to describe speech are frequency and spectral characteristics, including the fundamental frequency characteristics of laryngeal tone and formant frequencies [31]. Widely used are coefficients using a perceptual approach that mimics the speech signal recognition mechanisms found in humans—mel-frequency cepstral coefficients (MFCC). This is a spectrum approximation method that is a special case of mel generalized cepstral analysis [32]. The final step of the emotion recognition task is the classification process. This process determines the affiliation of an unknown object, described by an attribute vector, to one of the predefined classes. Classifying emotions based on speech is a difficult task, depending on many factors. One of the main problems of automatic emotion recognition in

speech is collecting the right data to train the model. Based on how emotional speech is generated, speech datasets are categorized into three groups: acted, evoked, and natural [33]. Speech emotion recognition systems that achieve high accuracy for actor datasets do not perform as well with realistic natural speech. Acting speech is the exaggeration of emotions, which is not characteristic of spontaneous speech. While emotion recognition in speech has been an active research area for the last three decades, techniques that deal with the natural environment have only emerged in the last decade. The difficulty of recognizing emotions is much greater in natural contexts with uncontrolled conditions. Specific aspects related to emotion recognition in natural settings are presented in the review paper [34]. The authors note the small number of datasets containing recordings taken in natural contexts. Moreover, the problem of natural databases is the imbalance of examples. An imbalance exists with respect to class, length of speech. In natural settings, long utterances are common. Detecting emotions in such utterances is much more difficult than in short statements. In long utterances, emotion may only be present in a part thereof. Therefore, there is a need to identify emotional regions in utterances. Training and testing should only be done on emotional regions. Another problem with emotion recognition in natural contexts is related to a speaker. Systems for emotion recognition in speech signals perform much better in the case of a small, fixed set of speakers compared to conditions where the number of speakers is large. In the first situation, the system is trained and tested using recordings of the same speakers. However, in natural settings, it is easy to encounter a new test speaker who has not previously participated in the system training process. When training and testing with different speakers, system performance usually drops significantly. This is because of anatomical and morphological differences in the geometry of the vocal cords of different speakers. In systems that recognize emotion in conversations conducted in natural contexts, the variability of environment-related conditions, namely noise, codec, and microphone distance must be taken into account. Additional difficulties arise for emotion recognition during hotline calls, where the quality of the audio signal is poor. The audio signals typically found in CC systems are extremely varied and often distorted. Such a signal usually has a sampling rate as low as 8 kHz, 16-bit resolution and PCM (pulse code modulation) format. In addition, calls are often carried out using VoIP (voice over Internet Protocol) technology, usually recorded in mono format without separate channels for the client and agent, so that the statements of interlocutors overlap. These channels are usually characterized by varying parameters, e.g., varying amplitude, or varying background noise level around the interlocutors [35]. When it comes to emotion recognition in audio files, some solutions based on the following classifiers are known: SVM (support vector machine), HMM (hidden Markov model), GMM (Gaussian mixture model), kNN (k-nearest neighbour), ANN (artificial neural network), DT (decision trees) and fuzzy classifiers. Recently, solutions based on DNN (deep neural networks), including CNN (convolutional neural networks), have become most popular. The second part of Table 1 presents the characteristics of selected emotion recognition solutions for audio conversations.

Table 1. Analysis of selected solutions for emotion classification methods in text and audio data.

No	Solution	Proposed Approach	Results	Ref.
		Text channel		
1.	Sentiment analysis of Twitter data	SVM model using the following descriptors: POS-tagging, tree kernel, senti features.	A validation level of about 56–60% was achieved, the model recognizes three groups of emotions: positive, negative, and neutral.	[18]
2.	MTC problem solution	The kNN algorithm combined with profile classifiers with improved performance.	A validation rate of approximately 67% was achieved. The model considers the relevance of the documents to the case, which are divided into two categories: positive and negative.	[36]
3.	Sentiment classification based on SVM and ANN	SVM and ANN models using the TF/IDF descriptor.	For ANN, the validation level is about 76–90% (depending on the training dataset). For the SVM classifier, the validation level is about 74–89%. Models recognize emotions: positive and negative.	[37]
4.	Twitter sentiment analysis based on ordinal regression	The following network types were tested: SVR, DT, RFC. The TF/IDF descriptor was used.	For DT, the validation level is about 92%, for RFC about 83% and for SVR–82%. The models listed above recognize positive and negative emotions.	[30]
5.	Sentiment representation model for Twitter analysis	The following network types were tested: SVM, RFC, CNN, and LSTM, and the Bayes classifier NBC. The QSR descriptor was used.	For the LSTM and CNN networks, the validation level is about 78%, the other solutions had validation levels in the range of about 64–65%. The models listed above recognize emotions: positive, negative, and neutral.	[38]
6.	Twitter text analysis using NBC, kNN and SVM	The following classifiers were tested: NBC, kNN, and SVM. They use the TF/IDF descriptor.	For the NBC classifier, results of about 63% were achieved, for the SVM classifier about 61% and kNN about 60%. The models classify: openness, conscientiousness, extraversion, agreeableness, neuroticism.	[39]
7.	Multi modal dialog act classification	DNN based model using GloVe vectors as a descriptor.	A validation rate of 39–65% was achieved depending on the training dataset used. The model recognizes the following 8 emotions: sadness, anger, fear, happiness, surprise, disgust, frustration, neutrality.	[40]
8.	Attention based word embeddings	The ABC algorithm with SVM classifier was used with the objective of maximizing classification accuracy. ATVs were used as descriptors.	A validation rate of 82–96% was achieved depending on the training dataset used. The model recognizes positive, negative, and neutral emotions.	[41]
		Voice channel		
1.	Emotion detection through speech	CNN network was tested. MFCC were used as descriptors.	For the CNN model, the best performance was achieved (at about 83%), the model recognizes the following types of emotions: anger, disgust, fear, happiness, sadness, and surprise	[42]
2.	Speech emotion recognition based on rough set and SVM	The training data for the SVM model is a total of 13 parameters, including three from the energy features group, five pitch features, four formant features, and speech rate.	A validation level of approximately 78% was achieved, the model recognizes the following 5 emotions: anger, happiness, sadness, fear, surprise. There is variation by gender.	[43]
3.	Speech emotion recognition for SROL database using weighted kNN algorithm	18 parameters were used as training data for kNN model: F0, standard deviation of F0, median fundamental frequency, mean 1–4 formant frequency, standard deviation of 1–4 formant frequency, median 1–4 formant frequency, local jitter, local absolute jitter, local shimmer.	A validation level of approximately 63% was achieved, the model recognizes the following three types of emotions: anger, happiness, sadness. There is variation by gender.	[44]
4.	A hierarchical framework for speech emotion recognition	The following network types were tested: PCA and LDA. 64 parameters were used as training data: 48 prosodic features and 16 formant frequency features.	Comparable results were obtained for the PCA and LDA models. The average recognition rate for males was 83.4% and 78.7% for females, the model recognizes the following five types of emotions: anger, happiness, sadness, fear, surprise.	[45]
		Fusion of text and voice methods		
1.	Deep neural networks for emotion recognition	An LSTM network and the average fundamental frequency, shimmer, jitter and MFCC descriptors were used to detect emotion based on acoustic features. In contrast, a multilayer CNN network was used to detect emotions in transcriptions.	A validation level of 65% was achieved, the fusion of the models recognizes: anger, joy, sadness, and neutrality.	[46]
2.	Emotion recognition system using speech features and transcriptions	A CNN network and MFCC spectrograms and parameters were used to detect emotions based on acoustic features. In terms of emotion classification from text, the CNN network was also used using the word2vec method.	A validation level of 76% was achieved, CNN model fusion recognizes: anger, joy, sadness, and neutrality.	[47]

Table 1. *Cont.*

No	Solution	Proposed Approach	Results	Ref.
3.	Emotion recognition based on bottleneck acoustic features and lexical features	A DNN network and the F0, MFCC, 40 mel filterbank energies parameters descriptors were used to detect emotions based on acoustic features. In terms of emotion classification from text, the DNN network was also used with word2vec and ANEW descriptors.	A validation level of 74% was achieved, DNN model fusion recognizes: anger, joy, sadness, and neutrality.	[48]
4.	Multimodal emotion recognition network with personalized attention profile	For emotion detection based on acoustic features, a BLSTM network and 45-dimensional vector consisting of MFCC features, F0, zero cross ratio along with the first and second derivatives of MFCC were used. Meanwhile, BLSTM using the GloVe method was also used to detect emotion in text.	A validation level of 70% was achieved, BLSTM model fusion recognizes: anger, joy, sadness, and neutrality.	[49]
5.	Speech emotion recognition based on attention weight correction using word level measure	The BLSTM network and the following descriptors were used to detect emotions based on acoustic features: MFCC, constant transform and average fundamental frequency. In terms of emotion classification based on text, the BLSTM network was also used with the BERT model.	A validation level of about 76% was achieved. The following emotions are recognized: anger, joy, sadness, and neutrality.	[50]

Ref.—references; TF/IDF—term frequency inverse document frequency; SVR—support vector regression; QSR—quantum-inspired sentiment representation; ATV—attention vector; PCA—principal component analysis; LDA—linear discriminant analysis; ANEW—affective norms for English words.

There are also known solutions that use a combination of methods used in the study of acoustic features and features extractable from textual records (shorthand) to determine selected types of emotions present in an audio recording [46–50]. In this case, in addition to the analysis of the sound, the text derived from the transcript is also studied. Examples of such solutions are shown in the third part of Table 1. Their mode of operation is usually a combination of methods used in audio material with methods used in textual material [51]. Considering the above discussion and the information in Table 1, the following classifiers were selected for the testing processes of the developed emotion recognition method dedicated specifically to the CC industry: (a) for the text channel: ANN, RFC (random forest classification), DT, kNN, SVM, ABC (artificial bee colony) and NBC (naive Bayes classifier) (b) for the voice channel: CNN, KNN, SVM and LDA.

3. Research Methodology

The main research work involved developing algorithms to recognize the emotions of clients contacting the CC hotline and agents handling incoming calls/messages. Separate emotion recognition for customers and agents is necessary from the point of view of practical applications of the proposed method in CC systems. Knowledge of agent emotions is needed, for example, for supervisors evaluating their work and monitoring selected calls. Emotional states of customers, in turn, are important, for example, in the aspect of increasing the effectiveness of bots, as explained in detail in the paper [15]. In this case, the ability to build a conversation flow tailored to the emotional states of the customer improves the quality of bot performance. Algorithms were developed in the form of relevant components included in the method proposed in this paper, which is described in detail in Section 5. The research work was divided into two main parts. Part one is a study of the emotion recognition component in the text channels of the CC system, while part two is a study of the emotion recognition component in the voice channels. In terms of emotion recognition algorithms based on voice conversations, multicriteria classifiers that consider the emotional states identified in the transcripts of these conversations were also additionally used. For this purpose, the classifiers developed in the first part of the research and a dedicated dictionary of emotional words [52] prepared for Polish language were used. The text transcriptions were derived from automatic transcriptions made for the individual fragments of the audio recordings under study. Transcriptions were made using the method described in paper [53]. Thus, the component that determines emotion from text data was ultimately used to support emotion classification in voice channels.

The purpose of the above papers was to determine the potential use of the developed algorithms as components of the emotion recognition method proposed in the paper. The illustrative block diagram of the test methodology is presented in Figure 1.

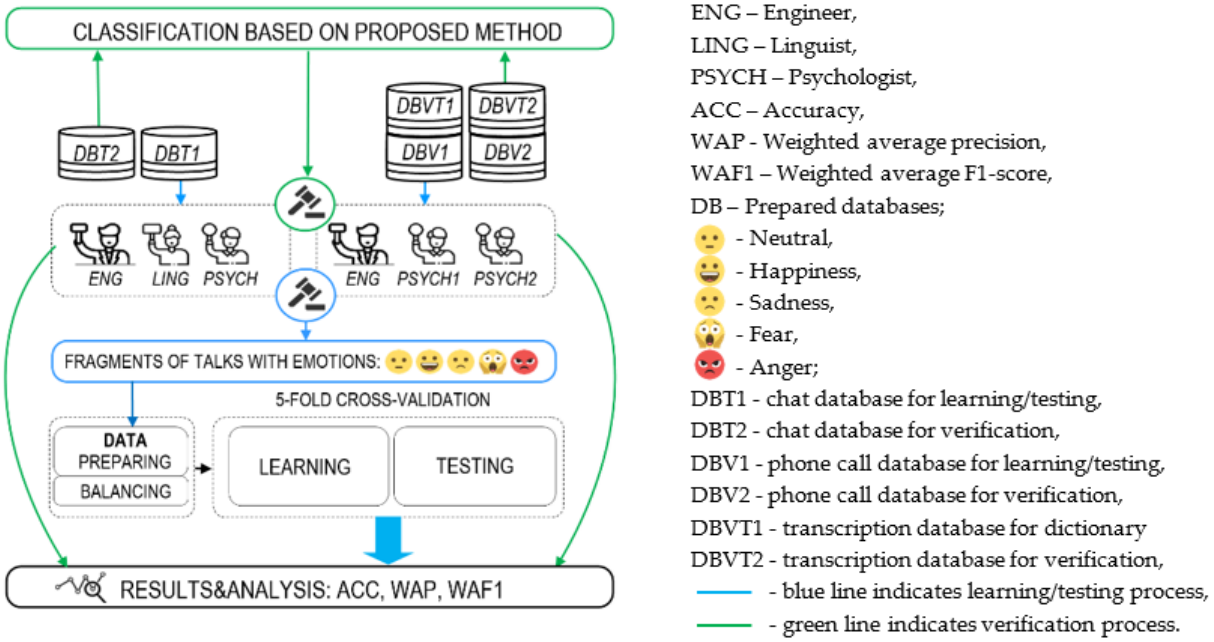

Figure 1. Research methodology.

In the initial phase of the research work, databases were created for text conversations and audio recordings. All data represented actual client–agent conversations in a large contact centre system operating on a commercial basis. The prepared databases are marked in the subsequent part of the paper as: DBT1, DBT2, DBV1, DBV2 as well as DBVT1 and DBVT2. A detailed analysis for the described datasets is presented in the paper [54], while Table 2 summarizes their most salient features.

Table 2. Used databases.

Name	Description	Application
DBT1	345 chat text conversations (7515 statements of clients and agents were extracted).	Data used in the process of training and testing of classifiers in part one.
DBT2	100 chat text conversations (3718 statements of clients and agents were extracted).	Database used in the verification process of the developed method.
DBV1	302 actual voice calls ranging from 3 to 20 min in duration. The total recording time of this database is 22 h 59 min and 11 s.	Data used in the process of training and testing of classifiers in part two.
DBV2	100 actual voice calls ranging from 3 to 20 min in duration. The total recording time of this database is 7 h 25 min and 15 s.	Database used in the verification process of the developed method.
DBVT1	Database of transcriptions of recordings from DBV1 set.	Additional functionality to improve recognition for the problem in part two.
DBVT2	Database of transcriptions of recordings from DBV2 set.	Database used in the verification process of the developed method.

The data collected in the DBT1 and DBV1/DBVT1 databases were further evaluated by experts who extracted emotionally charged fragments from individual conversations. Three independent judges participated in assessing the emotional states present in the collected databases in each part of the work. The evaluation of the text data was performed by a linguist, a psychologist working on emotion issues, and an engineer designing algorithms to detect emotions in text channels. In contrast, the evaluation of the audio files was performed by two psychologists working on emotion issues and an engineer designing algorithms

to detect emotions in voice channels. The judges independently evaluated the extracted data and, as a result of the final discussion, determined a single outcome score for each statement. Finally, the pool of audio and text files (used in the training and testing process) was joined by those for which there was 100% agreement in the judges' evaluation. Thanks to the evaluations obtained and taking into account the opinions of people with experience in the operation, functioning and implementation of CC systems, it was possible to propose the Emotion Classification for Machine Detection of Affect-Tinged Conversational Contents in voice and text channels of contact centre systems [54]. Accordingly, classes of basic emotions (neutral, happiness, sadness, fear, anger) were defined, corresponding to families of related emotions identified in the databases: DBT1 and DBV1/DBVT1. The emotion classes thus defined are recognized by the solution proposed in this paper. The problem is therefore multiclass, with five classes.

In the next phase of the work, sets for the training and testing processes were created from the conversation fragments selected in the data preparation process for the DBT1 and DBV1 databases. In terms of chat conversations (DBT1 database), the original set consisted of 7515 utterances, of which 3766 items were removed as a result of the evaluation procedure described above. The deleted items were 463 utterances containing text from the anonymization process and 3303 utterances that were not explicitly marked by the judges. Thus, the remaining 3749 utterances were used as the data for the training and testing processes. From these utterances, a total of the following emotion classes were selected: neutral = 2269, happiness = 761, anger = 312, sadness = 305, and fear = 102. In terms of voice conversations (DBV1 database), the judges convergently marked a total of 400 different passages in which specific emotion types were present. As the durations of the utterances marked by the judges were of variable length, the long passages were divided into several parts, which ultimately allowed the construction of a set consisting of 2935 elements. The class sizes were as follows: neutral = 1491, anger = 937, happiness = 196, fear = 166, sadness = 145.

The created datasets are characterized by relatively large discrepancies in sample sizes across classes. The problem of quantitatively unbalanced data in the training set can significantly affect the training processes, which in turn is later reflected in the classification results. Therefore, in the next step, the feasibility of using appropriate data balancing techniques was tested and their impact on the obtained classification results was investigated, as detailed in Section 4.1. In order to verify the prepared models, a 5-fold cross validation procedure was performed, for which the prepared datasets (text and voice) were randomly divided into five parts with similar numbers. Then, in five consecutive steps, one of the sets was selected as the testing set (20% of the data), while the other four were the training set (80% of the data). The same research methodology applied to text data and voice data.

Precision and F1-score metrics were used to evaluate the classification models proposed in this paper. These metrics can be referred to in three ways depending on the problem, namely, macro-averaged, micro-averaged or weighted-averaged. From the point of view of the target applications of the method proposed in this paper, as well as due to the high degree of variation in the amount of data in each class, it was decided to use weighted-averaged metrics. In the subsequent part of the paper, they are labelled as: WAP (Weighted Average Precision), WAF1 (Weighted Average F1-score). This ensures that the results obtained for each class also take into account its size in the actual data sample. This, in turn, influences a much more reliable evaluation of the models prepared for the classification problem in question. Additionally, the Accuracy metric labelled ACC was also examined. It shows the accuracy, which is defined as the averaged percentage of correctly classified cases (sum of diagonal values in the confusion matrix) over all cases. For weighted-averaged metrics, ACC values converge to weighted-averaged Recall values, therefore the Recall metric is not separately analysed in this paper. The metrics used in the verification process of the proposed method are integrated in the scikitlearn library [55,56].

The above metrics applied separately do not offer a complete picture of the classifier performance. Moreover, for unbalanced sets, some may even give an impression of misclas-

sification. Therefore, it is important to analyse all of the metrics listed above for the purpose of evaluating the proposed solution. This will allow for avoidance of potential controversy and provision of a complete picture of classification performance. DBT2 and DBV2/DBVT2 databases were used for the final verification of the developed method. These data were not used in the training process, they were only used to validate the method. For this purpose, emotional states were recognized by the classifier in the first phase, while in the second phase the results returned by the system were evaluated by the judges. The effectiveness of the method was evaluated according to the procedure described above for evaluating the data by the judges using the proposed metrics.

4. Methods

For the purpose of building the emotion recognition method proposed in this paper, dedicated specifically to the call/contact centre industry, it is advisable to include techniques related to dataset balancing. In addition, in terms of text data, appropriate vectorization methods, the occurrence of special characters in the form of emoticons, and word embedding techniques should be considered. In terms of voice data, on the other hand, proper selection of speech signal descriptors is important. A selection of appropriate classification methods suitable for both text and voice data should also be made. This section of the paper describes the various solutions that are the basic building blocks of the method proposed in this paper.

4.1. Data Balancing Techniques

The problem of unbalanced data in classification processes arises when the sample prepared for research is characterized by large discrepancies in abundance between classes. This situation occurs in the actual CC conversations analysed in the paper. In the prepared sample for the text channel, 61% of the data belong to the NEUTRAL class, 20% to the HAPPINESS class, 8% each to the ANGER and SADNESS classes, and only 3% to the FEAR class. For the voice channel on the other hand: 51% are NEUTRAL class, 31%—ANGER, 7%—HAPPINESS, 6%—FEAR and 5%—SADNESS. This is a natural and expected situation, as neutrality should prevail in CC conversations, while emotional conversations are much less frequent. However, with a random selection of training sets for such a large degree of imbalance, the decision function could favour classes with more samples, usually referred to as majority classes [57]. This is because in multiclass classification problems, most algorithms usually perform optimally when the sample sizes for each class are comparable [58].

One way to mitigate this problem is to be able to randomly extract sufficiently numerous subsets of the data from each class, which provides the required and appropriately proportional representation of all classes in the training process [59]. Such activities were conducted for the research, where 80% of the data were randomly selected from each class, which constituted the training data. Another approach in this area, in turn, is the ability to generate new samples in classes that are underrepresented [60]. This was also done by dividing longer portions of the recordings into smaller pieces in the dataset marked by the judges. The next methods are a combination of the approaches outlined above. Extensive ensemble classifiers using samplers internally are also used [61]. In the research, the Python imbalanced learn library [62] was used for data balancing, which allowed us to determine the applicability of different balancing methods for the problems discussed. The analysed methods are summarized in Table 3.

Table 3. Studied balancing methods.

Method Name	Type	Used Algorithm
RandomOS	OS	Random over-sampler
SMOTE	OS	Synthetic minority over-sampling technique
ADASYN	OS	Adaptive synthetic sampling
RandomUS	US	Random under-sampler
Near-Miss	US	Near-miss method based under-sampler
CondensedNN	US	Condensed nearest neighbor

OS—oversampling, US—undersampling.

Table 4 summarizes the values that determine the effect of the balancing methods on the correctness of the recognized emotions as a result of the classification with the SVM algorithm in the voice channel.

Table 4. The impact of balancing methods on the number of recognized emotions for SVM classifier.

Method	Number of Recognized Emotions [pcs]				
	Anger	**Sadness**	**Happiness**	**Fear**	**Neutral**
Labeled emotions	177	30	38	30	312
Recognition without using balancing techniques					
No balancing	125	1	1	1	283
Recognition with using balancing techniques					
RandomOS	139	13	15	17	177
SMOTE	139	13	15	12	192
ADASYN	142	15	14	12	185
RandomUS	119	22	14	13	124
NearMiss	78	15	13	16	62
CondensedNN	102	11	1	0	287

In the case of emotion recognition for CC systems, it is important that the recognition performance in classes specifying a particular type of emotion is as high as possible. When analysing Table 4, it can be seen that when the balancing methods are not used, the emotions that have less representation (SADNESS, HAPPINESS, FEAR) are hardly recognized. Similar results were obtained for the CondensedNN method. This situation is very unfavourable for the possibility of further applications of the proposed emotion recognition method in CC. Relatively good results in this regard were obtained from balancing methods RANDOMOS, SMOTE and ADASYN. In this case, the SMOTE method is attractive because, as can be seen, in addition to recognizing minority class emotions, it also has the largest number of NEUTRAL states recognized. This has a positive impact on the final classification results.

This is also supported by the data in Table 5, where the results of the research described by the metrics used in the evaluations of the classification models are presented. Analysing the results of Tables 4 and 5, it can be seen that better results for the discussed problem are obtained using methods from the oversampling group. In this regard, the SMOTE method proved to be the most optimal. This method was used further in the process of optimizing the training data.

Table 5. Classification results using the SVM algorithm.

Method	ACC	WAP	WAF1
RandomOS	0.61	0.69	0.64
SMOTE	0.63	0.68	0.65
ADASYN	0.63	0.68	0.64
RandomUS	0.50	0.64	0.53
NearMiss	0.31	0.58	0.35
CondensedNN	0.68	0.68	0.63

4.2. Techniques Used in Text Channel

Vectorization methods can be considered in the context of how to improve emotion recognition in text channels. In neural network classification, it is possible to use a complete utterance as the input of the classifier. However, for simpler classifiers, additional techniques are needed to prepare the input data adequately. For this purpose, a technique based on word weighting was used [63]. The method allows for preparation of statistics on the occurrence of individual words in the whole utterance. With this technique, it is possible to determine how often a word occurs in a given utterance. It also provides opportunities to identify the words that have the most meaning in an utterance. This is accomplished by assessing the frequency of specific words across all utterances.

The method described is shown in Figure 2. The first step performed is to evaluate the number of occurrences of each word in the utterance, denoted as TF. If certain words are repeated in a given utterance then ultimately these words may have a greater impact on the classification score. The second step is to calculate the IDF value. This factor is calculated according to the following formula [64,65]:

$$\mathrm{IDF}(w) = \ln\left(\frac{1+\mathrm{N}}{1+\mathrm{D}(w)}\right) + 1 \tag{1}$$

where: w—given word; N—number of all chats; D(w)—number of chats containing word w.

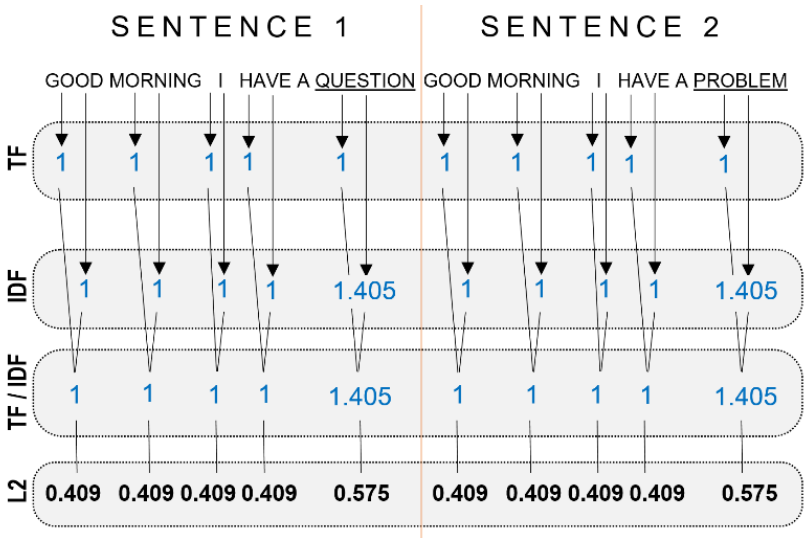

Figure 2. The vectorization technique used in the proposed approach. LEGEND: TF—term frequency; IDF—inverse document frequency; L2—normalization.

The use of IDF allows higher weights to be assigned to words that are most likely to be important in the classification process. On the other hand, popular words that appear in most chats will have low weights. To calculate the final word weight, both TF technique and IDF technique were used according to the formula [64,65]:

$$\mathrm{TF|IDF}(w) = \mathrm{TF}(w) \cdot \mathrm{IDF}(w) \tag{2}$$

The final step is to apply L2 normalization according to the formula [66]:

$$\mathrm{L2}(\mathrm{TF}|\mathrm{IDF}(w_i)) = \frac{\mathrm{TF}|\mathrm{IDF}(w_i)}{\sqrt{\mathrm{TF}\Big|\mathrm{IDF}(w_1)^2 + \ldots + \mathrm{TF}\Big|\mathrm{IDF}(w_k)^2}} \quad (3)$$

where: w_i—the *i-th* word in the utterance; k—the number of words in the utterance.

Emoticons constitute another element having a significant impact on the effectiveness of emotion recognition in CC text channels. In recent years, they have been very common and increasingly used during conversations in text channels. These types of signs are helpful in determining the emotions expressed by interlocutors, so it is essential to include them in the automatic emotion recognition method being developed. The assignment of emoticons identified in CC text channel conversations to specific classes defining emotions was made according to the assumptions developed in the paper [54].

Table 6 lists the groups of emoticons that are associated with a particular type of emotion. These signs were also included in the developed algorithms, which improved the efficiency of correct recognition of emotional states. Most of the icons presented were also identified by the judges in the process of evaluating the databases used for the purposes of this paper. The majority of identified signs were those associated with the emotion HAPPINESS, which is probably due to the conversation topics chosen for the study. This class is also one of the most numerous. However, if the topic of conversation was, for example, related to the area of debt collection then it can be assumed that the number of emoticons identifying the types of negative emotions would be much greater. Nevertheless, all studied emotion types were represented by emoticons, which also confirms the validity of including this element in the designed solution.

Table 6. Emotions in the CC text channel.

Emotion	Emoticon Type	Number of Emoticons
ANGER	>:(, >:-(, :P, :-P, :/:-/	11
FEAR	:O, :-O, :o, :-o	8
SADNESS	:(, :-(, :((, :-((, :’(, :’-(	13
HAPPINESS	:), :-), :)), :-)), :D, xD, XD	111
NEUTRAL	Not applicable	

Another approach to be considered in the method presented in this paper is word embeddings techniques. For the purpose of classifying text data using standard classifiers, it is necessary to convert individual input words into numerical form. In the simplest variant, this can be implemented using one hot vectorization coding. However, the major drawback of such simple solutions is the inability to take into account the semantic meaning of individual words. Codes assigned to words that occur in a similar context or are synonyms usually have random numeric values. Therefore, the word embedding method should be used to account for the semantic meaning of individual words. In this method, individual words are encoded using n-dimensional vectors. Using this technique allows for an effect where vectors representing collocated words are closer together than words with a completely different meaning. For example, the words “good” and “morning” will have collocated representations, while the words “good” and “red” will be further apart.

There are many well-learned word embedding models trained on sufficiently large corpora. Unfortunately, most of them are only available for English. The few examples available for the Polish language are usually trained using publicly available datasets (e.g., Wikipedia entries). However, these data are devoid of emotionally charged words, which make them unsuitable for use in the method proposed in this paper. Table 7 shows examples of words, the trained vectors of which are close to each other, indicating that the context of occurrence of these words was collocated in the studied corpus.

Table 7. Collocations of words.

Word	Group of Words
wish	nice, hours, day
works	slow, messenger, mostly
unfortunately	my, telephone, allows, important
very	Mr, thank
where	information, stated, important
please	prompt, find, can
help	can, somehow, something
unacceptable	disconnected, can, number
client	service, 24, disconnected, headquarters
case	weeks, last, 2
better	contact, attention, was
want	by, given, indicated
waiting	once, help, case
favour	short, window, mark
best	wishes, end, day, nice

In order to apply the embedding model that takes into account emotionally charged words, in particular to meet the requirements of CC systems, it was necessary to train one's own embedding model. For this purpose, the DBT1 database was used. One hundred-dimensional vectors were used, trained using the continuous-bag-of-words (CBOW) model [66]. This model uses windows of five words occurring next to each other, forming the context of an utterance. The CBOW model can predict the middle word based on the other words from the context. This method allowed for training a model that closely matched the vocabulary appearing in the actual CC text conversations studied. Collocations of words occurring in the same context were taken into account.

4.3. Speech Signal Descriptors

Different types of descriptors can be used when considering the issue of recognizing emotions occurring in audio conversations. They usually belong to one of three groups, namely: temporal, frequency, or cepstral descriptors. When examining a speech signal in the time domain, it can be parameterized using classical statistical parameters such as mean, variance, and standard deviation. Sometimes, signal energy, power, and RMS (root mean square) values are also determined [67]. In speech analysis systems, the average fundamental frequency denoted as F0, and the jitter and shimmer coefficients are also determined [68]. For the frequency representation of a speech signal, the parameters of the first and second formant frequency (formants F1 and F2) can be considered. Increased formant values may indicate a signal with a strong energy charge (e.g., a raised tone of voice). The audio signal can also be parameterized with different variants, cepstral coefficients. The most commonly used coefficients in the literature on emotion recognition problems [69,70] are: mel-frequency cepstral coefficients (MFCC), linear predictive cepstral coefficients (LPCC), and linear predictive coding (LPC). These coefficients reflect the acoustic nature of the signal.

The DBV1 database, described in detail in Section 2, was used to assess the relevance of each descriptor to the works presented in this paper. Based on the analyses performed, the following speech signal descriptors were finally selected: fundamental frequency F0, MFCC, jitter, and shimmer coefficients. As a result of literature analysis [71–75], it was determined that the first 13 MFCC coefficients describing the frequency parameters of the speech signal would be used for emotion recognition. These features are extended

to include the jitter parameter that determines changes in voice frequency during speech and the shimmer parameter that determines changes in amplitude. Pearson's correlation coefficients between individual characteristics and emotional states were determined. For this purpose, for each example in the dataset, a value of 1 was taken if it corresponded to the selected emotional state and a value of 0 when the emotional state differed from the selected one.

The values of the relevant Pearson correlation coefficients are summarized in Table 8. A significant linear relationship can be observed between the emotional state ANGER and the descriptors MFCC0, MFCC2, MFCC5 and MFCC8 (absolute value of Pearson's coefficient > 0.3). The SADNESS emotional state is moderately inversely correlated with the MFCC0 descriptor. In other cases, more complex, nonlinear relationships between traits and emotional states can be inferred from small Pearson coefficient values. Therefore, advanced classification tools, such as convolutional neural networks, should be used.

Table 8. Values of Pearson's correlation coefficients.

MFCC	Anger	Sadness	Happiness	Fear	Neutral
F0	−0.070	−0.002	0.118	0.035	−0.081
JITTER	0.157	−0.132	0.075	−0.087	−0.013
SHIMMER	0.289	−0.178	−0.001	−0.093	−0.016
MFCC0	0.335	−0.290	0.003	0.079	−0.128
MFCC1	0.218	0.117	−0.144	0.036	−0.227
MFCC2	−0.304	0.061	0.065	0.133	0.044
MFCC3	0.151	−0.089	−0.006	−0.014	−0.042
MFCC4	0.241	−0.021	−0.059	−0.083	−0.078
MFCC5	−0.338	0.110	0.054	0.024	0.150
MFCC6	−0.154	0.029	0.011	−0.034	0.148
MFCC7	0.002	−0.103	0.083	−0.106	0.123
MFCC8	−0.324	0.027	0.150	−0.033	0.181
MFCC9	−0.182	0.090	0.019	−0.024	0.096
MFCC10	−0.011	−0.108	0.028	0.016	0.075
MFCC11	−0.178	−0.135	0.151	−0.020	0.182
MFCC12	−0.000	−0.013	0.073	−0.146	0.086

5. Proposed Approach

Figures 3 and 4 propose a new approach to solve the problem of recognizing client and agent emotions in text and voice channels of CC systems. The proposed method has been integrated with various functional blocks used in conversation analysis to form a fully complementary hybrid environment. We can distinguish three basic components to the solution: TRANSCRIPTOR (a transcription module for calls originating from CC voice channels), TEXT ANALYZER and VOICE ANALYZER. The data from the text channel (CHAT TALKS) is routed directly to the TEXT ANALYZER module for further processing. In turn, the data coming from the voice channel (VOICE TALKS) goes simultaneously to the modules: VOICE ANALYZER and TRANSCRIPTOR. The TRANSCRIPTOR MODULE performs automatic transcription of conducted conversations according to the method described in paper [53]. Conversation transcripts are directed to the TEXT ANALYZER module, while audio signals go to the module that is designed for audio analysis.

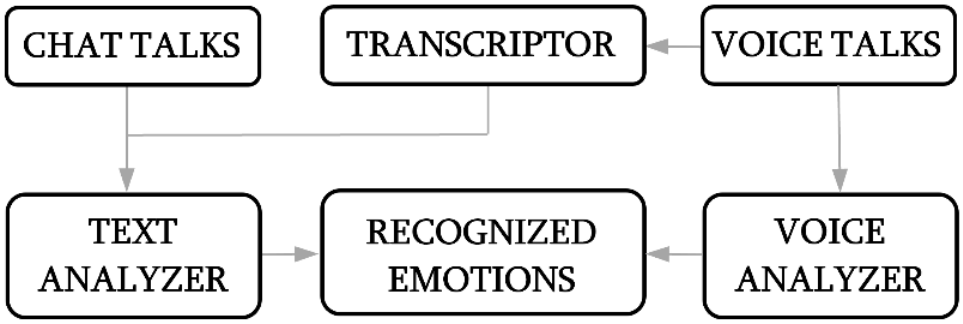

Figure 3. High level flowchart for emotion recognition method in text and voice channels.

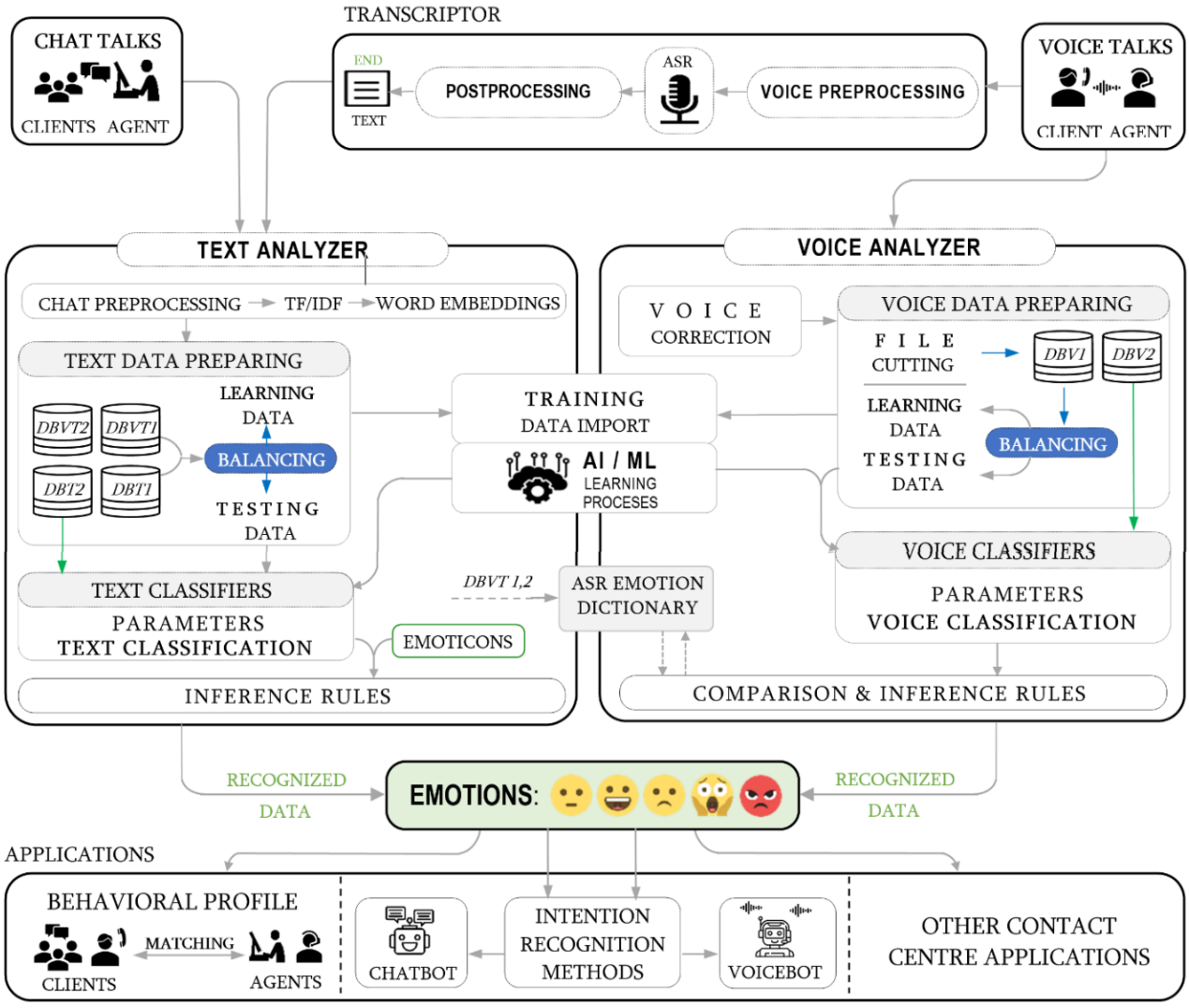

Figure 4. Method for emotion recognition in text and voice channels of CC systems.

The modules that analyse CHAT TALKS and VOICE TALKS provide the results of classifying emotional states according to the emotion classification for machine detection of affect-tinged conversational contents in voice and text channels of contact centre systems. A high level flowchart of our method is presented in Figure 3, with a detailed conception presented in Figure 4.

5.1. Transcriptor Module

This module uses a voice call transcription method dedicated directly to CC systems. This method allows for automatic transcription with an efficiency of 90–92%. It uses appropriate audio signal preprocessing algorithms and postprocessing algorithms that process the originally performed transcriptions. In terms of preprocessing, there are three modules: channels separation module, training of the ASR (automatic speech recognition) systems module, and audio signal correction module. In terms of postprocessing, depending on the needs, the following modules can be used: text correction module, close sounding and foreign words module, lemmatisation module. By using this method, it was possible to build the DBVT1 and DBVT2 databases.

5.2. Text Analyzer Module

The TEXT ANALYZER module receives calls from CC text channels and transcriptions of recordings made with the TRANSRIPTOR MODULE. The text analyser consists of the following components: CHAT PREPROCESSING, TF/IDF, WORD EMBEDDINGS, TEXT DATA PREPARING, TRAINING DATA IMPORT, AI/ML, TEXT CLASSIFIERS, EMOTICONS, and INFERENCE RULES.

For DBT1 data, using the CHAT PREPROCESSING component, operations are performed related to: change of Polish alphanumeric characters and tokenization. In the case of the DBVT1 database, these operations are performed by the TRANSCRIPTOR module. The processed text goes further into the TF/IDF and WORD EMBEDDINGS components, the functionality of which is described in Section 3. Then, the data is transferred to the TEXT DATA PREPARING component. The data collected in the DBT1 and DBVT1 databases is processed using balancing methods that are integrated into a component labelled BALANCING. The balancing is done according to certain parameters resulting in two balanced datasets named LEARNING DATA and TESTING DATA. The training data is imported into the TRAINING DATA IMPORT module, from where it is transferred to the machine learning mechanisms implemented in the AI/ML block. This block is responsible for preparing models that are further used by classifiers implemented in the TEXT CLASSIFIERS block. As described in Section 2, the following classifiers are integrated in this block: ANN, RFC, DT, kNN, SVM, ABC and NBC. After configuring the parameters for each classifier, the method provides the classification results. The EMOTICONS block verifies whether, in the classified utterances, there were defined emoticons indicating particular emotional states. In the next step, the collected data is transferred to the INFERENCE RULES module, where a specific emotion type is finally assigned to the studied utterance.

The data collected in the DBT2 and DBVT2 databases constitutes the validation data and is directly processed by the classifiers integrated in the TEXT ANALYZER module. On the other hand, the data collected in the DBVT2 database is used for verification tests using the ASR EMOTION DICTIONARY module.

For the text channel, the best results in the method verification process were obtained for the SVM classifier with the loss function squared l2 penalty applied. For SVM, in our solution: kernel = 'rbf', gamma = 'scale', shrinking = 'true', probability = 'false', cache size = 200 MB, decision function shape = 'ovr', tolerance = '0.001′ and there was no limit for maximum iteration. The selection of SVM classifier parameters was carried out by trial and error. For the studied ANN network, the best results were obtained when its structure consisted of the following layers: word embedding layer, BLSTM layer with 50 nodes, dropout layer with a step of 0.1, deep layer with 300 nodes, Relu activation function, dropout layer with a step of 0.1, deep layer with 120 nodes, linear Relu activation function layer, dropout layer with a step of 0.1, deep layer with 5 nodes, softmax layer. An "Adam" type optimizer with a learning rate of 0.001 and a categorical loss function cross entropy were used. For the kNN classifier, the best results were obtained with the number of neighbors equal to 5 and the Euclidean distance with uniform weights for all samples. For DT, the Gini coefficient was used, there was no maximum depth criterion, divisions were performed when the leaf contained more than two samples. On the other hand, for

the RFC classifier, 100 random trees were created, the Gini coefficient was used, with no maximum depth criterion, and divisions were performed when the leaf contained more than two samples.

5.3. Voice Analyzer Module

Conversations between agents and clients originating from the voice channel are transferred to the VOICE ANALYZER module. This module consists of the following components: VOICE CORRECTION, VOICE DATA PREPARING, TRAINING DATA IMPORT, AI/ML and VOICE CLASSIFIERS. In addition, it is complemented by the ASR EMOTIONS DICTIONARY and COMPARISON & INFERENCE RULES components. In the first phase, with the use of the VOICE CORRECTION component, the analysed audio signals are subjected to typical filtering and audio normalization related, for example, to the removal of any noise and ambient sound. This is an important step because, as explained in Section 2, the quality of the audio signal in CC systems is very poor, which can affect the performance of the method. Next, the data is transferred to the VOICE DATA PREPARING component, where longer fragments of recordings are divided into smaller parts. This resulted in the creation of DBV1 and DBV2 databases described in detail in Section 3.

The next step of the audio analysis module is analogous to that described earlier for text analysis. Data collected in the DBV1 database is processed using balancing methods. Training and testing sets are created and are subject to training and testing processes. As described in Section 2, the following classifiers are integrated in the VOICE CLASSIFIERS block: CNN, SVM, kNN and LDA. The data collected in the DBV2 database constitutes the validation data and is directly processed by the classifiers integrated in VOICE ANALYZER module. After configuring the parameters for each classifier the method provides the classification results, which are passed to the COMPARISON & INFERENCE RULES block. In the case of recognition of emotions like ANGER, FEAR, HAPPINESS, or SADNESS—this result is passed on to the system output. On the other hand, in the case of the recognition of the NEUTRAL state, the transcription of the studied fragment of the recording and the developed dictionary integrated in the ASR EMOTION DICTIONARY block are additionally used. In the first step, the transcription is classified using the classifiers integrated in the TEXT ANALYZER block, and the obtained results are sent to the ASR EMOTION DICTIONARY module. In parallel, the transcript is compared with the records contained in the dictionary of emotional expressions. If the studied expression (word or phrase) is found in the dictionary then the result of the algorithm's operation will be the emotion assigned to it. If the result of emotion recognition obtained at this stage is the same as with the result obtained with the use of the classifier, then the COMPARISON & INFERENCE RULES block will pass it to the system output. Otherwise (when the studied part of the transcription does not find a mapping in the dictionary or when these results will be different from the classification results) the final result of the system's operation will be the originally recognized NEUTRAL state. The task of the dictionary and the text channel classifiers used in the developed method is to optimize the efficiency of emotion recognition for the voice channel classifier. The integrated dictionary is based on data contained in the publicly available NAWL (Nencki Affective Word List) database [52] prepared by a team of Polish psychologists. This dictionary contains nearly three thousand phrases with emotional overtones (including: nouns, verbs, and adjectives). The information contained in the NAWL database was compiled on the basis of evaluations of more than 500 people [76]. The integrated dictionary, moreover, was expanded to include phrases identified by the judges as having an emotional tinge in the process of analyzing the DBVT1 database.

For the audio channel, the best results in the method verification process were obtained for the CNN network. The network structure consisted of the following layers: a weave layer consisting of 256 output filters in the weave, a layer of linear Relu activation function, a weave layer consisting of 128 output filters in the weave, another layer of linear Relu activation function, a dropout layer with a step of 0. 1, a MaxPooling1D layer for temporal data of size 8, a weave layer consisting of 128 output filters in the weave, a layer of linear

Relu activation function, a weave layer consisting of 128 output filters in the weave, a layer of linear Relu activation function, a Flatten layer flattening the input data to one dimension, a deep layer with five nodes, a layer of Softmax activation function. In addition, the following parameters were used: input size is chosen from {13, 18, 19, 24}—depending on parameter configuration for {MFFC − 13}, {MFCC + jitter − 18}, {MFCC + shimmer − 19} and {MFCC + jitter + shimmer − 24}, activation function is selected from {'relu', 'sigmoid'}, batch size is chosen from 8 to 128, learning rate is selected from {0.0001, 0.0005, 0.001, 0.01} and epochs number is selected from 50 to 2000. In order to avoid over-fitting the network, the effect of the learning parameters on the convergence of the algorithm and the error function values obtained for both the learning/testing and validation data were also evaluated in successive epochs. The implemented CNN obtains the best results for 1000 learning epochs; categorical cross-entropy class with an accuracy metric and RMSprop optimizer was used as a loss function. For the SVM classifier, the RBF (radial basis function) kernel function and L2 regularization with a weight of 1.0 were used. For the kNN classifier, the best results were obtained with the number of neighbors equal to 5, Euclidean distance and uniform weights.

6. Results

This section describes the findings obtained for the method proposed in Section 5. The following sections present (a) findings related to emotion recognition in the CC text channel (chat); (b) findings related to determining the effectiveness of emotion recognition from audio conversations; (c) findings related to determining the feasibility of improving emotion recognition performance in audio files through additional investigations of conversation transcripts derived from automatic transcriptions.

6.1. Text Channel (Chats)

Table 9 summarizes the findings obtained for the five-fold cross validation process, for the chat base designated as DBT1. The research used classifiers selected in Section 2, namely ABC, ANN, DT, kNN, NBC, RFC, and SVM. The performance and stability of the emotion classification models were evaluated through the mean values of ACC, WAP, WAF1 metrics and of standard deviation. The standard deviation values confirm the good stability of the prepared classification models. The results obtained for the studied metrics are usually in the range of about 50% to more than 60%, which is a satisfactory value from the point of view of the target applications of the proposed method. Only the NBC classifier produced weaker results.

Table 9. Cross validation test results for the text channel.

No	ABC			ANN			DT			kNN			NBC			RFC			SVM		
	ACC	WAP	WAF1	ACC	WAP	WAF1	ACC	WAP	WAF1	ACC	WAP	WAF1	ACC	WAP	WAF1	ACC	WAP	WAF1	ACC	WAP	WAF1
1	0.51	0.55	0.52	**0.60**	**0.60**	**0.61**	0.52	0.54	0.53	0.53	0.53	0.52	0.34	0.54	0.53	0.51	0.63	0.54	0.48	0.52	0.48
2	0.51	0.61	0.53	**0.67**	**0.67**	**0.66**	0.51	0.54	0.52	0.54	0.63	0.54	0.38	0.54	0.56	0.52	0.63	0.53	0.49	0.61	0.45
3	0.51	0.63	0.54	**0.64**	**0.64**	**0.65**	0.55	0.60	0.56	0.56	0.63	0.56	0.40	0.58	0.58	0.53	0.67	0.55	0.48	0.59	0.46
4	0.52	0.61	0.54	**0.63**	**0.63**	**0.63**	0.56	0.57	0.56	0.56	0.62	0.56	0.38	0.53	0.55	0.55	0.68	0.57	0.49	0.53	0.48
5	0.54	0.67	0.61	**0.66**	**0.66**	**0.68**	0.57	0.65	0.61	0.59	0.69	0.62	0.38	0.64	0.58	0.56	0.69	0.61	0.52	0.63	0.57
Av	0.52	0.61	0.55	**0.64**	**0.64**	**0.64**	0.54	0.58	0.56	0.56	0.62	0.56	0.38	0.57	0.56	0.53	0.66	0.56	0.50	0.58	0.50
Std	0.01	0.04	0.04	**0.03**	**0.03**	**0.03**	0.02	0.05	0.04	0.02	0.06	0.04	0.02	0.05	0.02	0.02	0.03	0.03	0.02	0.05	0.05

Av—average value; Std—standard deviation.

Table 10 shows the results obtained during the verification process of the developed method. For this purpose, data from the chat database designated as DBT2 was used. The experiment also detailed the different channels of interlocutors (agent/client). When analysing the results for the different interlocutors, it can be seen that in the case of chat conversations, better results are obtained for agents' statements. This is mainly influenced by the way they work. Agents in text channels very often use readymade response templates prepared by supervisors. It is common practice that readymade support texts used later by agents during conversations are prepared for the implementation of a given campaign. Thus, it is often the case that many of the statements made by different agents are identical, and this makes it easier to teach the classifiers. In the experiment to

verify the usefulness of the proposed method, similar to cross validation, the model based on the NBC classifier produced results well below expectations. Other models achieved satisfactory results.

Table 10. Verification of the method based on the DBT2 database.

Metric	Classifier						
	ABC	**ANN**	**DT**	**kNN**	**NBC**	**RFC**	**SVM**
Agent							
ACC	0.64	0.60	0.53	0.63	0.29	0.64	**0.75**
WAP	0.69	0.72	0.72	0.68	0.74	0.73	**0.72**
WAF1	0.66	0.65	0.60	0.65	0.26	0.67	**0.73**
Client							
ACC	0.53	0.53	0.48	0.50	0.32	0.51	**0.60**
WAP	0.51	0.57	0.51	0.50	0.56	0.51	**0.50**
WAF1	0.51	0.54	0.49	0.49	0.32	0.50	**0.54**

6.2. Voice Channel

Table 11 summarizes the test results obtained for the five-fold cross validation process obtained for the audio conversation base designated as DBV1. The research uses the classifiers selected in Section 2, namely: CNN, kNN, LDA and SVM.

Table 11. Cross validation test results for the voice channel.

No	CNN			kNN			LDA			SVM		
	ACC	**WAP**	**WAF1**	**ACC**	**WAP**	**WAF1**	**ACC**	**WAP**	**WAF1**	**ACC**	**WAP**	**WAF1**
1	**0.78**	**0.79**	**0.78**	0.71	0.66	0.68	0.66	0.58	0.61	0.72	0.70	0.67
2	**0.88**	**0.88**	**0.87**	0.70	0.69	0.68	0.67	0.62	0.63	0.71	0.64	0.65
3	**0.86**	**0.86**	**0.86**	0.69	0.66	0.65	0.62	0.64	0.57	0.69	0.63	0.63
4	**0.85**	**0.85**	**0.84**	0.69	0.69	0.66	0.66	0.58	0.61	0.64	0.74	0.57
5	**0.78**	**0.78**	**0.78**	0.71	0.69	0.69	0.66	0.57	0.60	0.70	0.76	0.65
Av	**0.83**	**0.83**	**0.83**	0.70	0.68	0.67	0.65	0.60	0.60	0.70	0.69	0.63
Std	**0.05**	**0.04**	**0.04**	0.01	0.02	0.02	0.02	0.03	0.02	0.03	0.05	0.04

The average values of the individual metrics obtained for the voice channel range from 60% to 80%. The weakest results were obtained using the LDA classifier, while the best results were obtained for CNN. The following speech signal descriptors were used in this test: F_0 MFCC, Jitter and Shimer. The obtained classification results are satisfactory for the applicability of the proposed method in the CC industry. The standard deviation values confirm the good stability of the prepared models. Table 12 summarizes the results obtained in the method verification experiment for emotion recognition in CC voice channels. For this purpose, the conversation base designated as DBV2 was used. As for the text data, the agent and client channels were specified in the experiment.

Analysis of the results of the experiments performed on the verification sample confirmed the effectiveness of the developed solution. In this case, the CNN classifier did a very good job of identifying the agents' emotions. In addition, the results for emotion recognition in voice channels are clearly better than for the text channel. This is due to the nature of the audio signal, which contains much more information that allows the identification of individual types of emotions using artificial intelligence algorithms.

Table 12. Verification of the method based on the DBV2 database.

Metric	Classifier			
	CNN	kNN	LDA	SVM
		Agent		
ACC	**0.68**	0.53	0.54	0.63
WAP	**0.97**	0.56	0.80	0.54
WAF1	**0.80**	0.54	0.64	0.58
		Client		
ACC	**0.67**	0.52	0.47	0.61
WAP	**0.71**	0.79	0.71	0.74
WAF1	**0.69**	0.61	0.55	0.66

6.3. Voice Channel with Text Method Combined

The last piece of research for the method proposed in this paper consisted of experiments in which automatic transcriptions of conducted voice conversations were also used in the classification processes. The results obtained in this case are presented in Table 13.

Table 13. Verification in voice channel including transcriptions and dictionary.

Metric	Classifier			
	CNN	kNN	LDA	SVM
		Client		
ACC	**0.68**	0.53	0.48	0.62
WAP	**0.80**	0.81	0.73	0.76
WAF1	**0.73**	0.62	0.57	0.67

Table 13 presents the results for the client channel, as applying this technique to the agent channel did not improve the final classification results. This is probably due to the fact that there are relatively few emotionally charged utterances in the agent channel, and the agents themselves often use the hints suggested to them by the system. In addition, transcriptions obtained from an audio conversation lack many of the specific features found in typical chat conversations. During chat conversations, interlocutors often use additional characters (e.g., emoticons, punctuation) to convey emotions. As a result, the efficiency of emotion recognition in chat conversations is at a relatively high level. The transcript is only a simple shorthand of the audio conversation devoid of much of the relevant information that is contained in the chats. Nevertheless, thanks to the proposed approach, the values of individual metrics in the customer's channel improved, a result worth noting.

7. Conclusions

This paper proposes a new method for recognizing emotions of interlocutors on a CC hotline. The method assumes recognition of emotional states in both text (chat type) and voice channels. The proposed solution provides capabilities to recognize the emotions of clients contacting the hotline as well as the agents handling these calls, which is important for practical applications. In verification experiments for the prepared classification models, the obtained emotion recognition results for the studied metrics in the text channel even reach values of over 70% for agent utterances and up to 60% for client utterances. In the voice channel, taking into account the conversation transcription and the dictionary integrated in the system, the results for the CNN classifier in both channels reach values above 68%. Taking into account the usual poor audio signal quality found in CC systems, it can be clearly stated that the obtained test results are very satisfactory and show a very high potential for commercial application of the proposed solution. A high level of customer service is becoming increasingly important in the CC industry, and the method proposed by the authors aims to help with these aspects. The solution is expected to improve the

quality of customer service and optimize relevant KPIs (key performance indicators). These conclusions apply to both the voice channel and the text channel.

The developed solution can be easily integrated with different functionalities used in CC systems. This makes it feasible to widely use the developed method in commercial applications. Functionalities in which the emotion recognition method dedicated to the CC industry can be used are, for example, solutions for chatbots and voicebots, methods of creating behavioural profiles of interlocutors (both agents and clients), components responsible for the best possible matching of interlocutors, or in solutions helping to evaluate the work of agents. The method proposed in the paper, after several modifications, can be generalized to applications with a language other than Polish. However, this requires optimization of its selected components. The first of the components requiring optimization is the transcription method implemented as TRANSCRIPTOR MODULE, which in the described solution has been adapted to the needs of selected campaign topics conducted in Polish. In this method, first of all, it would be necessary to optimize the algorithms operating on the postprocessing side. Another element requiring modification is the emotional word dictionary component, implemented in the form of the ASR EMOTION DICTIONARY block. In addition, in order to make the solution even more versatile, relevant retraining processes can be additionally implemented. The task of these solutions should be to constantly and dynamically correct the reference models used.

Further development of the proposed topic includes tasks related to the construction and implementation of virtual assistants, whose algorithms will be able to take into account the emotional states of the client occurring during the conversation. It is also anticipated that the method proposed in this paper can be used in CC systems of the future, where popular IoT (Internet of Things) technologies currently implemented in many areas of technology can be integrated. In addition, it can be assumed that the rapidly developing new information processing and transmission technologies, especially video technologies, which could also support emotion recognition mechanisms, will soon become much more important in customer service. Moreover, the limitation of the currently developed method is support only for the Polish language. The next step in development is extending the method to support conversations in other languages. It is also important to increase the size of the learning databases and prepare automated retraining methods, which would improve the accuracy of emotion recognition.

Author Contributions: M.P.: conceptualization, methodology, investigation, writing—original draft, supervision, project administration, funding acquisition; R.K.: formal analysis, investigation, resources, data curation, validation, writing—review and editing; Z.K.: conceptualization, methodology, writing—review and editing; M.K.: investigation, software, formal analysis, data curation; M.L.: conceptualization, methodology; K.S.: software; J.S.: software. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by European Union's Smart Growth Operational Programme 2014–2022 under grant agreement no POIR.04.01.04-00-0079/19.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The research was conducted based on real talks obtained from a largecontact centre system.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jyoti, R.; Wardley, M. Unlocking the Transformative Power of AI for Contact Centers. IDC InfoBrief, October 2020. Available online: https://inthecloud.withgoogle.com/idc-infobrief/CCAI-IDC-Infobrief.html (accessed on 22 April 2022).
2. DMG Consulting LLC. *The State of Artificial Intelligence in the Contact Center*; Report; DMG Consulting LLC: West Orange, NJ, USA, 2022.
3. Kask, S.; Fitterer, R.; Anshelm, L. *Augmenting Digital Customer Touchpoints: Best Practices for Transforming Customer Experience Through Conversational AI*; Marketing Review; University St. Gallen: Sankt Gallen, Switzerland, 2019; Volume 5, pp. 64–69.

4. Plaza, M.; Pawlik, L. Influence of the Contact Center Systems Development on Key Performance Indicators. *IEEE Access* **2021**, *9*, 44580–44591. [CrossRef]
5. Google Cloud. Natural Language API Basics. 2021. Available online: https://cloud.google.com/naturallanguage/docs/basics#sentiment_analysis (accessed on 10 May 2022).
6. Google Cloud. Language Reference. 2021. Available online: https://cloud.google.com/dialogflow/es/docs/reference/language (accessed on 1 August 2022).
7. Stolletz, R.; Helber, S. Performance analysis of an inbound call center with skills-based routing. *OR Spektrum* **2004**, *26*, 331–352. [CrossRef]
8. Jahangir, R.; Teh, Y.W.; Hanif, F.; Mujtaba, G. Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimed. Tools Appl.* **2021**, *8*, 23745–23812. [CrossRef]
9. Roy, T.; Marwala, T.; Chakraverty, S. A Survey of Classification Techniques in Speech Emotion Recognition. In *Mathematical Methods in Interdisciplinary Sciences*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2020; pp. 33–48. [CrossRef]
10. Swain, M.; Routray, A.; Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: A review. *Int. J. Speech Technol.* **2018**, *21*, 93–120. [CrossRef]
11. Rubin, V.L.; Stanton, J.M.; Liddy, E.D. Discerning Emotions in Texts, The AAAI Symposium on Exploring Attitude and Affect in Text AAAI-EAAI, Stanford, CA, 2004. Available online: https://www.aaai.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-023.pdf (accessed on 10 August 2022).
12. Sathe, J.B.; Mali, M.P. A hybrid Sentiment Classification method using Neural Network and Fuzzy Logic. In Proceedings of the 11th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, India, 5–6 January 2017; pp. 93–96. [CrossRef]
13. Khan, M.T.; Durrani, M.; Ali, A.; Inayat, I.; Khalid, S.; Khan, K.H. Sentiment analysis and the complex natural language. *Complex Adapt. Syst. Model.* **2016**, *4*, 1–19. [CrossRef]
14. Dragoni, M.; Tettamanzi, A.G.B.; Pereira, C.D.C. A Fuzzy System for Concept-Level Sentiment Analysis. In *Semantic Web Evaluation Challenge*; Springer: Cham, Switzerland, 2014; pp. 21–27. [CrossRef]
15. Pawlik, Ł.; Płaza, M.; Deniziak, S.; Boksa, E. A method for improving bot effectiveness by recognising implicit customer intent in contact centre conversations. *Speech Commun.* **2022**, *143*, 33–45. [CrossRef]
16. Ekman, P.; Friesen, W.V. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*; Malor Books: Los Altos, CA, USA, 2003.
17. Liu, B. Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167.
18. Agarwal, A.; Xie, B.; Vovsha, I.; Rambow, O.; Passonneau, R. Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media, Portland, Orego, 23 June 2011; pp. 30–38.
19. Phan, H.T.; Tran, V.C.; Nguyen, N.T.; Hwang, D. Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model. *IEEE Access* **2020**, *8*, 14630–14641. [CrossRef]
20. Sitaula, C.; Basnet, A.; Mainali, A.; Shahi, T.B. Deep Learning-Based Methods for Sentiment Analysis on Nepali COVID-19-Related Tweets. *Comput. Intell. Neurosci.* **2021**, *2021*, 2158184. [CrossRef]
21. Ortigosa, A.; Martín, J.M.; Carro, R.M. Sentiment analysis in Facebook and its application to e-learning. *Comput. Hum. Behav.* **2014**, *31*, 527–541. [CrossRef]
22. Jianqiang, Z.; Xiaolin, G.; Xuejun, Z. Deep Convolution Neural Networks for Twitter Sentiment Analysis. *IEEE Access* **2018**, *6*, 23253–23260. [CrossRef]
23. Zeng, J.; Ma, X.; Zhou, K. Enhancing Attention-Based LSTM With Position Context for Aspect-Level Sentiment Classification. *IEEE Access* **2019**, *7*, 20462–20471. [CrossRef]
24. Wang, J.; Yu, L.-C.; Lai, K.R.; Zhang, X. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 2, pp. 225–230.
25. Shekhawat, S.S.; Shringi, S.; Sharma, H. Twitter sentiment analysis using hybrid Spider Monkey optimization method. *Evol. Intell.* **2020**, *14*, 1307–1316. [CrossRef]
26. Bouazizi, M.; Ohtsuki, T. A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. *IEEE Access* **2017**, *5*, 20617–20639. [CrossRef]
27. Wang, Y.; Kim, K.; Lee, B.; Youn, H.Y. Word clustering based on POS feature for efficient twitter sentiment analysis. *Hum.-Cent. Comput. Inf. Sci.* **2018**, *8*, 17. [CrossRef]
28. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 6–8 July 2020; pp. 4040–4054. [CrossRef]
29. Kumar, P.; Raman, B. A BERT based dual-channel explainable text emotion recognition system. *Neural Netw.* **2022**, *150*, 392–407. [CrossRef]
30. Saad, S.E.; Yang, J. Twitter Sentiment Analysis Based on Ordinal Regression. *IEEE Access* **2019**, *7*, 163677–163685. [CrossRef]
31. Busso, C.; Lee, S.; Narayanan, S. Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 582–596. [CrossRef]
32. Kuchibhotla, S.; Vankayalapati, H.D.; Vaddi, R.S.; Anne, K.R. A comparative analysis of classifiers in emotion recognition through acoustic features. *Int. J. Speech Technol.* **2014**, *17*, 401–408. [CrossRef]

33. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. *Int. J. Speech Technol.* **2012**, *15*, 99–117. [CrossRef]
34. Fahad, S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **2020**, *110*, 102951. [CrossRef]
35. Smagowska, B. Noise at Workplaces in the Call Center. *Arch. Acoust.* **2010**, *35*, 253–264. [CrossRef]
36. Zadrozny, S.; Kacprzyk, J.; Gajewski, M. Multiaspect Text Categorization Problem Solving: A Nearest Neighbours Classifier Based Approaches and Beyond. *J. Autom. Mob. Robot. Intell. Syst.* **2015**, *9*, 58–70. [CrossRef]
37. Moraes, R.; Valiati, J.F.; Neto, W.P.G. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Syst. Appl.* **2013**, *40*, 621–633. [CrossRef]
38. Zhang, Y.; Song, D.; Zhang, P.; Li, X.; Wang, P. A quantum-inspired sentiment representation model for twitter sentiment analysis. *Appl. Intell.* **2019**, *49*, 3093–3108. [CrossRef]
39. Pratama, B.Y.; Sarno, R. Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In Proceedings of the International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, Indonesia, 1 November 2015. [CrossRef]
40. Saha, T.; Patra, A.; Saha, S.; Bhattacharyya, P. Towards Emotion-aided Multi-modal Dialogue Act Classification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 6–8 July 2020; pp. 4361–4372. [CrossRef]
41. Zhang, M.; Palade, V.; Wang, Y.; Ji, Z. Attention-based word embeddings using Artificial Bee Colony algorithm for aspect-level sentiment classification. *Inf. Sci.* **2020**, *545*, 713–738. [CrossRef]
42. Qayyum, A.B.A.; Arefeen, A.; Shahnaz, C. Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. In Proceedings of the IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, 28–30 November 2019; pp. 122–125. [CrossRef]
43. Zhou, J.; Wang, G.; Yang, Y.; Chen, P. Speech Emotion Recognition Based on Rough Set and SVM. In Proceedings of the 5th IEEE International Conference on Cognitive Informatics, Beijing, China, 17–19 July 2006; pp. 53–61. [CrossRef]
44. Feraru, M.; Zbancioc, M. Speech emotion recognition for SROL database using weighted KNN algorithm. In Proceedings of the International Conference on Electronics, Computers and Artificial Intelligence, Pitesti, Arkansas, 27–29 June 2013; pp. 1–4. [CrossRef]
45. You, M.; Chen, C.; Bu, J.; Liu, J.; Tao, J. A Hierarchical Framework for Speech Emotion Recognition. *IEEE Int. Symp. Ind. Electron.* **2006**, *1*, 515–519. [CrossRef]
46. Cho, J.; Pappagari, R.; Kulkarni, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Deep Neural Networks for Emotion Recognition Combining Audio and Transcripts. *arXiv* **2019**, arXiv:1911.00432. [CrossRef]
47. Tripathi, S.; Kumar, A.; Ramesh, A.; Singh, C.; Yenigalla, P. Deep learning based emotion recognition system using speech features and transcriptions. *arXiv* **2019**, arXiv:1906.05681. [CrossRef]
48. Kim, E.; Shin, J.W. DNN-based Emotion Recognition Based on Bottleneck Acoustic Features and Lexical Features. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6720–6724. [CrossRef]
49. Li, J.-L.; Lee, C.-C. Attentive to Individual: A Multimodal Emotion Recognition Network with Personalized Attention Profile. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019. [CrossRef]
50. Santoso, J.; Yamada, T.; Makino, S.; Ishizuka, K.; Hiramura, T. Speech Emotion Recognition Based on Attention Weight Correction Using Word-Level Confidence Measure. In Proceedings of the Interspeech 2021, Brno, Czechia, 30 August–3 September 2021. [CrossRef]
51. Atmaja, B.T.; Sasou, A.; Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* **2022**, *140*, 11–28. [CrossRef]
52. NAWL. Interactive Analysis of the NAWL Database. 2022. Available online: https://exp.lobi.nencki.gov.pl/nawl-analysis (accessed on 12 June 2022).
53. Plaza, M.; Pawlik, L.; Deniziak, S. Call Transcription Methodology for Contact Center Systems. *IEEE Access* **2021**, *9*, 110975–110988. [CrossRef]
54. Płaza, M.; Trusz, S.; Kęczkowska, J.; Boksa, E.; Sadowski, S.; Koruba, Z. Machine Learning Algorithms for Detection and Classifications of Emotions in Contact Center Applications. *Sensors* **2022**, *22*, 5311. [CrossRef]
55. Scikit-Learn User Manual. Available online: https://scikit-learn.org/stable/modules/model_evaluation.html# (accessed on 16 August 2022).
56. Behera, B.; Kumaravelan, G.; Kumar, P. Performance Evaluation of Deep Learning Algorithms in Biomedical Document Classification. In Proceedings of the 11th International Conference on Advanced Computing (ICoAC), Chennai, India, 18–20 December 2019; pp. 220–224. [CrossRef]
57. López, V.; Fernández, A.; García, S.; Palade, V.; Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **2013**, *250*, 113–141. [CrossRef]
58. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
59. Chawla, N.V. Data mining for imbalanced datasets: An overview. In *Data Mining and Knowledge Discovery Handbook*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 875–886.

60. Prusa, J.; Khoshgoftaar, T.M.; Dittman, D.J.; Napolitano, A. Using Random Undersampling to Alleviate Class Imbalance on Tweet Sentiment Data. In Proceedings of the IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 13–15 August 2015; pp. 197–202. [CrossRef]
61. Lemaitre, G.; Nogueira, F.; Aridas, C. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced satasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
62. The Imbalanced-Learn Developers. Imbalanced-Learn Documentation. 2022. Available online: https://imbalanced-learn.org/ (accessed on 1 June 2022).
63. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **2003**, *39*, 45–65. [CrossRef]
64. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
65. Scikitlearn. Tf–idf Term Weighting. 2022. Available online: https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting (accessed on 11 May 2022).
66. Gebre, B.G.; Zampieri, M.; Wittenburg, P.; Heskes, T. Improving Native Language Identification with TF-IDF Weighting. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Atlanta, GA, USA, 13 June 2013; pp. 216–223.
67. Murugappan, M.; Rizon, M.; Nagarajan, R.; Yaacob, S.; Hazry, D.; Zunaidi, I. Time-Frequency Analysis of EEG Signals for Human Emotion Detection. *IFMBE Proc.* **2008**, *21*, 262–265. [CrossRef]
68. Kong, J. A Study on Jitter, Shimmer and F0 of Mandarin Infant Voice by Developing an Applied Method of Voice Signal Processing. In Proceedings of the Congress on Image and Signal Processing, Sanya, China, 27–30 May 2008; pp. 314–318. [CrossRef]
69. Korkmaz, O.E.; Atasoy, A. Emotion recognition from speech signal using mel-frequency cepstral coefficients. In Proceedings of the 9th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, 26–28 November 2015; pp. 1254–1257. [CrossRef]
70. Ancilin, J.; Milton, A. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Appl. Acoust.* **2021**, *179*, 108046. [CrossRef]
71. Chamoli, A.; Semwal, A.; Saikia, N. Detection of emotion in analysis of speech using linear predictive coding techniques (L.P.C). In Proceedings of the International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 19–20 January 2017; pp. 1–4. [CrossRef]
72. Basu, S.; Chakraborty, J. Aftabuddin Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In Proceedings of the 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 19–20 October 2017; pp. 333–336. [CrossRef]
73. Wang, K.; An, N.; Li, B.N.; Zhang, Y.; Li, L. Speech Emotion Recognition Using Fourier Parameters. *IEEE Trans. Affect. Comput.* **2015**, *6*, 69–75. [CrossRef]
74. Aouani, H.; Ben Ayed, Y. Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder. In Proceedings of the 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, 21–24 March 2018; pp. 1–5. [CrossRef]
75. Saste, S.T.; Jagdale, S.M. Emotion recognition from speech using MFCC and DWT for security system. In Proceedings of the 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 20–22 April 2017; pp. 701–704. [CrossRef]
76. Riegel, M.; Wierzba, M.; Wypych, M.; Żurawski, Ł.; Jednoróg, K.; Grabowska, A.; Marchewka, A. Nencki Affective Word List (NAWL): The cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behav. Res. Methods* **2015**, *47*, 1222–1236. [CrossRef] [PubMed]