Combining a multi-feature neural network with multi-task learning for emergency calls severity prediction

Marianne Abi Kanaan[a,b,*,1], Jean-François Couchot[a], Christophe Guyeux[a], David Laiymani[a], Talar Atechian[b] and Rony Darazi[b]

[a]FEMTO-ST Institute, CNRS, Université de Franche-Comté, Besançon, France
[b]TICKET Lab, Université Antonine (UA), Baabda, Lebanon

ARTICLE INFO

Keywords:
Emergency Calls
Natural Language Processing
Speech Emotion Recognition
Deep Learning

ABSTRACT

In emergency call centers, operators are required to analyze and prioritize emergency situations prior to any intervention. This allows the team to deploy resources efficiently if needed, and thereby provide the optimal assistance to the victims. The automation of such an analysis remains challenging, given the unpredictable nature of the calls. Therefore, in this study, we describe our attempt in improving an emergency calls processing system's accuracy in the classification of an emergency's severity, based on transcriptions of the caller's speech. Specifically, we first extend the baseline classifier to include additional feature extractors of different modalities of data. These features include detected emotions, time-based features, and the victim's personal information. Second, we experiment with a multi-task learning approach, in which we attempt to detect the nature of the emergency on the one hand, and improve the severity classification score on the other hand. Additional improvements include the use of a larger dataset and an explainability study of the classifier's decision-making process. Our best model was able to predict 833 emergency calls' severity with a 71.27% accuracy, a 5.33% improvement over the baseline model. Moreover, we extended our tool with additional modules that can prove to be useful when handling emergency calls.

1. Introduction

In the case of an injury or illness, citizens usually contact emergency call centers to seek medical assistance. In France, the SDIS (Service Départemental d'Incendie et de Secours) department of a specific region handles the assistance of such emergencies around the clock. Following an emergency call, the operators usually have to determine the priority that should be assigned to this emergency, based on their assessment of the situation's severity and urgency.

Several factors can affect the decision-making process of the operator: the medical expertise of the operator handling the call, whether the operator is overloaded with calls and therefore is not in capacity to accurately assess the needs of the caller, etc. However, an inaccurate assessment of the situation's urgency could result in a late intervention, thereby increasing the risk of avoidable fatalities. Consequently, it is crucial to equip emergency center operators with effective methods that can assist them in the evaluation of an emergency's priority.

In one approach (Abi Kanaan, Couchot, Guyeux, Laiymani, Atechian and Darazi (2023)), a pipeline (Fig 1) is developed for processing and classifying emergency calls. The speech regions in the call are first extracted by a voice activity detection algorithm. Then, speaker diarization is applied on these signals in order to extract the caller's voice separately. The purpose of this process is to emulate a scenario where the operator is not able to assist the caller due to an overload of calls for instance. In such a case, the emergency center could set up a waiting machine that would request the caller to indicate the reason for their call, and to describe their emergency. The caller's audio signals are later passed into a speech-to-text system. Based on the transcribed text, the call is finally classified as either a "High Severity" or a "Low Severity" call. A "High Severity" label indicates that the potential outcome of the emergency might involve a dangerous medical condition or the passing of the victim. A "Low Severity" label on the other hand indicates that the emergency could amount to a minor or no medical condition. The classifier is a French version of BERT (Bidirectional Encoder Representations from Transformers) (Tenney, Das and Pavlick (2019)), CamemBERT (Martin, Muller, Suárez, Dupont, Romary, de La Clergerie, Seddah and Sagot (2019)), and was able to estimate the severity of 90 emergency calls with a 71.2% accuracy and a standard deviation of 3.02%.

Given that each improvement to this score can contribute to a saved life, we aim to improve upon this work with several contributions. First, we evaluate the accuracy improvement whilst training the text classifier on a larger dataset, which allows us to further evaluate the system's predictions on a larger sample. Second, we improve the accuracy of the system by augmenting the text classifier with additional inputs: the emotions exhibited by the caller, the call's time-based features, and the emergency victim's personal information (i.e., age, gender, and location). Moreover, we investigate the effect of incorporating multi-task learning into the model on the accuracy. We were able to further increase the accuracy in the prediction of the severity level when the additional tasks were correlated enough.

Our contributions in this work can therefore be summarized in the following way:

*Corresponding author
marianne.abi_kanaan@univ-fcomte.fr (M. Abi Kanaan)
ORCID(s): 0000-0001-7511-2910 (M. Abi Kanaan)

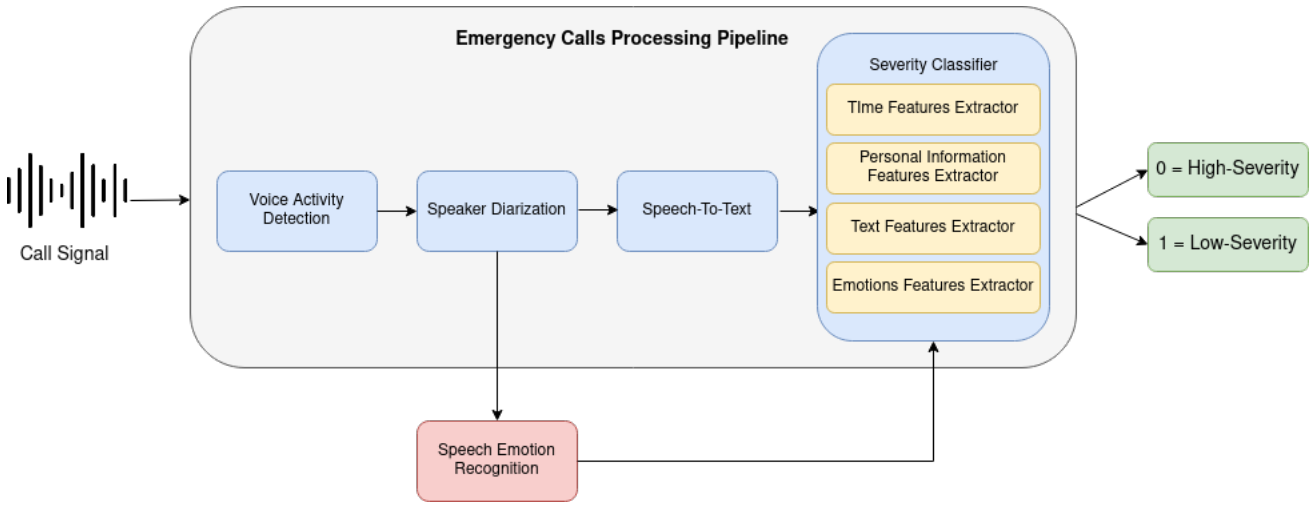


Figure 1: Emergency calls processing pipeline (Abi Kanaan et al. (2023)): improvements include the addition of an SER module and feature extractors in the severity classifier.

- We train a baseline system (Abi Kanaan et al. (2023)) and evaluate it on a larger dataset (an increase of 460.95% on the number of calls), thereby increasing its reliability.
- We increase the accuracy of the system by augmenting the speech classifier with additional features on the one hand, and by incorporating a multi-task learning approach on the other hand. Our best model achieves a mean accuracy of 71.27%, a 5.33% improvement over the baseline model (Abi Kanaan et al. (2023)) trained on the dataset of this study.
- We automate the speaker identification phase of the system, which improves its usability in a real-time setting.
- We include a study on the explainability of the deep learning models used in this work, in order to gain a better insight into the decision-making process of these algorithms.

The code related to this study will be available on the following URL `https://tinyurl.com/4pk4uf67`.

The remainder of this paper is organized as follows: Section 2 summarizes state of the art works related to emergency calls classification, speech emotion recognition, and multi-task learning. Section 3 covers the proposed improvements in detail. The experimental process as well as the results of this work, the explainability study, and a performance analysis of the system are reported in Section 4. In Section 5, we include a discussion of the obtained results and the current limitations of this study. Finally, Section 6 concludes this article and discusses the future directions for this work.

2. Related Work

This section first describes state of the art works in emergency calls classification. Second, an overview of speech emotion recognition applications in medical and emergency contexts is given, as well as a description of studies on the impact of multi-task learning.

2.1. Emergency Calls Classification

A first group of works have focused on the classification of calls into a medical diagnosis (Blomberg, Folke, Ersbøll, Christensen, Torp-Pedersen, Sayre, Counts and Lippert (2019)). The authors use a machine learning framework developed by the Danish company Corti.ai (cor (2023)) in the recognition of cardiac arrests in callers' speech extracted from automatically transcribed text. The study shows that the framework achieves a higher sensitivity rate (84.1%) compared to the dispatcher (72.5%). Another work (Gil-Jardiné, Chenais, Pradeau, Tentillier, Revel, Combes, Galinski, Tellier and Lagarde (2021)) uses a GPT-2 (Generative Pre-Trained Transformer) (Radford, Narasimhan, Salimans, Sutskever et al. (2018)) model in the classification of emergency call notes taken by medical experts, into one of several emergency categories, such as chest pain, violence, etc. The maximum F_1 scores on the categories range from 47.9% to 80%.

Several other works have described the development of tools that assist operators in the prioritization of calls. An emergency center support system is designed (Trujillo, Orellana and Acosta (2019)), with the combination of several modules. The calls in Spanish are first transcribed through an Automatic Speech Recognition (ASR) system, and are passed through a Named Entity Recognition (NER) module for the extraction of relevant entities. An additional classifier module detects the service type and priority of a specific call's transcription, using algorithms such as TF-IDF (Term Frequency-Inverse Document Frequency) and Support Vector Machines (SVMs). The emergency calls classifier is discussed in detail in another study Orellana, Trujillo and Acosta (2020). The texts goes through several

pre-processing steps, such as conversion to a lowercase format, stop-words removal, and lemmatization. Furthermore, the texts are subjected to "word pruning" in order to reduce the dimensionality of the features. The best model results in a recall score of 86%, a precision score of 75%, and an F_1-score of 80%. Another tool, which constitutes the basis for this work (Abi Kanaan et al. (2023)), was developed to assist emergency call operators in France. Emergency calls in the French language are first passed through audio processing blocks, such as voice activity detection and speaker diarization. Another block automatically transcribes the calls into text. The latter is then used to train a BERT model (Tenney et al. (2019)) on 904 emergency calls, for the prediction of the injury level of the victim concerned by the call. The latter achieves a 71.2% accuracy on the classification of the severity.

2.2. Emotion recognition applications

In emotion recognition applications, emotions are typically modeled in one of two ways: either through a discrete representation, or a dimensional representation (Akçay and Oğuz (2020)). A discrete emotion is based on one of six basic categories of emotions such as anger, happiness, fear, etc. The dimensional model on the other hand argues that since emotions constantly change, an alternative representation could be through continuous dimensions that encompass the pleasantness, i.e., the valence, of an emotion, and its intensity, or the arousal.

Furthermore, with the use of machine learning and deep learning algorithms, the emotions can be inferred based on either the acoustic signals of a speech, the textual contents of a speech, facial features (in the case of video recordings), or a fusion of these features. One work (Omar and Abd El-Hafeez (2023)) experiments with quantum computing and classic machine learning methods to perform sentiment classification on Arabic documents. Another study (Ayache and Alti (2020)) suggests a system for facial expression recognition. It performs feature selection on faces using Active Shape Model (ASM). These features are used to train several classifiers, such as a Quadratic classifier (DA), a Multi Layer Perceptron (MLP), etc. It was found that the Quadratic classifier provides the most accurate classification results. In this study, we focus on acoustic-based emotion recognition applications, as this is the one modality that is available to us in its unchanged format (no facial features are available, and the transcriptions are not 100% error-free). We expect that the voices of the callers can exhibit enough emotional information.

In speech emotion recognition applications, the most commonly used acoustic features are prosodic (e.g. pitch, loudness, duration) and spectral features (e.g. MFCC; Mel Frequency Cepstral Coefficients) (Akçay and Oğuz (2020)). In one speech emotion recognition study (Kumar, Haq, Jain, Jason, Moparthi, Mittal and Alzamil (2023)), the authors extract MFCC features from speech signals and use a multilayer perceptron (MLP) to classify the features into a category of emotion. In contrast, another work (Zhao, Mao and Chen (2019)) uses an alternative representation for speech, the log-mel spectrogram, which represents the frequency changes in the signal over time. These spectrograms are used to train a CNN-LSTM (Convolutional Neural Network with an LSTM) on the task of speech emotion recognition. This architecture achieves the following scores in a speaker-independent setting: a 95.89% accuracy for EmoDB database Burkhardt, Paeschke, Rolfes, Sendlmeier, Weiss et al. (2005) and 52.14% on IEMOCAP (Busso, Bulut, Lee, Kazemzadeh, Mower, Kim, Chang, Lee and Narayanan (2008)).

Many works have studied the impact of emotions in a medical or emergency context. In one study (Deschamps-Berger, Lamel and Devillers (2021)), a network based on Convolutional Neural Networks and Bidirectional LSTMs (CNN-BiLSTM) is developed for speech emotion recognition and trained on the acoustic signals found in emergency calls. The model predicts one of four categorical emotions: anger, sadness, happiness, and neutrality. The authors demonstrated the difficulty of accurately recognizing real-life emotions through neural networks, compared to the performance when using the improvised section of IEMOCAP database (Busso et al. (2008)). They obtain a 45.6% unweighted accuracy on the four classes using the real-life dataset, compared to 63% obtained on IEMOCAP. The previous work is extended (Deschamps-Berger, Lamel and Devillers (2022)) with several improvements, such as the use of transformers (Minaee, Kalchbrenner, Cambria, Nikzad, Chenaghlu and Gao (2021)), and the fusion of textual and acoustic features for the classification of emotions. This improved the previously obtained 45.6% unweighted accuracy to 77.1%. The authors also mention that the use of textual features improved the recognition in complex calls, as text complemented the acoustic features when the callers attempted to exaggerate or control their emotions.

As opposed to the previously described studies that have relied on discrete emotions, one work (Perez-Toro, Vasquez-Correa, Bocklet, Noth and Orozco-Arroyave (2021)) utilizes dimensional emotions in a clinical context. More specifically, the emotional features are used in the detection of depression in Parkinson's patients, and the detection of Alzheimer's disease. The classifier is based on a fusion of linguistic and acoustic features. This results in F_1 scores of up to 82% for the depression detection, and up to 80% for Alzheimer's detection. In this study, we equally rely on a dimensional representation of emotions, as we seek to collect this information on a continuous level (for several intervals of the calls).

2.3. Multi-task Learning

Multi-task learning (MTL) is an approach for training machine learning models, in which the same model can be trained on multiple tasks simultaneously, while several loss functions are optimized at once. The purpose of such a training approach is to allow the model to leverage the features that are relevant in multiple tasks. This way, the input data can be represented more efficiently, which can improve

Table 1
Summary of existing works on emergency calls classification.

Study	Goal	Dataset	Methods	Results	Limitations
Orellana et al. (2020)	Classification of emergency calls priority	1000 emergency call transcripts in Spanish provided by a security service	Use of text pre-processing techniques. A TF-IDF-based representation of the texts. Use of SVMs for the classifications of the texts	Precision of 75%, F_1-score of 80%, Recall of 86%	Limited dataset (only 1000 calls). Text requires several pre-processing steps.
Gil-Jardiné et al. (2021)	Classification of emergency calls' reason	Manually annotated notes of French 888,469 emergency calls	Use of GPT-2 for the classification of the texts	F_1 scores ranging from 47.9% to 80%	The work is based on manually annotated notes of the calls. No severity classification.
Abi Kanaan et al. (2023)	Classification of emergency calls severity	Automatically transcribed 904 emergency calls in French	Use of CamemBERT for the classification of the texts	71.2% accuracy	Limited dataset (only 904 calls).

the performance when the tasks have some correlation. The method of multi-task learning has been used in a variety of applications, such as text classification, medical image analysis, and speech emotion recognition.

In a work aiming at automating the evaluation of peer assessments (Jia, Cui, Xiao, Liu, Rashid and Gehringer (2021)), a multi-task learning BERT model was employed for detecting features in assessments such as the tone, suggestions, etc. It was shown that this joint training approach, as opposed to dedicating a separate model for each task, improved performance in terms of accuracy, memory usage, and response time. Multi-task learning was also successfully applied in another study (Goncharov, Pisov, Shevtsov, Shirokikh, Kurmukov, Blokhin, Chernina, Solovev, Gombolevskiy, Morozov et al. (2021)) to improve the detection of Covid-19 and its severity, based on CT images of patients. Another application of MTL involves a model for speech emotion recognition in emergency call centers (Deschamps-Berger et al. (2021)). The involved tasks in this approach are the prediction of the emotion and the gender of the caller. As opposed to the previously mentioned works, the joint learning of an auxiliary task (the gender recognition), did not seem to single-handedly improve the performance in the recognition of emotions.

In Table 1, we summarize some of the most relevant works related to our goal. The table shows some of the limitations of the mentioned studies, such as the evaluation of classifiers on manually annotated notes of emergency calls (which are not available in our case), or the evaluation on a limited dataset. In the current study, we rely on automatically annotated transcriptions of calls, and evaluate our models on a bigger number of samples. Moreover, we do not preprocess our transcriptions, so as to avoid an additional slowdown of our pipeline.

3. Methods

In this section, we include an overview of the datasets used in this work. We then describe the contributions in this study, specifically in the extension of the text classifier with additional feature extractors (see Fig. 2).

3.1. Datasets

3.1.1. Emergency calls dataset

The SDIS 25, an emergency department in the French Doubs region, provided the emergency calls used in this work. The calls had taken place during the range of the years 2016 to 2021. Some of the recordings were filtered out for being irrelevant to our task, as they included conversations between operators, medical professionals, or policemen discussing the details of an emergency intervention. In this work, we will rather focus on the analysis of calls that have been initiated by a civilian who is directly involved in the situation. In the recorded conversations, the caller is describing their situation, while the operator attempts to assist them and get a clearer understanding of the emergency.

As shown in Abi Kanaan et al. (2023), using only the caller's parts in these recordings contained enough information to amount to a similar performance in the classification of the severity compared to using the complete recordings. For this reason, we limit our experiments to the caller's speech that was extracted from the conversations using a speaker diarization tool (Bredin, Yin, Coria, Gelly, Korshunov, Lavechin, Fustes, Titeux, Bouaziz and Gill (2020)). The calls are labeled with the reason for the call (e.g., automobile accident, loss of consciousness...) and the victim's condition following the intervention of the team. This condition can either be what is called "Lightly injured", meaning the emergency resulted in minor or no medical conditions, or "Highly injured", meaning the resulting medical condition was dangerous, and "Deceased", which indicates that the victim(s) passed. In this work, our main task consists of

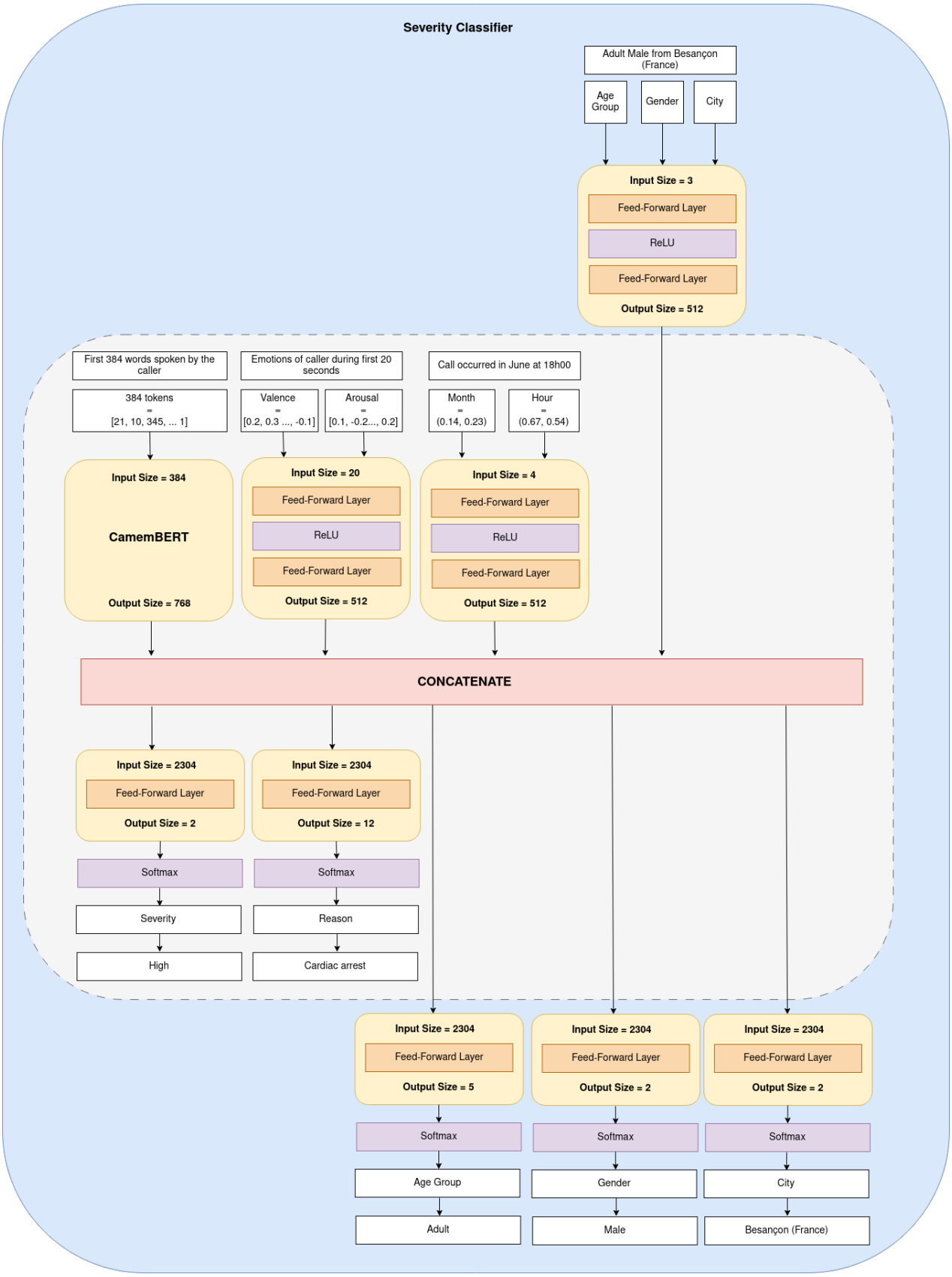


Figure 2: Severity classification network architecture in two settings: using personal information as input VS as output.

predicting the condition of the victim, which we consider to be equivalent to the "severity" of the emergency. So a prediction of a "Highly injured" state indicates that the emergency at-hand is urgent and requires the attention of the intervention team. Given that light injuries are the most common types of conditions, we group the "Highly injured" and "Deceased" categories into one, and randomly remove a portion of the most common class to balance out the dataset. In some cases, several victims with different levels of injuries might be involved in the same emergency (a fire for example). For these situations, we only include the call once, labeled by the most severe injury. An auxiliary task in the context of multi-task learning consists of classifying the call into a "reason" of emergency.

We expand the number of calls previously used (Abi Kanaan et al. (2023)) from 904 to 4167 recordings, made up of 49.96% "high severity" cases (2081 calls) and 50.03% "low severity" cases (2085 calls). The final recordings are then automatically transcribed into text using automatic speech recognition (Radford, Kim, Xu, Brockman, McLeavey and Sutskever (2023)). The average recording's length is 138.40 seconds, whilst the longest recording is 496.34 seconds long, and the shortest 6.69 seconds long. The average transcription's length is 2555 $\pm$46 words, whereas the minimum and maximum length is 110 and 8279 words respectively.

According to a confidentiality agreement that was signed with the SDIS 25, neither the dataset nor any of the models can be disclosed since doing so could expose the callers' personal information.

3.1.2. RECOLA database

To develop our speech emotion recognition model, we train a deep neural network using RECOLA (REmote COLlaborative and Affective) database (Ringeval, Sonderegger, Sauer and Lalanne (2013)). This dataset contains 46 different recordings in Frencg, where each recording is accompanied by its audio, video, electrocardiogram (ECG), and electrodermal activity (EDA). The recordings were collected following the collaboration of several participants in a task where they had to discuss how to survive in a disaster scenario. These conversations were annotated with arousal and valence values each 40 ms by 6 annotators. Since we only have access to the speech of callers in the context of this work, we only use the audio file of each recording.

3.2. Model Outputs

In the context of the multi-task learning approach, we select the following tasks as auxiliary outputs in regards to the severity output:

- *Reason of the call*: We theorize that the reason of the call is correlated with the severity of the emergency. A heart attack emergency for instance is usually more urgent and dangerous than a fall. We group the reasons of the call into 12 categories. Table 2 presents an overview of the different emergency incidents in the dataset, as well as the sample size of each class. The dataset can be considered imbalanced in regards to the number of samples for each emergency reason. However, since the classification of the reason remains a complementary task in this paper, we do not currently take actions in order to handle this imbalance, and leave this for a future work.
- *Age of the victim*: The age of the victim is not always mentioned in a call, as sometimes the urgency of the emergency makes it more difficult for the caller to accurately describe the situation. Moreover, some emergencies might involved more than one victim, such as a car accident emergency. Based on this, we don't aim to extract the age of the victim in the discussion, but rather classify the speech into one of 5 age groups (Table 3) (of Health (2023)).
- *Gender of the victim*: The gender detection task is reduced to a simple binary classification ("Male" or "Female").
- *City*: We theorize that the city where the emergency originates from might impact the outcome of the situation. For instance, some cities' roads might be more prone to road accidents than others. The dataset is highly imbalanced in regards to the city of the emergency, as most calls originate from one dominating city (Besançon, France). For this reason, we also reduce the city detection task into a binary classification, where the 0 label represents the dominating city, and 1 represents any other city.

Table 2
Emergency reasons sample size distribution in dataset.

Emergency Reason	Sample Size
Violence	77
Wounds/Trauma	776
Faintness	494
Fall	429
Public Road Accidents	479
Suicide Attempt	129
Respiratory Distress	314
Heart Failure	472
Delivery problems	24
Individual who is not answering calls	132
Fire	23
Others	818

Table 3
Victims' ages sample size distribution in dataset.

Age Range	Sample Size
0-1	17
2-12	238
13-17	205
18-64	2294
64+ and unknown ages	1413

Moreover, we compare the results obtained when including the age, gender, and city as auxiliary task outputs, as opposed to including these values as inputs to the network (see Fig. 2).

3.3. Speaker Identification

Based on the baseline emergency calls processing pipeline (Abi Kanaan et al. (2023)), the calls go through a speaker diarization block, in order to extract the caller's speech and discard any other intervening side, such as the operator. One limitation in this phase is the speaker identification process, which was completed manually in the previously described work. Once the speakers are separated, a re-identification of the speech signals linked to callers is required.

Given the nature of the conversations, the operator's speech is of an interrogating nature, where the same questions are typically repeated in most of the calls. Based on this, we take advantage of the manually labeled dataset to train a French BERT (Tenney et al. (2019)), CamemBERT (Martin et al. (2019)), to automatically distinguish between an operator's and a caller's speech. The best model was able to label 3263 calls with a 96.87% accuracy. Since the accuracy is not 100%, an imperfection is added to our dataset as 3.13% of the transcriptions are mis-labeled. We consider this a necessary trade-off as it enabled us to expand the size of our collection of transcriptions (a 460.95% increase in number of samples) to slightly less than five times the equivalent of that of the previous dataset.

3.4. Classification Model

In this section, we describe the various contributions that were made to the baseline classifier (as illustrated Fig. 2) to improve the accuracy on the severity classification. The outputs of four feature extractors, denoted as v_i, each one treating a different modality of data (text , emotions, time, and victim's personal information), were concatenated into one layer v, such as $v = (v_1, v_2, v_3, v_4)$. This layer in turn is followed by the output layers, which will predict the severity of the call, alongside several potential outputs (the reason of emergency, age, gender, and city of the victim).

3.4.1. Text Classification

We fine-tune the base version of the CamemBERT model with one classification layer on our dataset (Abi Kanaan et al. (2023)). We either pad or truncate the callers' speech transcriptions to match the maximum sequence length of 384 words, which was found to lead to a higher accuracy compared to the maximum of 512 supported by CamemBERT (more details in Section 4.1). This shows that, even though the average sequence is much longer than 384 words (see Section 3.1.1), the most relevant information for CamemBERT are found at the beginning of the caller's speech. The sequences are tokenized using the uncased CamemBERT tokenizer. Finally, attention masks are set to differentiate between real and padded tokens.

3.4.2. Speech Emotion Recognition

We train a deep neural network on the RECOLA database, and use the trained model to infer the emotions in the emergency calls, such as shown Fig. 1. Our primary goal is to extract the most relevant emotional features in the caller's voice, which would provide a more accurate idea of the situation's urgency and priority. For example, in a low-urgency situation, such as hitting a boar on the road, the caller would exhibit calmer emotions than one who's experiencing symptoms of a heart-attack. We only focus on the acoustic features to build the emotions classifier, as opposed to some works that have also used the linguistic features (Section 2.2). Although this has been proven to improve the network's performance, we leave this for a future extension of this work.

The network (illustrated in Fig. 3) is based on a state-of-the-art speech emotion recognition architecture (Zhao et al. (2019)). We first apply some pre-processing to the RECOLA audio files. We fragment them into 4-second long segments (and pad the fragments that are shorter than 4 seconds), with an overlap of 2 seconds between successive fragments. Each segment is re-sampled from 44100 Hz to 8000 Hz (similar to the emergency calls' speech rate), then converted to a Mel spectrogram (Smyth (2019)). The Mel spectrogram is an efficient method for audio feature extraction which mimics the way humans perceive sounds. This is achieved by adopting the Mel-scale which allows the distance between scales of pitches to be perceived in the same way by the listener. The process of extracting the Mel spectrogram for each audio segment is as follows:

- The segment is separated into windows of 2048 samples with a hop length of 512.
- Fast Fourier Transform (FFT) is applied on each window, which allows us to pass from the time domain to the frequency domain.
- The frequency spectrum is converted to the Mel-scale. It is separated into 128 frequency bands.
- Each window is decomposed using the frequencies in the Mel-scale.

As for the network's architecture, it consists of a collection of what its creators (Zhao et al. (2019)) call a "Local Feature Learning Block" (LFLB). The LFLB is used to extract local features for speech emotion recognition. It consists of one 2D convolutional layer, followed by a batch normalization layer (Ioffe and Szegedy (2015)), the ELU activation function (Rasamoelina, Adjailia and Sinčák (2020)), and a 2D max-pooling layer to reduce the dimensionality of the features. We use four LFLBs with 64 convolution kernels in the first two layers, and 128 convolution kernels in the last two layers. Similarily to one emotion recognition study in an emergency center (Deschamps-Berger et al. (2021)), we add a bidirectionnal LSTM layer with 32 units in order to allow the network to learn the temporal aspect of the signals.

The network is trained in a multi-task learning manner for the simultaneous prediction of the arousal and valence.

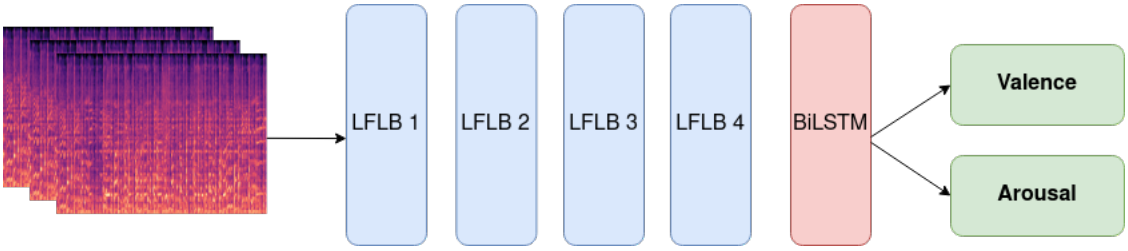


Figure 3: Speech emotion recognition network architecture.

Since dimensional speech emotion recognition is a regression task, we use a correlation-based loss function, the concordance correlation coefficient (CCC) (Lawrence and Lin (1989)).

We use the trained model to extract the callers' emotions based on the first 20 seconds of each caller speech (Fig. 1). We believe that the core emotional features reside in the beginning of the call, i.e., the first 20 seconds of the recordings, as the callers are usually less emotional towards the end of the call once they receive the operators' help. Calls that are shorter than 20 seconds are padded to match this duration. The calls undergo the same pre-processing steps as the RECOLA dataset. Once speech emotion recognition is applied on the Mel spectrograms, we obtain two vectors (the valence vector and the arousal vector) of 10 values each, representing the emotional value at the end of each 4-second long fragment. These vectors are later used as additional features to the severity classifier (Fig. 2).

3.4.3. Time-based Features

We extract each call's time-based features, mainly the month and the hour of the day (in a 24-hour format), during which the call occurred. We theorize that such features are highly correlated with the reason of the emergency and its outcome. For instance, a call that occurs at a late hour in the month of February, which is one of the coldest months in France, could likely be linked to a car accident emergency, whereas an emergency that occurs at noon during summer has a higher probability of being linked to a cardiac arrest, as such conditions commonly occur during hot weathers.

In order to represent the cyclic nature of these features, we opt for a cyclic-based representation (Chakraborty and Elzarka (2019)) as opposed to a classic one-hot encoding modeling. Such a representation reduces the input dimensionality on the one hand, as encoding the hours of the day for instance results in a 24 dimensionality vector. On the other hand, this approach also incorporates the cyclical continuity aspect of the time-based values. To model the time-based values in a cyclic representation, each value, which we denote as t, is reduced to a feature vector of two values $[x, y]$, using trigonometric functions. If we consider *max_value* equal to 12 when representing months, and equal to 24 when representing hours, the x and y of the features are computed in the following way:

$$x = \sin\left(\frac{2 * \pi * t}{max\_value}\right) \tag{1}$$

$$y = \cos\left(\frac{2 * \pi * t}{max\_value}\right) \tag{2}$$

3.5. Multi-task Learning

Multi-task learning can be implemented in one of two ways: either by hard-parameter sharing or through soft-parameter sharing (Caruana (1997)). In the hard-parameter sharing approach, which is the more commonly used approach, the hidden layers are shared among all tasks, while a few task-specific layers are kept. When using the soft-parameter sharing approach, a separate model is dedicated to each task. However, in order to minimize the distance between the parameters of the models, the latter are subjected to regularization during training. In this work, we opt for the hard-parameter sharing method, as this allows us to reduce resources consumption.

In this study, all tasks share the same inputs, but have distinct task-specific labels. We use the following loss function to optimize the model when including all the values described in Section 3.2 as auxiliary tasks:

$$Loss = Loss_{severity} + Loss_{reason} + Loss_{age} + Loss_{gender} + Loss_{city} \tag{3}$$

with *Loss* being the Negative log-likelihood function (as described in 4.1), whereas the following loss function is used when only including the reason of emergency as an auxiliary task:

$$Loss = Loss_{severity} + Loss_{reason} \tag{4}$$

4. Experiments and Evaluation

4.1. Experiments and Hyperparameter Selection

In terms of computational complexity, we completed the computations in this study on an NVIDIA Tesla V100 GPU with 32 GB of memory. We used the PyTorch framework

Table 4
Classification scores for the severity classification for the different combinations of features and tasks with a 95% confidence interval. Abbreviations: PI, stands for the victim's Personal Information; ⋄, indicates the score of the baseline (Orellana et al. (2020)) using text pre-processing methods from the same study; ⋆, indicates the baseline (Abi Kanaan et al. (2023)) score on the current dataset

Inputs				Auxiliary outputs		Accuracy	Recall	Precision	F_1-score
Text	Emotions	PI	Time	PI	Reason				
+	-	-	-	-	-	67.18 ±1.49%⋄	67.30 ±1.43%⋄	67.25 ±1.46%⋄	67.15 ±1.49%⋄
+	-	-	-	-	-	67.47 ±1.49%⋆	67.50 ±1.49%⋆	67.60 ±1.49%⋆	67.42 ±1.50%⋆
+	+	-	-	-	-	67.67 ±1.31%	67.55 ±1.39%	68.1 ±1.26%	67.36 ±1.52%
+	+	-	+	-	-	70.19 ±1.36%	70.19 ±1.34%	70.48 ±1.36%	70.07 ±1.37%
+	+	-	+	+	-	69.81 ±1.44%	69.77 ±1.43%	69.82 ±1.44%	69.77 ±1.43%
+	+	-	+	+	+	70.55 ±1.83%	70.51 ±1.80%	70.72 ±1.79%	70.45 ±1.83%
+	-	-	-	+	+	67.30 ±1.96%	67.27 ±1.98%	67.38 ±1.94%	67.22 ±1.99%
+	+	+	+	-	-	69.83 ±1.27%	69.76 ±1.24%	70.33 ±1.46%	69.59 ±1.27%
+	+	-	+	-	+	71.14 ±1.44%	71.08 ±1.42%	71.39 ±1.44%	71.01 ±1.46%
+	+	+	+	-	+	**71.27 ±1.63%**	**71.23 ±1.64%**	**71.40 ±1.64%**	**71.19 ±1.63%**

(Paszke, Gross, Massa, Lerer, Bradbury, Chanan, Killeen, Lin, Gimelshein, Antiga et al. (2019)) to implement the models, and split the dataset into 80% training (3333 samples) and 20% testing (833 samples). We trained all models using a 10-fold cross validation technique. The estimated training duration of a single model was 1 hour and 22 minutes for all 10 folds (or about 16 minutes per fold).

For the optimization of CamemBERT, we select our hyperparameters (Table 5) based on the range of recommended values in one study (Devlin, Chang, Lee and Toutanova (2018)). As for the maximum text sequence length parameter, we found that, surprisingly, setting its value to the maximum supported number of 512, did not have a significant impact of the network's accuracy, compared to using a lower value of 384. For this reason, we select the lower value of 384, as it enables us to complete the training faster, without impacting the performance.

Each one of the remaining networks, i.e., emotions, age, time, gender, and city networks, are made up of an input layer of size 512, followed by the ReLU activation function (Rasamoelina et al. (2020)), and another hidden linear layer of size 512. Using a linear learning rate scheduler with warmup further improved the model's performance. This allows the learning rate to first linearly increase from 0 to the initial learning rate of the optimizer during the warmup period, to then linearly decrease from its initial value to 0.

We found that training for 12 epochs using a batch size of 8 led to the highest accuracies. We used the Adam optimizer (Kingma and Ba (2014)) to optimize the network, with a learning rate value of 3E-5, and an epsilon of 1E-7. We use the Negative log-likelihood (NLLL) as our loss function Contributors (2023).

Table 5
List of hyperparameters used to train the severity classifier.

Hyperparameter	Value
Sequence length	384
Last CamemBERT layer learning rate	5E-5
Concatenated network learning rate	3E-5
Epsilon	1E-7
Neural networks Layers size	512
Batch size	8
Num. of epochs	12

4.2. Results

4.2.1. Evaluation metrics

For the severity classification task, we evaluate the models' performance using the metrics of Accuracy (Eq. 5), Precision (Eq. 6), Recall (Eq. 7), and F_1-score (Eq. 8). These metrics are calculated based on the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

True positives are instances of a class that were correctly predicted as belonging to this class, whereas true negatives are when the instances are correctly predicted as belonging to the other class. False positives are when the classifier incorrectly predicts that an instance belong to a class, whereas false negatives are when the instance is incorrectly predicted to belong to the other class.

The accuracy metric therefore (Eq. 5) indicates the proportion of all instances that are classified correctly, and is a good indicator of the model's overall performance. The precision (Eq. 6) is an adequate metric to see how often our model is predicting false positives. In our case, the lower the precision rate, the more frequently the model is predicting the "high severity" class when it shouldn't. The recall is a more relevant metric to our work, as it is associated with the prediction of false negatives. In our study, the lower the recall rate, the more likely the model is not predicting the "high severity" classes when it should have. The F_1-score (Eq. 8) is a combination of the recall and precision scores, and similarly to the accuracy, shows an overall idea of the model's performance. It is often used when the dataset is

imbalanced, which makes it a good metric to evaluate the performance on the reason classification task.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{8}$$

4.2.2. Evaluation results

Table 6
Mean confusion matrix of the severity classification of the best performing model with victims' personal information as input.

			Support	Recall
High Severity	TP = 292	FP = 120	412	71.04%
Low Severity	FN = 119	TN = 302	421	**71.56%**

Table 7
Mean confusion matrix of the severity classification of the best performing model without victims' personal information as input.

			Support	Recall
High Severity	TP = 275	FP = 137	412	**72.75%**
Low Severity	FN = 103	TN = 318	421	69.89%

We report in Table 4 the mean classification scores with a 95% confidence interval for the severity classification task. We obtain it by performing 10-fold cross-validation runs. We include the evaluation metrics described in Section 4.2.1 for several combinations of inputs and auxiliary tasks. This allows us to demonstrate the impact of each one of these inputs and tasks on the scores.

Moreover, we compare our models' performance on the severity classification task to that of two baseline classifiers:

- The CamemBERT classifier used in a previous work (**?**). We retrain this classifier on our enlarged dataset of emergency calls transcriptions.
- A work (Orellana et al. (2020)) that is concerned with the classification of high-priority calls. This study is the most similar to ours as the high-severity calls in our application can also be considered "high-priority" calls. We reproduced the text pre-processing code and trained the SVM classifier with Radial Basis Function (RBF) kernel on our dataset. We kept the default values of the gamma and C parameters suggested by sklearn library (skl (2023)). We found that the default values resulted in better scores in our case compared to the values recommended by the baseline (Orellana et al. (2020)))

We first focus on the accuracy metric as it gives us an overall idea of a model's performance. The first baseline (Orellana et al. (2020)) results in the lowest scores (67.18% accuracy) on the severity classification task. The baseline CamemBERT model (Abi Kanaan et al. (2023)) on the other hand results in a higher accuracy of 67.47%. This confirms the BERT-based models' (Devlin et al. (2018), Tenney et al. (2019)) robustness on text classification tasks, which can lead to decent results with minimal to no data pre-processing..

We can see that concatenating CamemBERT's output with the emotions network's output slightly improved the accuracy (to 67.67%). This is an indication that the emotions of the caller do not always reflect the severity of an emergency due to many reasons. Some situations that can be perceived as non-urgent can result in dangerous outcomes, if not handled properly. Moreover, there are many cases where the caller is distantly or not related to the victim (e.g., in the case where an intoxicated individual is found on the streets). In such cases, the caller is not expected to exhibit many emotions.

On the other hand, the addition of the time-based features network improved the accuracy more significantly. This proves the correlation between the severity of an emergency and its time of day and year. Using the previously mentioned network to train on the auxiliary tasks of age, gender, and city detection, the accuracy decreased compared to only training on the main task. This means that training the network on determining the personal information of the victim, i.e., the age, gender, and location, is a difficult task for the network. This may be due to the imbalance in the dataset regarding these labels (see Table 3), which would make it more difficult for the network to learn the correlations properly. However, it is with the inclusion of the reason of emergency classification task that we were able to increase the accuracy more significantly.

The highest score was obtained by using the age, gender, and city values as inputs as opposed to using them in the context of multi-task learning. This accuracy (71.27%) was slightly higher than the accuracy obtained without using this information as input (71.14%). This shows that the high level of correlation between the classification of the reason and its severity was enough to estimate the severity with a very close accuracy compared to when having access to the victim's age, gender, and city. It also proves the effectiveness of the multi-task learning approach in this context. These results can be interpreted in the following way:

- In some emergency situations where the caller is apparently agitated, operators would often prioritize collecting details on the emergency itself, rather than

wasting time on collecting the victim's personal information, i.e., age and gender. However, our experiments prove that the availability of this information slightly increases the chances of accurately estimating the priority of the emergency. As such, it is crucial that operators attempt to collect this information at the beginning of the call.

- In other cases, the age, gender, and city information cannot be explicitly inputted. Such a scenario can take place when the emergency center is overloaded, and the developed tool is used to automatically assign priorities based on some of the caller's speech. In such a case, the system would attempt to infer this information if needed, and determine the severity and the reason simultaneously, before dispatching this report to the operator

We can conclude that our approach effectively increased the accuracy of the baseline CamemBERT classifier by 5.15% when the personal information was unavailable (to 71.15%), and by 5.33% when such information was available (to 71.27%).

As for the remaining metrics, the results indicate that the model with the highest accuracy (71.27%) has also the best recall, precision, and F_1-score. It has the best precision score (71.40%), meaning it predicts false severe cases the least among all models. However, since the precision is slightly higher than the recall rate (71.23%), this means that the model tends to underestimate emergencies and focuses more on avoiding false negatives. This is further confirmed in Tables 6 and 7, which show the number of false negatives obtained on each class. Interestingly, the recall rate for the "High Severity" class is higher (72.75% > 71.04%) without the victim's information.

Table 8
Max precision, recall, and F_1-score obtained on each reason of emergency.

Reason	Precision	Recall	F_1-score	Support
Violence	75%	20%	32%	15
Wounds/Trauma	67%	77%	72%	155
Discomfort	53%	39%	45%	99
Fall	60%	62%	61%	86
Public road accident	80%	91%	85%	96
Suicide Attempt	73%	62%	67%	26
Respiratory distress	62%	65%	64%	63
Heart failure	65%	54%	59%	94
Delivery problems	100%	40%	57%	5
Individual not answering	60%	69%	64%	26
Fire	100%	60%	75%	5
Other	53%	58%	55%	164

Table 8 shows the maximum F_1-score obtained one each one of the reasons of emergencies. The scores show that the minority classes (e.g. Fire, Individual not answering), do not necessarily have the lowest scores. Some reasons with a high number of samples, such as "Heart Failure" and "Discomfort" are rather difficult to detect. Such reasons may be associated with a wide range of symptoms, an might not be immediately diagnosed. This is not the case for other more obvious emergencies, like "Public road accident" or "Fire".

We illustrate the mean confusion matrix of the reason of emergency classification task in Fig. 4 when including the personal information as input. Some difficulties were found in distinguishing the following categories of emergencies:

- "Violence" and "Wounds/Trauma".
- "Public road accident" and "Wounds/Trauma".
- "Discomfort" and "Fall"/"Public road accident"/"Respiratory distress"/"Heart Failure".
- "Respiratory distress", "Discomfort" and "Heart Failure".

Table 9
Scores obtained on the auxiliary classification tasks.

Task	Metric	Score
Age	Macro F_1	49%
City	Accuracy	85%
Gender	Accuracy	90%

As for the remaining auxiliary tasks of age, city, and gender detection, we summarize the classification scores of the best model in Table 9. It is clear that the network faces difficulties when determining the age group, as in some cases, the caller does not explicitly mention an age for the victim. In addition, some age groups are less prominent than others, such as the age group of 0-1 (see Table 3) and therefore constitute minority classes. The tasks of city and gender detection remain relatively easy binary classification tasks. The gender can be determined from the pronouns that the caller is using, and the caller will mention their location so that they can receive assistance.

4.3. Predictions Explainability

In order to gain a better understanding of the impact of each one of the given inputs, we use SHAP (SHapley Additive exPlanations) (Lundberg and Lee (2017)), a library that offers both local and global explainability for machine learning models. SHAP's results are based on Shapley values, a game theory concept.

In this study, we consider the global explainability aspect, as we seek to understand the model's decision making overall, rather than on specific samples. We create a model explainer using 100 randomly selected samples from the training set. We then plot the SHAP values for 25 random samples of each type of prediction (see Table 10) from the test set using the explainer.

The SHAP values demonstrate the contribution of the "Month" and "Gender" features, the highest among all features for all types of outputs. The "Age" feature is the

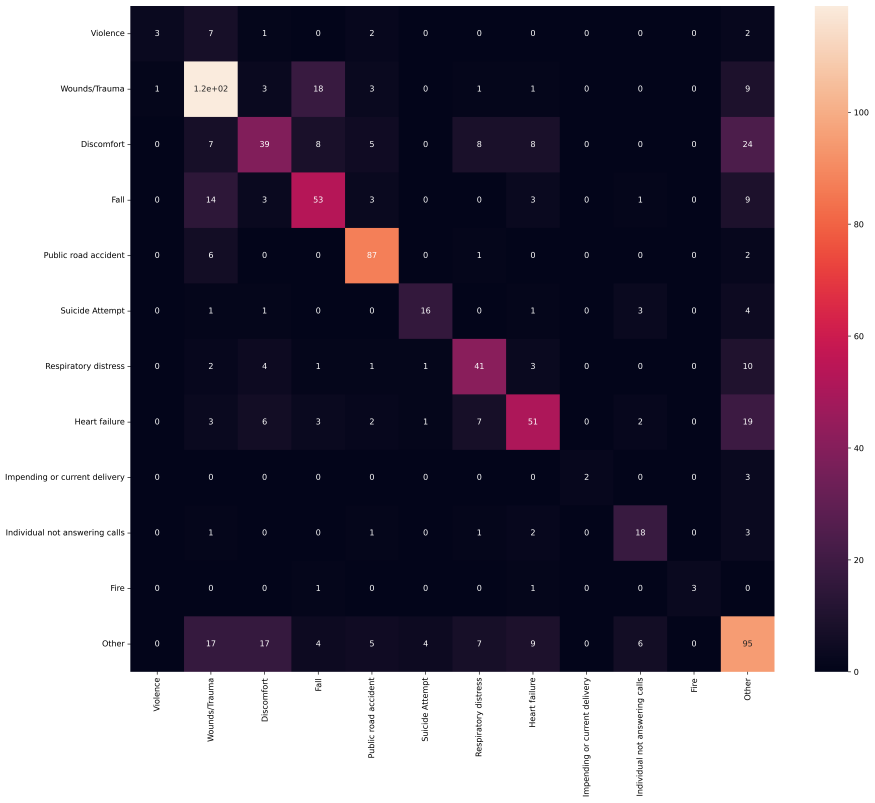
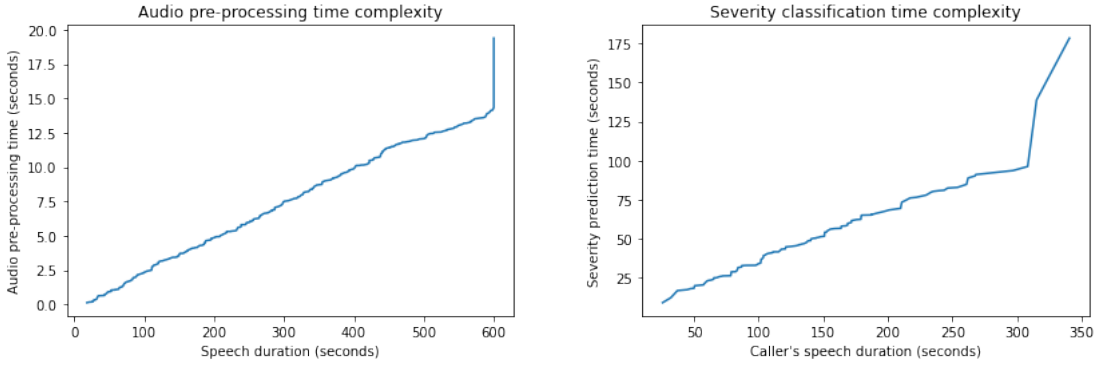


Figure 4: Reason of emergency classification mean confusion matrix.



(a) Audio pre-processing task time analysis

(b) Severity classification task time analysis

Figure 5: Time complexity analysis of the system on the audio pre-processing and severity classification tasks.

third most contributing feature to both "Highly Injured" and "Lightly Injured" outputs. Interestingly, we can note the "Hour" feature is a bigger contributor to the model's faulty predictions than the age and city of the victim.

4.4. Performance Analysis

We conduct a performance analysis on 124 randomly selected calls to evaluate the time required, in seconds, to analyze an emergency call from start to end using our approach. We group the tasks involved into two broad groups:

- *Audio pre-processing*: this involves the voice activity detection and speaker diarization tasks conducted on the calls.
- *Severity classification*: the performance of these tasks is evaluated on the caller's speech obtained from the previous phases. They include the speech transcription, the caller identification, the emotions extraction, features encoding, and the final severity prediction task.

The analysis shows that on average, the audio pre-processing task requires 7.63 $\pm$0.42 seconds, whereas the

Table 10
The five most impactful input features for each output, in descending order based on the mean of the absolute values of SHAP.

Output	Features
Lightly Injured	**Month**
	Gender
	Age
	City
	Hour
Highly Injured	**Month**
	Gender
	Age
	Hour
	City
Mistaking Highly Injured for Lightly Injured	**Month**
	Gender
	Hour
	Age
	City
Mistaking Lightly Injured for Highly Injured	**Month**
	Gender
	Hour
	City
	Age

severity classification requires 48.69 ±3.64 seconds. If we don't consider the audio pre-processing phase (in the case where no operator is involved and the caller is giving a first description of their emergency), we can consider that it takes 48.69 seconds to determine the severity of an average caller's speech that lasts 137.14 ±11.68 seconds. These durations vary depending on the call's duration, as illustrated in Fig. 5.

These results are highly promising for a future real-time implementation of this system, as no model inference or hardware optimizations were made yet in the current work.

5. Discussion and Limitations

5.1. Discussion

Based on our experiments, we can conclude that our approach allows for the classification of an emergency call's severity with a 71.14% accuracy when functioning autonomously (no operator intervention). This score is further improved (71.27%) when an operator is able to intervene and indicate the victim's personal information as additional inputs. If we reflect these results to a real-life scenario, the operator using this tool would be able to correctly predict the level of injury of 71/100 emergency victims, and therefore undertake the necessary procedures to avoid these severe injuries. Moreover, the predicted severity could allow for the enhancement of the call center's queuing system. During the periods where operators are overloaded and the callers' waiting time is increased, the caller would briefly describe their situation to the system, thereby enabling the inference of a level of urgency, which can be used to reorder the waiting queue from most to least urgent. There clearly is room for improvement of the score to correctly prioritize a bigger number of callers. Nevertheless, since the developed system's intended use is to assist (and not replace) emergency center operators in their evaluation of each situation, we hope that the severity and reason of emergency predictions can be used as reference in confusing situations.

5.2. Limitations

One aspect of this work that can be considered as a limitation is that it does not currently function in a real-time setting. The current system is a proof-of-concept with no hardware or software optimizations. As it is shown in Section 4.4, an average 137-seconds long emergency call would require an additional 56.32 seconds (7.63 + 48.69 seconds), for the system to be able to infer the severity. This suggests that the operator is not currently capable of assessing the situation in real-time with the developed tool, and would have to wait an additional minute for a prediction. If we take into account the fact that no optimizations were made to the system in terms of computational complexity, the inference time of 1 minute seems to be an acceptable duration. Nonetheless, some processes can be further improved in a future work, such as for instance the speech emotion recognition and the speech transcription, which can be executed in parallel in real-time as the call is going. It is also worth mentioning that our current approach to speech emotion recognition can be further improved by extracting emotional information from the textual features alongside the acoustic features. In some cases where the callers are not exhibiting their real emotions (such as when attempting to hide their emotions, or exaggerating them), or when the callers are not related to the victim, analyzing the text for alternative emotional cues can be helpful. One approach to this (Deschamps-Berger, Lamel and Devillers (2023)) would be using pre-trained transformers, one on each type of data, and then fusing the extracted vectors to detect the emotions related to a call.

6. Conclusion

In this study, we implemented several improvements and extensions to an emergency calls analysis tool. Specifically, we investigated the effect of combining the CamemBERT-based calls transcriptions classifier with multiple feature extractors of different modalities of data.

The results show that time-based features and the emergency's victim's personal information improve the accuracy of the emergency severity classification the most, while the inclusion of the emotions exhibited by the caller slightly increases the accuracy score.

Furthermore, we explored the use of a multi-task learning approach in the training of the network. Our experiments showed that such an approach can effectively further improve the accuracy, as we included tasks that were highly correlated with the severity classification task. These tasks

included, on one hand, the classification of the call into a reason of emergency (e.g., cardiac arrest, accident, etc.). On the other hand, we modified the network's architecture to detect the personal information as an additional auxiliary task, as opposed to using these data as input.

With the implementation of the described methods, our classifier predicted the severity of 833 emergency calls with a 71.14% accuracy, a 5.15% increase over the baseline classifier, when the personal information was unavailable. This score increased to 71.27%, when such information was available, a 5.33% improvement over the baseline, Such a tool can be considered useful when used in an autonomous way, without human intervention, to get a first evaluation of the emergencies.

As future work, we aim to handle some of the limitations in this study. First, it is worth delving into the implementation of a questions generation algorithm, as emergency call center operators would highly benefit from such suggestions in tough situations. This would be relatively easy to incorporate with the use of the more recent robust large language models (LLMs), such as LLaMA (Meta (2023)). In addition, we can employ some state-of-the-art methods (Mamdouh Farghaly and Abd El-Hafeez (2023), Mamdouh Farghaly and Abd El-Hafeez (2022)) for feature selection to extract more meaningful and less redundant features from the texts. Moreover, our performance analysis showed that the system requires further optimization, so as to improve its performance in a real-time setting. Finally, we aim to improve our speech emotion recognition inference model by fusing textual features with acoustic features (Deschamps-Berger et al. (2022), Deschamps-Berger et al. (2023)).

Acknowledgments

We would like to thank the SDIS 25 (Service Départemental d'Incendie et de Secours du Doubs) for their invaluable assistance, for sharing their needs and ideas, and for trusting us with the dataset used in this work. Computations have been performed on the supercomputer facilities of the "Mésocentre de calcul de Franche-Comté". This work is (partially) supported by the EIPHI Graduate School (contract ANR-17-EURE-0002). It is also supported by the Safar scholarship program, co-funded by the French embassy in Lebanon, and the Agence Universitaire de la Francophonie AUF-BMO, in the framework of the Interuniversity Scientific Cooperation Project WeBel.

CRediT authorship contribution statement

Marianne Abi Kanaan: Conceptualization of this study, Methodology, Software, Writing - Original draft preparation, Data curation. **Jean-François Couchot:** Conceptualization, Supervision, Writing - review & editing, Data curation, Resources, Validation. **Christophe Guyeux:** Conceptualization, Writing - review & editing, Methodology, Project administration, Investigation, Validation. **David Laiymani:** Software, Methodology, Writing - review & editing, Resources, Supervision, Validation. **Talar Atechian:** Supervision, Funding acquisition. **Rony Darazi:** Supervision, Funding acquisition.

References

, 2023. Corti.ai. https://www.corti.ai/. Accessed: 2023-11-07.

, 2023. Scikit-learn: Machine learning in python. https://scikit-learn.org/stable/. Accessed: 2023-11-09.

Abi Kanaan, M., Couchot, J.F., Guyeux, C., Laiymani, D., Atechian, T., Darazi, R., 2023. A methodology for emergency calls severity prediction: From pre-processing to bert-based classifiers, in: Artificial Intelligence Applications and Innovation, Springer Nature Switzerland. pp. 329–342.

Akçay, M.B., Oğuz, K., 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116, 56–76.

Ayache, F., Alti, A., 2020. Performance evaluation of machine learning for recognizing human facial emotions. Revue d'Intelligence Artificielle 34.

Blomberg, S.N., Folke, F., Ersbøll, A.K., Christensen, H.C., Torp-Pedersen, C., Sayre, M.R., Counts, C.R., Lippert, F.K., 2019. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation 138, 322–329.

Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., Gill, M.P., 2020. pyannote.audio: neural building blocks for speaker diarization, in: ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al., 2005. A database of german emotional speech., in: Interspeech, pp. 1517–1520.

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, 335–359.

Caruana, R., 1997. Multitask learning. Machine learning 28, 41–75.

Chakraborty, D., Elzarka, H., 2019. Advanced machine learning techniques for building performance simulation: a comparative analysis. Journal of Building Performance Simulation 12, 193–207.

Contributors, P., 2023. Nlll. https://pytorch.org/docs/stable/generated/torch.nn.NLLLoss.html. Accessed: 2023-10-31.

Deschamps-Berger, T., Lamel, L., Devillers, L., 2021. End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings, in: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE. pp. 1–8.

Deschamps-Berger, T., Lamel, L., Devillers, L., 2022. Investigating transformer encoders and fusion strategies for speech emotion recognition in emergency call center conversations., in: Companion Publication of the 2022 International Conference on Multimodal Interaction, pp. 144–153.

Deschamps-Berger, T., Lamel, L., Devillers, L., 2023. Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Gil-Jardiné, C., Chenais, G., Pradeau, C., Tentillier, E., Revel, P., Combes, X., Galinski, M., Tellier, E., Lagarde, E., 2021. Trends in reasons for emergency calls during the covid-19 crisis in the department of gironde, france using artificial neural network for natural language classification. Scandinavian journal of trauma, resuscitation and emergency medicine 29, 1–9.

Goncharov, M., Pisov, M., Shevtsov, A., Shirokikh, B., Kurmukov, A., Blokhin, I., Chernina, V., Solovev, A., Gombolevskiy, V., Morozov, S., et al., 2021. Ct-based covid-19 triage: Deep multitask learning improves joint identification and severity quantification. Medical image analysis 71, 102054.

of Health, N.I., 2023. Age | national institutes of health (nih). https://www.nih.gov/nih-style-guide/age. Accessed: 2023-08-01.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International conference on machine learning, pmlr. pp. 448–456.

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, M.P., Gehringer, D., 2021. All-in-one: Multi-task learning bert models for evaluating peer assessments.

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

Kumar, S., Haq, M.A., Jain, A., Jason, C.A., Moparthi, N.R., Mittal, N., Alzamil, Z.S., 2023. Multilayer neural network based speech emotion recognition for smart assistance. Computers, Materials & Continua 75.

Lawrence, I., Lin, K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics , 255–268.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems 30.

Mamdouh Farghaly, H., Abd El-Hafeez, T., 2022. A new feature selection method based on frequent and associated itemsets for text classification. Concurrency and Computation: Practice and Experience 34, e7258.

Mamdouh Farghaly, H., Abd El-Hafeez, T., 2023. A high-quality feature selection method based on frequent and correlated items for text classification. Soft Computing , 1–16.

Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de La Clergerie, É.V., Seddah, D., Sagot, B., 2019. Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 .

Meta, 2023. Introducing llama 2. https://ai.meta.com/llama/. Accessed: 2023-08-23.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J., 2021. Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR) 54, 1–40.

Omar, A., Abd El-Hafeez, T., 2023. Quantum computing and machine learning for arabic language sentiment classification in social media. Scientific Reports 13, 17305.

Orellana, M., Trujillo, A., Acosta, M.I., 2020. A methodology to predict emergency call high-priority: Case study ecu-911, in: 2020 Seventh International Conference on eDemocracy & eGovernment (ICEDEG), IEEE. pp. 243–247.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Perez-Toro, P.A., Vasquez-Correa, J.C., Bocklet, T., Noth, E., Orozco-Arroyave, J.R., 2021. User state modeling based on the arousal-valence plane: applications in customer satisfaction and health-care. IEEE Transactions on Affective Computing .

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR. pp. 28492–28518.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training.

Rasamoelina, A.D., Adjailia, F., Sinčák, P., 2020. A review of activation function for artificial neural network, in: 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI), IEEE. pp. 281–286.

Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D., 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions, in: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE. pp. 1–8.

Smyth, T., 2019. The mel scale. http://musicweb.ucsd.edu/~trsmyth/pitch2/Mel_Scale.html. Accessed: 2022-09-30.

Tenney, I., Das, D., Pavlick, E., 2019. Bert rediscovers the classical nlp pipeline. arXiv preprint arXiv:1905.05950 .

Trujillo, A., Orellana, M., Acosta, M.I., 2019. Design of emergency call record support system applying natural language processing techniques, in: Information and Communication Technologies of Ecuador (TIC. EC). Springer, pp. 53–65.

Zhao, J., Mao, X., Chen, L., 2019. Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomedical signal processing and control 47, 312–323.