Rui Miguel Boneco Novais

A Framework for Emotion and Sentiment Predicting Supported in Ensembles

UNIVERSIDADE DO ALGARVE
Instituto Superior de Engenharia

2022

Rui Miguel Boneco Novais

A Framework for Emotion and Sentiment Predicting Supported in Ensembles

Master's Dissertation in Electrical and Computer Engineering

Work done under the supervision of:
Professor Pedro Jorge Sequeira Cardoso
Professor João Miguel Fernandes Rodrigues

UNIVERSIDADE DO ALGARVE
Instituto Superior de Engenharia

Faro, September of 2022

A Framework for Emotion and Sentiment Predicting Supported in Ensembles

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

I hereby declare to be the author of this work, which is original and unpublished. Authors and works consulted are properly cited in the text and included in the reference list.

(Rui Novais)

ABSTRACT

Humans are prepared to comprehend each other's emotions through subtle body movements or facial expressions; using those expressions, individuals change how they deliver messages when communicating between them. Machines, user interfaces, or robots need to empower this ability, in a way to change the interaction from the traditional "human-computer interaction" to a "human-machine cooperation", where the machine provides the "right" information and functionality, at the "right" time, and in the "right" way. This dissertation presents a framework for emotion classification based on facial, speech, and text emotion prediction sources, supported by an ensemble of open-source code retrieved from off-the-shelf available methods. The main contribution is integrating outputs from different sources and methods in a single prediction, consistent with the emotions presented by the system's user.

For each different source, an initial aggregation of primary classifiers was implemented: for facial emotion classification, the aggregation achieved an accuracy above 73% in both FER2013 and RAF-DB datasets; For the speech emotion classification, four datasets were used, namely: RAVDESS, TESS, CREMA-D, and SAVEE. The aggregation of primary classifiers, achieved for a combination of three of the mentioned datasets results above 86 % of accuracy; The text emotion aggregation of primary classifiers was tested with one dataset called EMOTIONLINES, the classification of emotions achieved an accuracy above 53 %. Finally, the integration of all the methods in a single framework allows us to develop an emotion multi-source aggregator (EMsA), which aggregates the results extracted from the primary emotion classifications from different sources, such as facial, speech, text etc. We describe the EMsA and results using the RAVDESS dataset, which achieved 81.99% accuracy, in the case of the EMsA using a combination of faces and speech. Finally, we present an initial approach for sentiment classification.

Keywords: Facial Emotions, Speech Emotions, Text Emotions, Sentiment Classification, Ensembles, Machine Learning.

RESUMO

Os humanos estão preparados para compreender as emoções uns dos outros por meio de movimentos subtis do corpo ou expressões faciais; i.e., a forma como esses movimentos e expressões são enviados mudam a forma de como são entregues as mensagens quando os humanos comunicam entre eles. Máquinas, interfaces de utilizador ou robôs precisam de potencializar essa capacidade, de forma a mudar a interação do tradicional "interação humano-computador" para uma "cooperação homem-máquina", onde a máquina fornece as informações e funcionalidades "certas", na hora "certa" e da maneira "certa".

Nesta dissertação é apresentada uma estrutura (um ensemble de modelos) para classificação de emoções baseada em múltiplas fontes, nomeadamente na previsão de emoções faciais, de fala e de texto. Os classificadores base são suportados em código-fonte aberto associados a métodos disponíveis na literatura (classificadores primários). A principal contribuição é integrar diferentes fontes e diferentes métodos (os classificadores primários) numa única previsão consistente com as emoções apresentadas pelo utilizador do sistema. Neste contexto, salienta-se que da análise ao estado da arte efetuada sobre as diferentes formas de classificar emoções em humanos, existe o reconhecimento de emoção corporal (não considerando a face). No entanto, não foi encontrado código-fonte aberto e publicado para os classificadores primários que possam ser utilizados no âmbito desta dissertação. No reconhecimento de emoções da fala e texto foram também encontradas algumas dificuldades em encontrar classificadores primários com os requisitos necessários, principalmente no texto, pois existem bastantes modelos, mas com inúmeras emoções diferentes das 6 emoções básicas consideradas (tristeza, medo, surpresa, repulsa, raiva e alegria). Para o texto ainda possível verificar que existem mais modelos com a previsão de sentimento do que de emoções. De forma isolada para cada uma das fontes, i.e., para cada componente analisada (face, fala e texto), foi desenvolvido uma *framework* em Python que implementa um agregador primário com *n* classificadores primários (nesta dissertação considerou-se *n* igual 3). Para executar os testes e obter os resultados de cada agregador primário é usado um *dataset* específico e é enviado a informação do *dataset* para o agregador. I.e., no caso do agregador facial é enviado uma imagem, no caso do agregador da fala é enviado um áudio e no caso do texto é enviado a frase para a correspondente *framework.*

Cada *dataset* usado foi dividido em ficheiros treino, validação e teste. Quando a *framework* acaba de processar a informação recebida são gerados os respetivos resultados, nomeadamente: nome do ficheiro/identificação do input, resultados do primeiro classificador primário, resultados do segundo classificador primário, resultados do terceiro classificador primário e *ground-truth* do *dataset*. Os resultados dos classificadores primários são depois enviados para o classificador final desse agregador primário, onde foram testados quatro classificadores: (a) **voting**, que, no caso de *n* igual 3, consiste na comparação dos resultados da emoção de cada classificador primário, i.e., se 2 classificadores primários tiverem a mesma emoção o resultado do **voting** será esse, se todos os classificadores tiverem resultados diferentes nenhum resultado é escolhido. Além deste "classificador" foram ainda usados (b) ***Random Forest***, (c) ***Adaboost*** e (d) ***MLP*** (*multiplayer perceptron*). Quando a *framework* de cada agregador primário foi concluída, foi desenvolvido um super-agregador que tem o mesmo princípio dos agregadores primários, mas, agora, em vez de ter os resultados/agregação de apenas 3 classificadores primários, vão existir *n* × 3 resultados de classificadores primários (*n* da face, *n* da fala e *n* do texto).

Relativamente aos resultados dos agregadores usados para cada uma das fontes, face, fala e texto, obteve-se para a classificação de emoção facial uma precisão de classificação acima de 73% nos *datasets* FER2013 e RAF-DB. Na classificação da emoção da fala foram utilizados quatro *datasets*, nomeadamente RAVDESS, TESS, CREMA-D e SAVEE, tendo que o melhor resultado de precisão obtido foi acima dos 86% quando usado a combinação de 3 dos 4 *datasets*. Para a classificação da emoção do texto, testou-se com o um *dataset* EMOTIONLINES, sendo o melhor resultado obtido foi de 53% (precisão).

A integração de todas os classificadores primários agora num único *framework* permitiu desenvolver o agregador multi-fonte (*emotion multi-source aggregator* - EMsA), onde a classificação final da emoção é extraída, como já referido da agregação dos classificadores de emoções primárias de diferentes fontes. Para EMsA são apresentados resultados usando o *dataset* RAVDESS, onde foi alcançado uma precisão de 81.99 %, no caso do EMsA usar uma combinação de faces e fala. Não foi possível testar EMsA usando um *dataset* reconhecido na literatura que tenha ao mesmo tempo informação do texto, face e fala. Por último, foi apresentada uma abordagem inicial para classificação de sentimentos.

Keywords: Emoções Faciais, Emoções da Fala, Emoções do Texto, Sentimento, Ensembles, Aprendizagem Automática.

ACKNOWLEDGEMENTS

For the successful completion of this master's thesis, there were several participants that I would like to thank: To my parents and my brother who has always supported and encouraged me throughout my academic life, especially for the completion of the dissertation.

To the guidance of professors João Rodrigues and Pedro Cardoso who guided me in the dissertation, and who were always available to support its conclusion, without them this thesis would not have been completed.

Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AAC	Advanced Audio Coding
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
biLM	Bidirectional language model
CAER-NET	Context-Aware Emotion Recognition Networks
CaFE	Canadian French Emotional Speech Dataset
CNN	Convolutional Neural Network
CREMA-D	Crowd-sourced Emotional Multimodal Actors Dataset
C-RNN	Continuous Recurrent Neural Network
DNN	Deep Neural Network
ELMo	Embeddings from Language Models
Emo-DB	Berlin Emotional Speech Database
EMsA	Emotion Multi-Source Aggregator
ERC	Emotional Recognition in Conversation
ESC-50	Dataset for Environmental Sound Classification
FER	Facial Expression Recognition
FER2013	Facial Expression Recognition 2013 Dataset
GloVe	Global Vectors for Word Representation
GPT-2	Generative Pre-trained Transformer 2
GRU	Gated Recurrent Unit
HCI	Human-Computer Interaction
HCII	International Conference on Human-Computer Interaction

HMC	Human-Machine Cooperation
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LBP	Local Binary Patterns
LHC	Local (multi) Head Channel
LLDS	Low-level Descriptors
LSTM	Long short-term memory
MELD	Multimodal EmotionLines Dataset
MFCCs	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Multiplayer Perceptron
MMER	MultiModal-Emotion-Recognition
MOOCs	Massive Open Online Courses
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neural Network
OAENet	Oriented Attention Ensemble for Accurate Facial Expression Recognition
RAF-DB	Real-world Affective Faces Database
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
RMS	Root-Mean-Square
SAR	Socially Assistive Robot
SAVEE	Surrey Audio-Visual Expressed Emotion
SEDC	Speech Emotion Detection Classifier
SER	Speech Emotion Recognition

SEWA	Audio-Visual Emotion and Sentiment Research in the Wild
SVM	Support Vector Machines
TESS	Toronto Emotional Speech Set
TF-IDF	Term Frequency-Inverse Document Frequency
TIMIT	Texas Instruments/Massachusetts Institute of Tecnology
UWA	Unweighted Accuracy
VGG-16	Convolutional Neural Network
WA	Weighted Accuracy

1 INTRODUCTION

Abstract. Emotion and sentiment analysis methods are the automated processes of analyzing information to determine the emotion or sentiment expressed by the user. The present chapter introduces the subject of study and presents the goal of the dissertation, as well as the main contributions. Finally, it is introduced the contents of the following chapters.

Keywords: Facial Emotions, Ensembles, Machine Learning.

1.1 Contextualization

Emotion and sentiment analysis methods are the automated processes of analyzing information to determine the emotion (e.g., happiness, sadness, fear, surprise, disgust, anger, and neutral) or sentiment (e.g., positive, negative, and neutral) expressed by the user. The sentiment influences the emotion, and the emotions influence the sentiment. Humans are prepared to comprehend each other's emotions through subtle body movements, facial expressions, the way they speak, or simply by the tone of voice. They use this capacity when communicating between them, changing the way they pass the message based on those responses/emotions/sentiments. Humans can express affection through facial, vocal, or gestural behaviours. E.g., affect is the designation of a result of an emotion of a human being who has had an interaction of stimuli, referring to mental counterparts of the internal bodily representations associated with emotions [1].

We are living in the so-called Information Society, Society 4.0. However, we are starting to notice that the cross-sectional sharing of knowledge is not enough. So, in Japan appeared a new designation, Society 5.0, which should be one that "through the high degree of merging between cyberspace and physical space, will be able to balance economic advancement with the resolution of social problems by providing goods and services that granularly address manifold latent needs regardless of locale, age, sex, or language" [2]. Simplifying, Society 5.0 is a super-smart people-centric society.

To achieve this degree of development, one of the keys is to empower machines, user interfaces, or robots with the same communication capabilities that humans have to communicate between them. This changes the interaction between machines and humans from the traditional "human-computer interaction" (HCI) to a "human-machine cooperation" (HMC) [3], where the machine should now provide the "right" information and functionality, at the "right" time, and in the "right" way.

One of the solutions to achieve HMC relies on machine learning algorithms, with a performance that depends greatly on the quality of the algorithm (and proper tuning), but also on the data's (high) quality. There are several ways to improve algorithms' results, being the more usual way to train them repeatedly with all available data, with different settings, until the best possible result is achieved (fine-tuning the algorithm). Training might be extremely time-consuming, as well as it implies spending a lot of energy during the training phase, also increasing the algorithm's "carbon footprint".

Of course, there are ways in the literature to mitigate this problem, one of those is Active Learning [4]. The idea behind Active Learning is that the algorithms can achieve greater accuracy with fewer labelled training instances if they are allowed to pick the training data from which they learn, achieved by letting the learners ask queries in the form of unlabelled instances to be labelled by an oracle (e.g., human annotator). This filtered use of data might even have a greater effect on performance and costs since many times labelled data is scarce and extremely expensive to obtain (unlabelled data may be abundant but labels are difficult, time-consuming, or expensive to acquire).

A different solution is applying ensemble techniques, using for instance the results from algorithms previously thought and available in the community, e.g., open-source code. This use of hybridization and ensemble techniques allows empowering computation, functionality, robustness, and accuracy aspects of modelling [5], as well as allows us to reduce the "carbon footprint" of the algorithm once we use already trained model(s).

The ensemble solution is the approach adopted in this dissertation, as it will be explained in the next chapters. The present ensemble aggregates results from other models, by (possibly) training with the results returned from them. For instance, as will be the case in this dissertation, the ensemble/aggregator method uses floating-point numbers returned by running established algorithms over an image, sound (speech), or text, instead of using the original image/sound/text. In the present case, each floating-point number corresponds to a class of the emotion detection algorithm, but other alternatives are also admissible.

In short, this dissertation explores a framework that uses the aggregation of open-source methods (primary classifiers) to develop an emotion classifier supported by several sources, such as facial expression, speech, and text. Each of these “groups of methods” will be developed individually and integrated in the end, in a single final classification model, implementing an emotion multi-source aggregator (EMsA). Finally, with the emotions classified, they were divided into four groups: positive, negative, unexpected, and neutral, corresponding to the (predicted) sentiment of the user.

1.2 Objectives

As already mentioned, the main goal of the dissertation is to **(a)** ***develop a framework for emotion classification based on facial, speech, and text emotion prediction supported by an ensemble of open-source code retrieved from off-the-self available methods***, returning a single prediction consistent with the emotions presented by the system’s user. The secondary goal is to **(b)** ***propose a user sentiment classifier*** from the emotions presented.

Figure 1.1 shows the emotion classification framework scheme, that is going to be addressed and explained in detail in the different chapters of the dissertation.

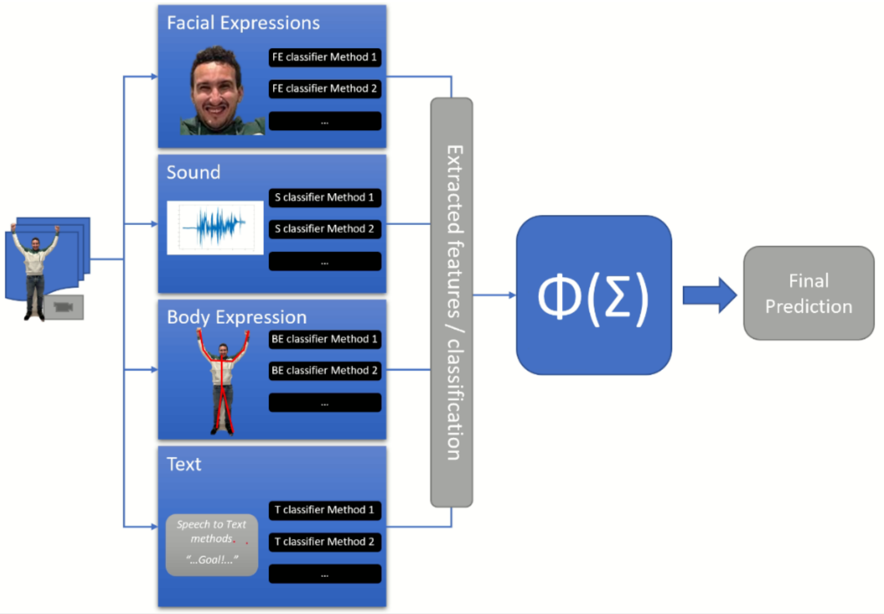


Figure 1.1 - Emotion classification framework scheme.

To achieve the main goals several subgoals can the mentioned:

i. Develop a framework for facial emotion classification (Fig. 1.1, “Facial expression” module)
ii. Develop a framework for speech emotion classification (Fig. 1.1, “Sound” module)
iii. Develop a framework for text emotion classification (Fig. 1.1, “Text” module)
iv. Combine an undetermined number of classifiers to predict emotions from video sources (Fig. 1.1, “Final Prediction” module).
v. Propose a simple sentiment classifier (Fig. 1.1, “Final Prediction” module).
vi. Allow dynamically add/remove/update classifiers to the framework.
vii. Publish the attained results in, at least, a journal or conference in the research field.

It is important to stress that, despite Fig. 1.1 shows a module for “Body Expression” this will not be addressed in this dissertation. The same applies to other modules, not presented in the figure, but important for the emotion and sentiment analysis, such as the analysis of the environment where the person stands.

1.3 Contributions

The main contributions of the dissertations are:

i. A framework for facial emotion classification that presents a classification accuracy above the open-source methods used.
ii. A framework for speech emotion classification that presents a classification accuracy above the open-source methods used.
iii. The integration of outputs from different methods and sources (image, sound, and text) in a single prediction that is consistent with the emotions presented by the system’s user.
iv. Two papers in indexed International Conferences.

1.4 Contents

The chapters of this dissertation are written in the format of a conference paper, except for the Introduction and Conclusions. This means that the state of the art for each subject is addressed in the corresponding chapter, as well as can appear repetitions in the text between chapters and sections, including the Introduction and Conclusions. Our decision was not to remove those repetitions in a way to allow the text to be more readable and self-contained. Two chapters are the done publications with minor adjustments.

In short, this present chapter introduces the reader to the context and goals of the dissertation. Chapter 2 addresses the facial emotion classification from video clips or live streaming. The framework receives information in the form of images (or frames) that will be passed to different types of facial primary emotion classifiers (available as open-source code), returning the same type or different types of results (corresponding to the emotions classes), which are then combined (by the aggregator) to return a (single) result. This chapter was published in: "Novais, R., Cardoso, P.J.S., Rodrigues, J.M.F. (2022). Facial Emotions Classification Supported in an Ensemble Strategy. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies. HCII 2022. Lecture Notes in Computer Science, vol 13308. Springer, Cham. https://doi.org/10.1007/978-3-031-05028-2_32"

Chapter 3, like the previous one, consists of the development of a framework supported by the aggregation of methods (primary classifiers) to make the speech emotion classification from audio files or videos. The framework will receive the results of emotion classification methods and aggregate them returning one single result. This chapter is accepted and will be published in: "Novais, R., Cardoso, P.J.S., Rodrigues, J.M.F. (2022). Emotion Classification from Speech by an Ensemble Strategy. Procs of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, Lisbon, Portugal, 31 Aug. 31 - 2 Sept."

Chapter 4, like in Chapters 2 and 3, focuses on the same principles, but now on text emotion classification. In Chapter 5 is presented the multi-source aggregator, i.e., the integration of all the previous sources and primary classifiers, as well as a proposal for a sentiment classification base on the emotions' classification. As future work, these last two chapters (4 and 5) are going to be integrated into a single paper to submit to the 25th International Conference on Human-Computer Interaction to be held in Copenhagen, Denmark, 23-28 July 2023.

Finally, Chapter 6 draws some conclusions and defines some potential future work.

2 FACIAL EMOTIONS CLASSIFICATION

Abstract. Humans are prepared to comprehend each other's emotions through subtle body movements or facial expressions, and from those, they change the way they deliver messages when communicating between them. Machines, user interfaces, or robots need to empower this ability, in a way to change the interaction from the traditional "human-computer interaction" to a "human-machine cooperation", where the machine provides the "right" information and functionality, at the "right" time, and in the "right" way. This chapter presents a framework for facial expression prediction supported by an ensemble of facial expression methods, being its contribution the integration of outputs from different methods in a single prediction consistent with the expression presented by the system's user. Results show a classification accuracy above 73 % in both FER2013 and RAF-DB datasets.

Keywords: Facial Emotions, Ensembles, Computer Vision, Machine Learning.

2.1 Introduction

Emotion and sentiment analysis methods are the automated processes of analysing information to determine the emotion (e.g., happiness, sadness, fear, surprise, disgust, anger, and neutral) or sentiment (e.g., positive, negative, and neutral) expressed by the user. The sentiment influences the emotion, and the emotions influence the sentiment. Humans are prepared to comprehend each other's emotions through subtle body movements, facial expressions, the way they speak, or simply by the tone of voice. They use this capacity when communicating between them, changing the way they pass the message based on those responses/emotions/sentiments.

We are living in the so-called Information Society, Society 4.0. However, we are starting to notice that the cross-sectional sharing of knowledge is not enough. So, in Japan appeared a new designation, Society 5.0, which should be one that "through the high degree of merging between cyberspace and physical space, will be able to balance economic advancement with the resolution of social problems by providing goods and services that granularly address manifold latent needs regardless of locale, age, sex, or language." [2]. Simplifying, Society 5.0 is a super-smart, people-centric society.

To achieve this degree of development, one of the keys is to empower machines, user interfaces, or robots with the same communication capabilities that humans have between them. This changes the interaction between machines and humans from the traditional "human-computer interaction" (HCI) to a "human-machine cooperation" (HMC) [3], where the machine provides the "right" information and functionality, at the "right" time, and in the "right" way.

One of the solutions to achieve HMC relies on machine learning algorithms with a performance that depends greatly on the quality of the algorithm (and proper tuning), but also the data's (high) quality. There are several ways to improve the algorithm's results, being the more usual way to train them repeatedly with all available data, with different settings, until the best possible result is achieved (fine-tuning the algorithm). Training might be extremely time-consuming, as well as it implies spending a lot of energy during the training phase, also increasing the algorithm's "carbon footprint".

Of course, there are ways in the literature to mitigate this problem, one of those is Active Learning [4]. The idea behind Active Learning is that the algorithms can achieve greater accuracy with fewer labelled training instances, if they are allowed to pick the training data from which they learn, achieved by letting the learners ask queries in the form of unlabelled instances to be labelled by an oracle (e.g., human annotator). This filtered use of data might even have a greater effect on performance and costs since many times labelled data is scarce and extremely expensive to obtain (unlabelled data may be abundant but labels are difficult, time-consuming, or expensive to acquire).

A different solution is applying ensemble techniques, using, for instance, the results from algorithms previously thought and available in the community (e.g., open-source code). This use of hybridization and ensemble techniques allows empowering computation, functionality, robustness, and accuracy aspects of modelling [5], as well as allows us to reduce the "carbon footprint" of the algorithm, once we use already trained algorithm(s), and now we are working with those results to develop the ensemble model. In short, the ensemble aggregates models by (possibly) training with the results from the adopted methods. For instance, as will be the case in this chapter, the aggregator method uses floating-point numbers returned by running established methods (primary classifiers) over an image, each number corresponding to a class of the emotion detection method/algorithm, instead of using an original colour image.

So, this chapter explores the latter solution, a framework supported in the use of aggregation of methods/algorithms to make the facial emotion classification from video clips or live streaming. The complete framework receives information in the form of images (or frames) that will be passed to different types of facial emotion classifiers (primary classifiers), available as open-source code, returning the same type or different types of results (corresponding to the emotions classes), which are then combined (aggregated) to return a (single) final result. The main contribution of the paper is the ensemble tool, which shows generically better results than using the methods individually.

In this Section, it was introduced the goals of the chapter. The next sections present some related work (Sec. 2.2) and the proposed ensemble facial expression classification framework (Sec. 2.3), followed by the developed tests and results in Sec. 2.4. Section 2.5 draws some conclusions and defines some potential future work.

2.2 Contextualization and Related Work

Expression recognition to interpersonal relation prediction needs input from different sources, e.g., sound, body, and facial expressions, as well as age or cultural environment. Zhang *et al.* [6] devise an effective multitask network that is capable of learning from rich auxiliary attributes such as gender, age, and head pose, beyond just facial expression data. Noroozi *et al*. [7] presented a survey on emotional body gesture classification. While works based on facial expressions or speech abound, recognizing the effect of body gestures remains a less explored topic. The authors in [7] present a new comprehensive survey hoping to boost research in the field. They first introduce emotional body gestures as a component of what is commonly known as “body language” and comment on general aspects, such as gender differences and cultural dependence. Then, they define a complete framework for automatic emotional body gesture recognition.

Other solutions were also presented, such as, the fusing of body posture with facial expressions for the classification of affect in child-robot interaction [8]. The opposite also exists, i.e., the dissociation between facial and body expressions (in emotion classification), as in a study done with impaired emotion recognition through body expressions and intact performance with facial expressions [9]. Further recent examples exist in the literature, such as mood estimation based on facial expressions and postures [10] or, e.g., in the following works [11]–[15].

In the present case, we are focusing on a single aspect which is facial expression. Ekman and Friesen demonstrated that facial expressions of emotion are universal, i.e., the human way of expressing an emotion is supposed to be an evolutionary, biological fact, not depending on the specific culture [16]. Nevertheless, different methods for facial expression classification return different results when presented with the same input (face). The idea of facial expression recognition (FER) using an ensemble of classifiers is not new. For example, Zavaschi *et al.* [17] presented in 2011 a pool of base classifiers created using two feature sets: Gabor filters and Local Binary Patterns (LBP). Then a multi-objective genetic algorithm has used to search for the best ensemble using as objective functions the accuracy and the size of the ensemble. Later (in 2019), Renda *et al.* [18] compared several ensemble deep learning strategies applied to facial expression recognition (for static images only). Ali *et al.* [19] presented an ensemble approach for multicultural facial expressions analysis. Intending to get high expression recognition accuracy, the study presents several computational algorithms to handle those variations. They use facial images from participants in the multicultural dataset that originate from four ethnic regions, including Japan, Taiwan, “Caucasians”, and Morocco.

Wang *et al.* [20] presented OAENet (oriented attention ensemble for accurate facial expression recognition). The authors used an oriented attention pseudo-Siamese network that takes advantage of global and local facial information. Their network consists of two branches, a maintenance branch that consisted of several convolutional blocks to take advantage of high-level semantic features, and an attention branch that possesses a UNet-like architecture to obtain local highlight information. The two branches are fused to output the classification results. As such, a direction-dependent attention mechanism is established to remedy the limitation of insufficient utilization of local information. With the help of the attention mechanism, their network not only grabs a global picture but can also concentrate on important local areas. In [21] the authors present a facial emotion recognition system that addresses automatic face detection and facial expression classification separately, the latter is performed by a set of only four deep convolutional neural networks concerning an ensembling approach, while a label smoothing technique is applied to deal with the miss-labelled training data.

The LHC is a Local (multi) Head Channel (self-attention) method [22], which is based on two main ideas. First, the authors hypothesize that in computer vision the best way to leverage the self-attention paradigm is a channel-wise application, instead of the more explored spatial attention, and that convolution will not be replaced by attention modules, like recurrent networks were in NLP (natural language processing); second, a local approach has the potential to better overcome the limitations of convolution than global attention. With LHC, the authors managed to achieve a new state-of-the-art over the FER2013 dataset [23], with significantly lower complexity and impact on the “host” architecture in terms of computational cost.

Py-Feat [24] is an open-source Python toolbox that provides support for detecting, pre-processing, analyzing, and visualizing facial expression data. Py-Feat allows experts to disseminate and benchmark computer vision models and also for end-users to quickly process, analyze, and visualize face expression data. For two recent (2021) surveys on various deep learning algorithms for efficient facial expression classification and human face recognition techniques, please refer to the works of Banerjee *et al.* [25] and Revina & Emmanuel [26].

All the above-mentioned methods need a huge amount of data (images) from which they learn from. Differently, we intend that our method learns from the result of previously established methods, simplifying the learning phase and reducing the time required to teach the classification model, as well as the computing power that is needed for that (decreasing this way the local “carbon footprint” of the framework). The next section explores the proposed framework in more detail.

2.3 Emotions Prediction Supported in Ensembles

As mentioned before, the framework to develop an emotion classifier should be supported by several sources, such as facial expression, body expression, speech, text, environment etc. Figure 1.1 illustrates that principle: the combination of an “undetermined” number of primary classifiers, that are dynamically added/removed/updated to/from the framework, to return a final prediction. The main idea behind the frameworkdepicted in Fig. 1.1 is that the primary methods (primary classifiers) are off-the-shelf methods, i.e., methods that have their code publicly available and can be easily added into the ensemble model, by providing final and raw classification results, that will be processed by the aggregator for the final classification prediction.

It should be stressed that the framework does not intend to improve any of the primary models, but only to work with the results they return. In this context, the emotions classifications models used in this chapter had their code extracted from some repository and no changes of any kind were done in the code. This means that the individual results presented by the emotion classifier, when applied to the datasets, are then collected and results are summarized, despite many times those are not coincident with the ones in the original publication.

2.4 Facial Emotion Classifier

To facilitate the description, this chapter only addresses the use of static images (i.e., it does not consider sequences of images, sounds, text, and body expressions) and considers that primary methods return 7 floating-point values corresponding to the different emotions, as seen next. However, the framework (Fig. 2.1) is easily adaptable to different outputs from the primary classifiers, e.g., the number and type of returned features.

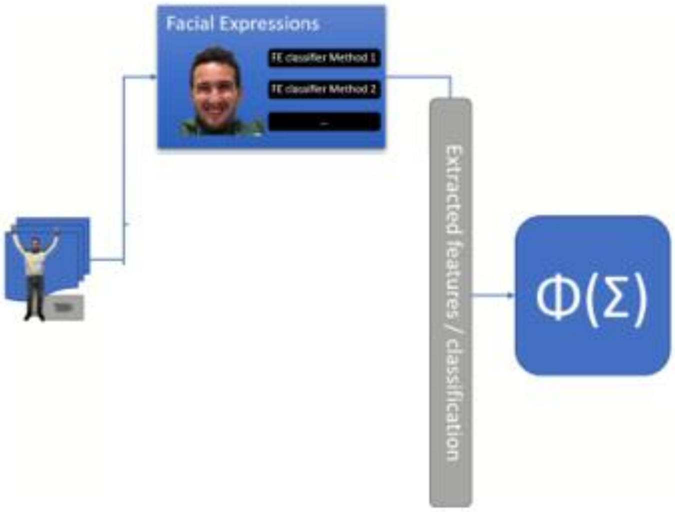


Figure 2.1 - Facial emotions classification framework.

Within the expressed restriction, the pipeline of the framework consists in presenting the same image to (i) *n* primary emotions classifiers (in the present case *n* = 3). Then, from the input image, each primary classifier returns a value between 0.0 (less likely to be) and 1.0 (more likely to be) for each of the seven classes/emotions considered (namely, happiness, sadness, fear, surprise, disgust, anger, and neutral). (ii) The returned values are then injected into the aggregation model, to produce a single final classification. This means that for each presented image there will be *n* × 7 inputs to the aggregation model and the emotion as output.

For the initial part, (i), of the framework pipeline, the following primary emotion classifiers were used: (1) LHC [22], with its code available at [27]; (2) Py-Feat [24], with its code available at [28]; and (3) FERjs, which is a free implementation done by Justin Shenk and has its code available at [29]. The reasons to choose these three methods to build the baseline were: (a) they present state-of-the-art results, (b) are recent methods, from 2021, (c) have publicly available code (implementation), and (d) represent different architectures. Again, it is important to stress that there is a huge number of different methods that could be used, as mentioned in [25], [26].

For the second part of the framework pipeline, (ii), the models/classifiers used were: (a) Voting; (b) Random Forest [30]; (c) AdaBoost [31]; and (d) a Multi-layer Perceptron/Neural Network (MLP/NN) [32]. Models (b-d) were tested without ranking the values returned by the initial classifiers, as we will see later. In more detail, the (a) Voting method used the classes predicted by the primary classifiers to predict if there is a "majority" of opinion between the guesses, that is, if at least two of the predictors guess the same emotion (given *n* equal to 3). If all return different emotions, then the Voting method returns no prediction. The Voting method can be considered a naïve method but serves as a baseline for building more advanced aggregation strategies.

The (b) Random Forest [30] is *per se* an aggregator of predictors. It starts by the drawing of *k* bootstrap samples from the original data and then, for each of the bootstrap samples, grows an unpruned classification tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample *m* of the predictors and choose the best split from among those. The (c) AdaBoost method [31], as the name suggests, uses boosting which involves combining the predictions from many weak learners, being a weak learner a (very) simple model, although it has some skill on the dataset. The AdaBoost algorithm uses short (one-level) decision trees as weak learners that are added sequentially to the ensemble. Each subsequent model attempts to correct the predictions made by the model before it in the sequence. This is achieved by weighing the training dataset to put more focus on training examples in which prior models made prediction errors. Finally, the (d) Multi-layer Perceptron [32] (MLP) is a feedforward artificial neural network model that maps sets of input data onto a set of outputs. An MLP consists of multiple layers and each layer is fully connected to the following one. There can be one or more non-linear hidden layers between the input and the output layer.

The next section details on the tests and achieved results.

2.5 Tests and Results

For the tests, it was used, as already mentioned, two different datasets, namely: (i) FER2013 [23], [33], where the data consists of 48 × 48-pixel grayscale images of faces. The faces have been automatically registered so that they are centered and occupy about the same amount of space in each image. Each face is annotated with a facial expression, one of the seven previously mentioned categories. The training set consists of 28,709 examples and the public testing set consists of 3,589 examples.

The second dataset used is (ii) RAF-DB [34], which is a facial expression database with 29,672 facial images, downloaded from the Internet. Images in this database have great variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions (e.g., glasses, facial hair, or self-occlusion), post-processing operations (e.g., various filters and special effects) etc. The images were classified into two different subsets: single-label subset, including the 7 classes of basic emotions (same as FER2013), and two-tab subset, including 12 classes of compound emotions. It also includes 5 accurate landmark locations, 37 automatic landmark locations, bounding boxes, race, age range, and gender attributes annotations per image. As usual, to be able to objectively measure the performance of the followers' entries, the database has been split into a train set and a test set, where the size of the training set is five times larger than the one of the testing set, and expressions in both sets have a near-identical distribution.

It is important to stress that in the case of RAF-DB, before applying the emotion classifier, a face detector was applied. As the goal of the chapter is not to select the best detector to apply in this situation, it was applied one of the most well-known face detectors - Haar-Cascate face detector [35].

Regarding both datasets, Table 2.1 shows the accuracy for each primary emotion classifier and for the Voting model, serving as a reference for the latter tested and more advanced aggregation models.

Table 2.1 - Accuracy for the individual emotion classifiers and Voting model.

Dataset	LHC	Py-Feat	FERjs	Voting
FER2013	70.59%	**77.98%**	65.67%	73.37%
RAF-DB (2)	60.68%	62.04%	62.19%	**62.60%**

Besides the Voting model, the other aggregators' methods were tuned using a grid search stratified 5-fold cross-validation, i.e., for each set of classifier parameters, the training data was divided into 5 folds, with the same classes ratio as the training set, and then the algorithms were trained and tested 5 times, where each time a new set (a fold) is used as testing set while remaining sets are used for training. The scoring (the accuracy in this case) for each set of parameters is computed as the mean score over the 5 train-test runs. Next, the full training dataset and best scored set of parameters are used to obtain the final model for each of the methods (i.e., Random Forest, AdaBoost, and MLP/NN).

Scikit-learn machine learning library for the Python programming language [36] (version 1.0.1) was used to develop the models. Besides the default values, the parameters used to tune the methods are summarized in Table A.1 e A.2 (Appendix). The remaining chapters's results were attained in a personal computer running Kubuntu 21.10 over an Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz with 16 GiB of RAM.

Considering the methods and datasets above introduced, different combinations of those were used, obtaining different aggregation models. So, Table 2.2 shows the results of the different models applied to the different datasets where the results per row were obtained in the following manners. The FER2013 row shows the results considering that the outputs of the primary methods (LHC, Py-Feat, and FERjs), namely the methods' estimated confidence of being each of the emotions, are injected into the aggregators' methods (Random Forest, AdaBoost, and MLP/NN). In this case, 21 features are injected, since 3 primary methods returning 7 expressions were used, either directly with no transformation or ranked within each method (resulting in the "Without ranking" and "With ranking" table's columns). Furthermore, in the aggregators training phase, the injected values were the results of applying the primary methods to the FER2013 training dataset, and the results shown in the table are the values obtained by applying the final aggregator model (the model with parameters obtained from the grid-search cross-validation phase and trained over the full FER2013 training dataset) to the FER 2013 testing dataset. Row RAF-DB (1) shows the results of applying the above models (the model trained for table's RAF2013 row) directly to the full RAF-DB dataset, which in this case can be considered as a testing dataset since the model never saw that data.

Finally, row RAF-DB (2) was built similarly to row RAF2013, with the difference being that the training and testing datasets were RAF-DB training and testing datasets, respectively. The parameters obtained from the grid-search cross-validation are summarized in Table A.2 (Appendix).

Table 2.2 - Results of the different models applied to the different datasets.

	Without ranking			With ranking		
Dataset	Random Forest	AdaBoost	MLP/ NN	Random Forest	AdaBoost	MLP/ NN
FER2013	71.08%	**71.41%**	70.95%	70.68%	70.79%	70.84%
RAF-DB (1)	**60.24%**	60.17%	60.01%	59.47%	59.53%	59.75%
RAF-DB (2)	**76.17%**	64.94%	74.14%	38.98%	38.98%	67.07%

Some conclusions can be drawn from Table 2.1 and Table 2.2. The first conclusion is that, although in some cases the values are close, the ranking of the values is not justified as it always returned worst accuracy than the corresponding method without ranking, i.e., it seems to be a better solution to inject the values from primary methods directly into the aggregation methods. Considering the results over the FER2013 test dataset, the best aggregation method was the Voting model with an accuracy of 73.37 %, which is worse than the accuracy of the Py-Feat model (77.98 %). As a curiosity, which is not presented in the tables, is the fact that the accuracy of the models over the FER2013 training dataset was 99.18 %, 81.84 %, and 74.96 %, respectively. This seems to indicate overfitting of the first method since it drops from 99.18 % accuracy over the training dataset to 70.59 % accuracy in the testing dataset. In the reverse, Py-Feat suffered a very small drop from 81.84 % to 77.98 % of accuracy. This is relevant since having a 99.18 % accuracy, the aggregation methods might have had some somehow misleading predictions from method LHC – this was an expectable risk and provides us with further studies to mitigate this threat. Applying the aggregated model trained for FER2013 to RAF-DB is interesting of the fact that it produces results very similar to the primary methods without the need to train them with that dataset. In more detail, LHC, Py-Feat and FERjs trained with RAF-DB training dataset produced an accuracy of 60.68 %, 62.04 %, and 62.19 %, respectively, which is very similar to the aggregation methods trained with FER2013 accuracies (60.24 %, 60.17 %, and 60.01 %, respectively), but without the need to train a new model.

If the aggregation models were trained using the prediction from LHC, Py-Feat and FERjs for the RAF-DB training set, then their prediction improve the base methods in all (the without ranking) cases, i.e., the best accuracy was 62.19 % for method FERjs, and the aggregation methods attained an accuracy of 76.17 %, 64.94 %, and 74.14 % (for Random Forest, Ada-Boost, and MLP/NN, respectively).

Between the aggregation methods, AdaBoost was the one performing worst. Random forest and MLP/NN had similar results, being the Random Forest slightly better in the tested cases.

2.6 Conclusions

This chapter presented a simplified version of a facial expression/emotions predictor framework supported in ensembles. The pipeline of the frameworks consists in presenting an image, to several (primary, pre-trained) emotion classifiers. Then, each classifier returns, for each image and, for each of the seven considered classes/emotions (happiness, sadness, fear, surprise, disgust, anger, and neutral) its confidence values. Those results are then fed to an aggregator model returning a single predicted class.

The best results for the aggregators' methods in the case of FER2013 dataset were achieved with the Voting model (supported on the majority of the model's predictions), being above two of the primary emotion classifiers but below one of them. This result is achieved probably because the emotion classifiers were taught with FER2013 but have different accuracy behaviors (one of the classifiers is probably overfitted since the accuracy dropped from almost 100% on the training set to nearly 70% on the test set). In the case of RAF-DB, the best result was achieved with the model Random Forest aggregator, and the result is above all the results achieved individually by the primary emotion classifiers.

In future work, we intend to explore different datasets, like the ones mentioned in [24], [37], and datasets that have motion (video or streaming). We will also try to improve the final results by increasing the number of emotion classifiers and studying the influence of their characteristics, like the fact that they are over or under-fitted.

3 EMOTION CLASSIFICATION FROM SPEECH

Abstract. Humans are prepared to comprehend each other's emotions through subtle body movements and speech expressions, and from those, they change the way they deliver and understand messages when communicating between them. Socially assistive robots need to empower their ability in recognizing emotions in a way to change the interaction with humans, especially with elders. This chapter presents a framework for speech expression prediction supported by an ensemble of distinct out-of-the-box methods, being the main contribution the integration of the outputs of those methods in a single prediction, consistent with the expression presented by the system's user. Results show a classification accuracy of 75.56 % over the RAVDESS dataset and 86.43 % in a group of datasets constituted by RAVDESS, SAVEE, and TESS.

Keywords: Speech Emotion Recognition, Ensembles, Machine Learning, Emotions

3.1 Introduction

Humans are prepared to comprehend each other's emotions through subtle body movements, facial expressions, the way they speak, or simply by the tone of voice. On the other hand, computationally speaking, several methods are being developed to automate the processes of analyzing information from media sources to determine the emotions (e.g., happiness, sadness, fear, surprise, disgust, anger, and neutral) [38] or sentiments expressed by users [39].

This research is fundamental as, the worldwide elderly population is set to be more than double by 2050 [40] and robots are expected to assume new roles in health and social care, to meet that higher demand [41], [42]. Emotion and sentiment analysis are therefore fundamental in the development of socially assistive robot (SAR) technologies for elderly care. Nevertheless, a robot can only really interact with a person (elder), if it achieves some degree of emotional recognition in conversation (ERC) [43], i.e., if it understands the person's emotions and sentiments in a way to, on the fly, adjust its behavior/interaction in function of it. There are very recent studies which analyze emotional intelligence in SAR for elders [44] or which aspects may influence human-robot interaction in assistive scenarios [45].

To implement SAR technologies, state-of-the-art results in emotions and sentiments classification and needed and, for that, it is necessary to rely on machine learning algorithms. Several ways are known to improve algorithms' results, being the more usual way to train them repeatedly, with available data, with different settings, until the best possible result is achieved (fine-tuning the algorithm). Training might be extremely time-consuming, as well as it implies spending a lot of energy during the training phase, also increasing the model's "carbon footprint". A solution to mitigate this is applying ensemble techniques, i.e., using the results from various algorithms previously thought and available in the community [46]. This use of hybridization and ensemble techniques allows empowering computation, functionality, robustness, and accuracy aspects of modelling [5], as well as it allows to reduce the referred "carbon footprint" of the models.

This chapter explores a framework supported by the aggregation of models (primary classifiers) to make speech emotion recognition (SER) from video clips or audio file samples. Similar, but complementary, to our previous study that explores an ensemble in facial expression classification [46], the present method receives audio file samples that will be passed to different types of speech emotion recognition methods (primary classifiers, available as open-source code), returning the same or different types of results (corresponding to the emotions classes), which are then aggregated to return a final result. In this chapter, the authors also present a primary classifier for speech emotion classification. The main contribution of the chapter is the ensemble tool, which shows generically better results than using the primary classifiers individually.

The goals of the chapter were introduced in this section. The next sections present some related work (Sec. 3.2) and the proposed ensemble speech expression classification method (Sec. 3.3), followed by the developed tests and results (Sec. 3.4). Section 3.5 draws some conclusions and defines some potential future work.

3.2 Contextualization and Related Work

As mentioned, the framework to develop an emotion classifier should be supported by several sources, such as facial expression, body expression, speech, text, environment etc. (see Figure 1.1). Novais *et al.* [46] illustrated that principle: the combination of an "undetermined" number of primary classifiers, that are dynamically added/removed/updated to/from the framework, to return a final prediction. The main idea behind the framework is that the primary classifiers are off-the-shelf methods, that have their code publicly available, and can be easily added to the aggregation model. As in most studies and emotion recognition methods/algorithms, this study focuses on Paul Ekman and his colleagues' basic emotions study [38] which, despite being done for facial expressions, were generalized to other emotions analyses, such as speech. They are anger, disgust, fear, happiness, sadness, and surprise. Given the used datasets, the neutral emotion was added to them. In the next subsections, the datasets used in this study and a brief state-of-the-art about speech emotion classification methods will be presented.

3.2.1Speech Emotions Databases

There are several databases for speech emotion testing. In this chapter, we refer to four of them, mostly due to their diversity and length. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [47] comprises 7,356 files. The database contains 24 professional actors (12 female and 12 male), vocalizing two lexically matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprised, and disgusted emotions and songs contain calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal and strong), with an additional neutral expression. All conditions are available in three modality formats: audio-only (16bit, 48kHz .wav), audio-video (720p H.264, AAC 48kHz, .mp4), and video-only (no sound).

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [48] comprises 7,442 original clips. These clips were recorded from 48 male and 43 female actors, between the ages of 20 and 74, and coming from a variety of races and ethnicities (namely, African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (anger, disgust, fear, happy, neutral, and sad) and four different emotion levels (low, medium, high, and unspecified). Participants rated the emotion and emotion levels based on the combined audiovisual presentation, the video alone, and the audio alone. Due to the large number of ratings needed, this effort was crowd-sourced and a total of 2,443 participants rated each 90 unique clips, 30 audios, 30 visuals, and 30 audio-visuals (95 % of the clips have more than 7 ratings).

The Surrey Audio-Visual Expressed Emotion (SAVEE) [49] was recorded from four native English male speakers, of the University of Surrey, aged between 27 and 31.

Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness, surprise, and neutral. The text material consisted of 15 TIMIT sentences per emotion: 3 common, 2 emotion-specific, and 10 generic sentences, that were different for each emotion and phonetically balanced. The 3 common and 2 × 6 = 12 emotion-specific sentences were recorded as neutral to give 30 neutral sentences. This resulted in a total of 120 utterances per speaker, i.e., 480 samples.

The Toronto Emotional Speech Set (TESS) [50] has a set of 200 target words spoken by two actresses (aged 26 and 64 years) in the carrier phrase "Say the word _____". Recordings were made portraying each of seven emotions (anger, disgust, fear, happiness, pleasant, surprise, sadness, and neutral), resulting in a total of 2,800 stimuli. The two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

3.2.2State of the Art from Speech Emotions Classification

Popova *et al.* [51] presented an approach in which the classification of a sound fragment is reduced to the problem of image recognition. The authors use the waveform and spectrogram as visual representations of the sound. They test the results with RAVDESS, achieving an accuracy of 71 % when combining a Mel-spectrogram and a convolutional neural network (VGG-16). Chen *et al.* [52] presented a 3D attention-based convolutional recurrent neural network to learn discriminative features for SER, where the Mel-spectrogram with deltas and delta-deltas are used as input. Experiments on IEMOCAP and Emo-DB corpus demonstrate state-of-the-art performance in terms of unweighted average recall.

Palanisamy *et al.* [53] showed that ImageNet pre-trained standard deep convolutional neural network (CNN) models can be used as strong baseline networks for audio classification, showing 92.89 % validation accuracy on the ESC-50 dataset and 87.42 % validation accuracy on the UrbanSound8K dataset. de Pinto *et al.* [54] presented a CNN-based classification model of emotions elicited by speeches (using the RAVDESS dataset). The model has been trained to classify Ekman's [38] emotions plus the neutral and calm ones. They achieved a weighted average F1 score of 0.91, with the best performances in the "Angry" class, with an F1 score of 0.95, and the worst observed in the “Sad” class, with an F1 score of 0.87. El Seknedy and Fawzi [55] presented a study for speech emotion recognition tested on RAVDESS, Emo-DB, and CaFE datasets. They use 4 machine learning classifiers (Multi-Layer Perceptron, Support Vector Machine, Random Forest, and Logistic Regression), and a newly developed feature set was introduced consisting of main speech features as prosodic 7 features, spectral features, and energy. Furthermore, feature importance techniques were used to study the feature importance per each classifier across each corpus. The model achieved an accuracy of 70.56 % on RADVESS, 85.97 % on Emo-DB, and 70.61 % on CaFE. Kumaran *et al.* [56] presented a deep continuous recurrent neural network (C-RNN) approach to classifying the effectiveness of learning emotion variations in the classification stage. They use a fusion of Mel–Gammatone filter in convolutional layers to first extract high-level spectral features. Then recurrent layers are adopted to learn the long-term temporal context from high-level features. The authors achieved an accuracy of 80 % in RADVESS.

More recently, in 2021, Abbaschian *et al.* [57] reviewed deep learning approaches for SER with available datasets, followed by conventional machine learning techniques for speech emotion recognition. The authors presented a multi-aspect comparison between practical neural network approaches in speech emotion recognition. Also in the same year, Lieskovská [58] reviewed SER approaches using deep learning and different attention mechanisms. Finally, Prasanth *et al.* [59] provided a general outline of strategies for machine learning, pre-processing, feature extraction techniques, and determining the accuracy of suitable classifiers. The authors described and addressed various SER tactics and ideas, giving a detailed survey of each one’s existing literature, where these approaches are used for the identification of speech-based emotions. The analysis and experimental studies include databases used, emotions collected, extraction of the features, and the improvements made in the precision to the identification of speech emotions and pertinent shortcomings.

All the above-mentioned methods are based on training a completely new method from the start. Differently, we intend that our method learns from the result of previously established models, simplifying the learning phase and reducing the time required to teach the classification model, as well as the computing power that is needed for that (this way decreasing the local “carbon footprint” of the framework). The next section explores the proposed framework in more detail.

3.3 Speech Emotions Classifier

Within the restriction that this chapter focuses on Ekman's [38] defined emotions (added neutral), the pipeline of the framework consists of (see Fig. 3.1): (1) presenting the same audio file to *n* primary emotions classifiers (in the present case *n* = 3), each one returning a value between 0.0 (less likely to be) and 1.0 (more likely to be) for each of the seven classes/emotions (i.e., happiness, sadness, fear, surprise, disgust, anger, and neutral); (2) The returned values are then injected into the ensemble/aggregation model, to produce a single final classification. This means that, for each presented audio file there will be *n* × 7 inputs to the aggregation model and an expression/emotion as output.

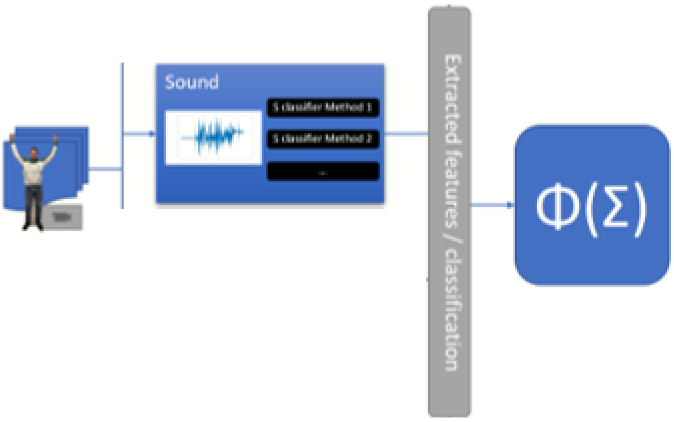


Figure 3.1 - Speech emotions classification framework.

For the initial part of the framework pipeline, (1), the following primary emotion classifiers were used: (i) MDP [54], with its code available in [60]; (ii) SB, which is a free implementation done by Shivam Burnwal and has its code available at [61]; and (iii) SEDC (Speech Emotion Detection Classifier), which is an implementation done by the authors, summarized below in this section (and Fig. 3.2).

The reasons to choose these initial methods as primary classifiers were: (a) they are recent methods, (b) have publicly available code (implementation), and (c) they present different architectures. Furthermore, MDP was reviewed by peers and resulted in a scientific publication [54], the second one (SB) was done in a Bachelor project and presented on Kaggle platform, and the third (SEDC), as stated, is a method proposed by the authors. Again, it is important to stress that there is a huge number of different methods that could be used, as mentioned in [57]–[59].

The ***Speech Emotion Detection Classifier*** (SEDC) follows the same principles as presented in [54], i.e., SEDC is based on a deep learning strategy with a CNN and dense layers. The key idea is to consider features obtained from the computation of: (i) the chromatogram from the waveform/power spectrogram, (ii) the zero-crossing rate of the audio time series, (iii) the Mel-frequency cepstral coefficients (MFCCs), and (iv) the root-mean-square (RMS) value for each frame. For further details, please refer respectively to methods *chroma_stft*, *zero_crossing_rate*, *mfcc*, and *rms* in [29]. So, features were generated by converting each audio file to a floating-point time series, as done by [54], followed by the application of the mentioned models, resulting in 162 extracted features. Consequently, the proposed network will work on vectors of 162 features for each audio file provided as input. The model has five 1D convolution blocks and two dense blocks (fully connected layer), see Fig 3.2, and the hyperparameters are presented in Tab. 3.1. The next section details the tests and the achieved results.

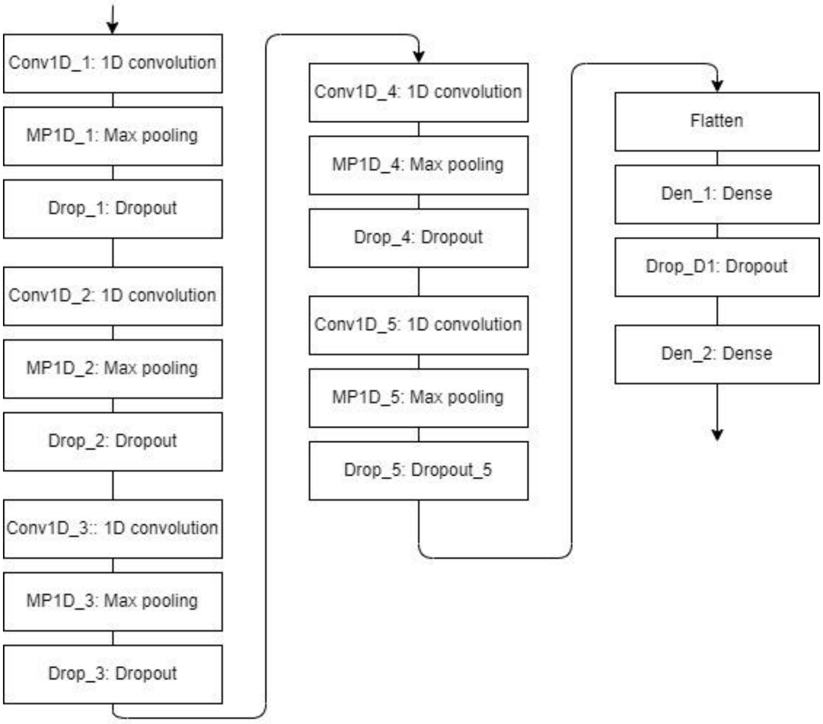


Figure 3.2 - Speech Emotion Classifier.

Table 3.1 - Hyperparameters for SEDC.

1D Convolution	Filters	Kernel size	Stride	Activ.	**Max Pooling**	Size	Stride	**Dropout**	rate	**Dense**	unit	Activ.
Conv1D_1	256	5	1	ReLU	MP1D_1	5	2	Drop_1	0.2	Den_1	32	ReLU
Conv1D_2	256	5	1	ReLU	MP1D_2	5	2	Drop_2	0.2	Den_2	7	SoftMax
Conv1D_3	128	5	1	ReLU	MP1D_3	5	2	Drop_3	0.2			
Conv1D_4	128	3	1	ReLU	MP1D_4	5	2	Drop_4	0.2			
Conv1D_5	64	3	1	ReLU	MP1D_5	3	2	Drop_5	0.2			
								Drop_D1	0.2			

Optimizer:	rmsprop (Root Mean Squared Propagation)	**Batch size:**	32
Loss:	Categorical cross-entropy	**Epochs:**	300

For the second part of the framework's pipeline, (2), the used models were: (a) Voting; (b) Random Forest (RF) [30]; (c) AdaBoost [31]; and (d) a Multi-layer Perceptron/Neural Network (MLP/NN) [32]. Models (b-d) were tested without ranking the values returned by the primary classifiers, as we will see later. In more detail, (a) the Voting method uses the classes estimated by the primary models to predict its own class if there is a majority of opinion between the guesses. E.g., in the present case, where 3 primary classifiers were considered, if at least two of the primary predictors guess the same emotion, then that emotion is return as the prediction. If the primary predictors return different emotions (i.e., there is no majority), then the Voting method returns no prediction. The Voting method is a naïve method but serves as a baseline for building more advanced aggregation strategies. The (b) Random Forest [30] is *per se* an aggregator of predictors. It starts by drawing *k* bootstrap samples from the original data and then, for each of the bootstrap samples, grows an unpruned classification tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample *m* of the features and choose the best split from among those.

The (c) AdaBoost method [31], as the name suggests, uses boosting which involves combining the predictions from many weak learners, being a weak learner a (very) simple model, although it has some skill on the dataset. The AdaBoost algorithm uses short (one-level) decision trees as weak learners that are added sequentially to the ensemble. Each subsequent model attempts to correct the predictions made by the model before it is in the sequence. This is achieved by weighing the training dataset to put more focus on training examples in which prior models made prediction errors. Finally, the (d) Multi-layer Perceptron [32] is a feedforward artificial neural network model that maps sets of input data onto a set of outputs. An MLP consists of multiple layers and each layer is fully connected to the following one. There can be one or more non-linear hidden layers between the input and the output layer.

The next section presents the tests and results using a different combination of the datasets presented in Sec. 3.2.1

3.4 Test and Results

Three different training/testing scenarios were implemented: (a) ***RAVDESS*** – the full dataset has 8 emotions but, for compatibility with the other ones in this study, the "calm" emotion was removed, both for training and testing. In this sense, only 2,075 file samples were considered since 377 ("calm" files) were dropped from the total of 1,440 speech and 1,012 song files in the dataset; (b) ***All** dataset* – conjugates RAVDESS (with the restriction presented in (a)), CREMA-D, SAVEE, and TESS, corresponding to 12,797 (= 2,075 + 7,442 + 480 + 2,800) audio samples; and (c) ***3DataSet*** – which considers RAVDESS (same as in (a)), SAVEE, and TESS corresponding to 5,355 audio samples. For all cases, 70 % of the audio files were used for training, 15 % for validation, and 15 % for testing. Furthermore, before training, data augmentation was also used, generating syntactic data for audio (namely: noise injection, shifting time, changing pitch, and speed), which increased the number of training samples in the expected proportion. All training and testing followed the hyperparameters summarized in Tab. A.3 (Appendix A) and were conducted on a personal computer running Windows 10 over an AMD Ryzen 7 4800H @ 2.90 GHz with 16 GiB of RAM.

The metrics used for the evaluation of the models were accuracy and F1 for the primary models (see, e.g.,[54]) and accuracy for the aggregators. Table 3.2 shows the attained results of the primary models, i.e., before applying the aggregators. For all considered datasets, SEDC was the best performing method with an accuracy of 69.45 % for ***RAVDESS***, 61.59 % for ***All** dataset*, and 83.81 % for ***3DataSet***.

Table 3.2 - Baseline results for the 3 primary methods.

	RAVDESS			***All** dataset* = RAVDESS + CREMA-D + SAVEE + TESS			***3DataSet*** = RAVDESS + SAVEE + TESS		
	MDP	SB	SEDC	MDP	SB	SEDC	MDP	SB	SEDC
Accur.	62.37%	64.63%	**69.45%**	53.25%	59.35%	**61.59%**	78.45%	80.82%	**83.81%**
F1	0.63	0.65	**0.69**	0.53	0.59	**0.62**	0.79	0.81	**0.84**

Except for the Voting method, which, as already explained, receives as input the three predicted classes, the aggregators' methods receive as input the values estimated by the primary methods, i.e., the values between 0.0 (less likely to be) and 1.0 (more likely to be) for each of the seven classes/emotions, resulting in 21 values. In the tunning phase, done using a grid search stratified 5-fold cross-validation, the aggregators use as input the values achieved by primary methods inference over training and validation sets. Then, founded the best set of hyperparameters (See also Tab. A.3 in Appendix A), the union of the training and validation sets are used to obtain the final model for each of the aggregation methods (i.e., Random Forest, AdaBoost, and MLP/NN). Finally, the tests were done over the values achieved by primary methods inference over the test set. Aggregators were implemented using the Scikit-learn machine learning library [36].

Table 3.3 shows the results of the different models applied to the different datasets. The best result for the ***RAVDESS*** dataset has been achieved using Random Forest without ranking (the values returned by the primary methods are used as they are) and MLP with ranking (the values returned by the primary methods are ranked before being injected into the aggregators) with an accuracy of 75.56 % (6.11 % above the best baseline result, 69.45 %). Using ***All** datasets*, the best result was achieved again using Random Forest (without ranking) with an accuracy 63.64 % (2.05 % above the best baseline result, 61.59 %), and for the ***3DataSet*** the best result was achieved using MLP without ranking with an accuracy of 86.43 % (2.62 % above the best baseline result, 83.81 %).

Table 3.3 - Results of the different models applied to the different datasets.

		Without ranking			With ranking		
Dataset	Voting	Random Forest	AdaBoost	MLP/ NN	Random Forest	AdaBoost	MLP/ NN
RAVDESS	70.41 %	**75.56 %**	62.37 %	74.91 %	74.27 %	66,55 %	**75.56 %**
All datasets	58.05 %	**62.63 %**	61.07 %	62.42 %	61.85 %	60.91 %	62.11 %
3DataSet	83.93 %	86.05 %	81.69 %	**86.42 %**	85.18 %	85.67 %	85.55 %

3.5 Conclusions

This chapter presents a speech expression/emotions predictor framework supported in an ensemble. The pipeline of the framework consists in presenting an audio file to selected methods of the *Librosa* library which extracts features that are then sent to the three audio emotions primary classifiers. Then, each primary classifier returns its confidence values for the seven classes/emotions (happiness, sadness, fear, surprise, disgust, anger, and neutral) considered. Those confidence values are then fed to an aggregator model returning a single predicted class.

The best results for the aggregator's methods in the case of ***RAVDESS*** dataset were achieved using Random Forest and MLP with ranking. In all tests, the accuracy was always above the accuracy of the primary method. Similar results are achieved when considering the ***All** datasets* and ***3DataSet***, although, in the latter case, AdaBoost with ranking performed better than the primary methods. In conclusion, the aggregation methods are a promising solution as they provide a solution to improve the performance achieved by primary methods.

In future work, we intend to explore different datasets. We will also try to improve SEDC model and the results by increasing the number of emotion classifiers and studying the influence of their characteristics.

4 EMOTION CLASSIFICATION FROM TEXT

Abstract. Human emotions and sentiment classification can be achieved for instance from text analysis. Currently, the classification of expressions in texts is fundamental in social media, blogs, articles, or product evaluation. In this chapter, it was used Ekman's six basic emotions for text classification (namely, fear, anger, joy, sadness, disgust, and surprise), and also included the neutral emotion. A set of aggregators for classifying text emotions will be presented, supported by a set of available open-source classification methods. The main models used in this chapter were trained and tested with the EMOTIONLINES dataset, being the test and aggregators results displayed at the end.

Keywords: Text Emotions, Ensembles, Machine Learning.

4.1 Introduction

Emotions are very important for the human being to know how to deal with decisions, interactions, and cognitive processes, being considered a psychophysiological process. Emotions can be motivated by the conscious and/or unconscious of objects, and situations associated with a multiplicity of factors, such as, mood, temperament, personality, disposition, and motivation [62]. There are several types of sources from where emotion, and sentiment classification can be retrieved. These include the human face and body, as well as speech or text.

Several methods use a combination of sources (e.g., text, speech, and image) for emotion classification but, unfortunately, not all sources are always available. For example, for online retailers, such as Amazon, user’s emotions classification can be obtained from text, whenever text review data is available. Other examples are the call centers, in which emotion classification can be retrieved from speech, or security cameras, having emotion classification taken from visual (image) information. Having all sources at the same time is not a fully common [63].

Emotion classification has some important points of focus, such as figuring out how a person feels about some event, person, or thing, based on some predefined emotion models according to psychological theories of emotion. These classifications can apply almost in all aspects of our daily life, such as [63]: monitoring the mental health of people;

modifying, or improving business strategies depending on the emotion of customers; detecting potential criminals or terrorists by analyzing the emotions of people after/before a terrorist attack or crime; improving the performances of chatbots and other automatic feedback systems; etc.

This chapter focuses on text analysis, being the text classification fundamental at the social media, in blogs, in news articles, or, in the already mentioned, product evaluations [64]. E.g., Twitter is one of the major social media platforms where sentiments about some topic or a concept are expressed [65], being an excellent way to obtain data used in various analyzes to better understand the human being emotions [64]. In short, the popularity of text sentiment analyses is increasing in researchers’ communities and the business world, the former because of the challenge and the latter because of the potential value to make profits [66].

It is also important to stress that such extracted/analyzed data is fundamental to improve human-computer interaction, allowing to “teach computers” to make better decisions in the helping of users. These improvements and usage include, e.g., the interaction between humans and robots, since getting the emotion right would make human-robot interaction more natural, once it can adapt its output in the function of the human’s emotions.

In this chapter, primary models were developed to obtain the classification of text emotion. The basis of those models were created by several authors who made codes available online. As was done in Chapters 2 [46] and 3 [67], the primary classifiers make their predictions of emotions, with the respective probability for each emotion. After getting the predictions from those primary classifiers, the values are then injected into an aggregator/ensemble classifier. Also as before, the used aggregators were Random Forest, Adaboost, and MLP/NN.

To train and test the framework, the EMOTIONLINES dataset was used [66], which consists of a collection of Friends TV scripts and private Facebook messenger dialogues. This dataset uses six of the Ekman’s basic emotions [66], plus the neutral emotion.

The chapter’s objectives were introduced in this section. The next section presents some related work (Sec. 4.2), Section 4.3 describes de database used in the study, and the proposed ensemble text emotion classification method is presented in Sec. 4.4, followed by the developed tests and results (Sec. 4.5). Section 4.6 draws some conclusions and defines some potential future work.

4.2 Contextualization and Related Work

Implementing Natural Language Processing (NLP) technologies requires state-of-the-art results in emotional classification, being necessary to rely on machine learning (ML) algorithms to achieve state-of-the-art models. To improve ML algorithms results, the more common way is to train the methods with the same data but with different settings, until getting the best possible results. Worth mentioning again, this chapter is not focused on studying/training new methods from the scratch, but on using methods already trained and tested, and trying to improve their result, by combining the output information from several of those (open-source) models.

We recall that, the final goal of the framework (Chap. 5) is to develop an emotion classifier that must be supported on several sources [46], such as facial expression, sound, body, and text. The framework will receive an indeterminate number of primary classifiers which are dynamically added/removed/updated to/from the framework thus returning a final prediction. The framework has its primary methods (primary classifiers), that were properly trained, preferentially with their source code is publicly available, and the results from those methods can be added to an aggregation model returning a final classification.

As mentioned in the Introduction section, this chapter focuses on text emotion classification, following the authors' works in previous papers, namely for facial [46] and speech [67] emotion classification. In a continuity context, this means that the text classification will also be based on Paul Ekman's six emotions, namely: angry, disgusted, afraid, sad, happy, and surprised. Having added an extra expression called neutral.

In this background, Cahyani and Patasik [68] presented a text classification method that uses Support Vector Machines (SVM). The authors show that when using SVM classification with Term Frequency-Inverse Document Frequency (TF-IDF), the accuracy can reach 85.17 % and 93.45 % of average precision in the datasets Commuter Line and Transjakarta, respectively. As a note, TF-IDF is a one-word/term weighting method that assigns a different weight to each word in a document based on the frequency of words per document and the frequency of words in all documents. However, Cahyani and Patasik's only have the five expressions (namely, happy, angry, sad, fear, and surprised). Another example of a relevant model is ELMo, which was developed by AllenNLP [69] and trained on the one Billion Word Benchmark [70]. This model has two layers of bidirectional language model (biLM), which consist of two stacked layers of long short-term memory (LSTM) networks and use in character level CNN to convert words of a given string into word vectors. ELMo is not like a traditional word embedding, since it employs the complete sentence for computing word embeddings [69], where bidirectional input is calculated by characters instead of words. The bidirectional approach consists in considering a relation of a word with the next and previous words. The method allows to establish the differences between antonyms from synonyms, solving the challenge in traditional language models, such as, GloVe and FastText, because of the distributional information of words [69].

Proposed by Radford, the GPT-2 model is trained on 40 GB of text from the internet [71]. The architecture is based on 1.5 billion parameters, exploiting text from eight million websites. In the same work, [71], the authors use a collection of massive open online courses (MOOCs) videos lecture as a dataset and annotated them into the eight general-level categories (art and humanities, physical sciences and engineering, computer science, data science, business, information technology, health, and social science). In this case, the GPT-2 model was used to solve the problem of a highly imbalanced dataset.

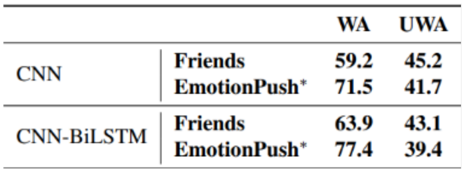
RoBERTa method [72] was developed by Facebook AI, being based on Google's Bidirectional Encoder Representations from Transformers (BERT). The method was designed to increase the comprehension of human language, being used in Google search. BERT is a neural network capable of learning human language forms of expression, it is based on an NLP model called Transformer that can understand the relationships between words in a sentence. The BERT model is a deep bidirectional pre-training which means that the dataset that the model knows will not forget when it is used to develop various systems. Facebook made some changes and retrained BERT for a longer period and with more data [72].

Models using transformers are also a known solution. The transformers are a novel architecture that is used to encode a sequence of textual tokens into a large vectors-based representations for each token [73]. These models are made up of encoder and decoder blocks with a *softmax* activation function to normalize the output probabilities [74]. In short, these models take as input a sequence of data, which are embedded and passed through to the positional encoders, assigning vectors to words based on their position in the sentence. The encoder captures the contextual relation between the words in a phrase and sends one attention vector at a time to the decoder. The decoder receives this attention vector and sends it to the linear layer and, finally, for the *softmax* activation function, which converts it into a probability distribution for the output. The transforms model was designed for machine translation, but it is also used for language modeling, making it applicable for other NLP tasks, such as, for text classification, document summarization, question answering, and others [74].

As before, three primary classifiers were used to build the text classification aggregator, namely: (a) The original project of the MultiModal-Emotion-Recognition (MMER) [79] method used the dataset MELD (Multimodal EMOTIONLINES Dataset) [75]. MELD is divided into three parts: text, audio, and video parts for the facial model [75]. (b) The End2End [76] primary classifier uses Neattext package for cleaning text, which means this package removes emails, numbers, emojis, and stop words in phrases [77]. The dataset used by the author of this project, Jesse Agbe, is unknown. The author made use of a csv with sentences available without reference to the name of the dataset. The author loads the csv with sentences, use the Neattext package to clean the sentences in the dataset, split the sentences 70 % to train and 30 % for the test, and used Sklearn library to train the model [76]. (c) The Gated Recurrent Unit (GRU) model [78] was trained and tested using the GRU authors' dataset, based on three sub-datasets, namely: Daily Dialogue, Emotion-Stimulus, and ISEAR. Only phrases that had the expressions joy, sadness, anger, fear, and neutral were selected from the dataset, that has a total 11,327 sentences [78]. To remove hashtags and usernames from sentences was used Regex [79], and to create tokenization of words it was used the library Natural Language Toolkit (NLTK) [80]. The tokenization created in this library just separates each word of the phrase into an array of words. After this is used a Tokenizer and *pad_sequences* from Keras before training the model [78].

Being an utterance a "bit" of spoken language (if you can hear then it is considered an utterance, if not it is not an utterance), the emotional expression of each utterance is dependent on the context. When it is classified as utterances, other utterances may provide important contextual information. To measure the flow, it applied to the contextual LSTM architecture [66]. Chen [66] conducted experiments on the EmotionLines dataset with a CNN model and a CNN-Bidirectional LSTM (CNN-BiLSTM) model. The performance of both models is evaluated by both the weighted accuracy (WA) and the unweighted accuracy (UWA). The attained results are shown in Table 4.1, varying the weighted accuracy between 59.2 % and 63.9 % on the Friends dataset, and between 71.5 % and 77.4 % on the Emotions dataset [66].

Table 4.1 - Weighted and unweighted accuracy of the methods proposed by Chen on Friends and EmotionPush dataset, adapted from [66].



		WA	UWA
CNN	**Friends**	**59.2**	**45.2**
	EmotionPush*	**71.5**	**41.7**
CNN-BiLSTM	**Friends**	**63.9**	**43.1**
	EmotionPush*	**77.4**	**39.4**

As a final note, the EMOTIONLINES [66] dataset was the first dataset with emotions labelling on all utterances, in each dialogue, and based on their textual content. Dialogues in EMOTIONLINES are collected from Friends TV scripts and private Facebook messenger dialogues. This dataset has a total of 29,245 utterances from 2,000 labelled, being that one dialogue can have several utterances (phrases) [66].

4.3 Text Emotions Classifier

Following once again Fig. 1.1, and as already mentioned, the classification focus is on Ekman’s [38] basic emotions plus neutral emotion. The framework for text classification is divided in 2 steps: (a) send the text to each of primary emotions classifiers (in this case, *n* = 3), Fig. 4.1 left side, each one returning a value between 0.0 and 1.0 for each of the seven emotions (neutral, joy, sadness, disgust, fear, anger, and surprise). (b) Then, those returned values are injected into the aggregation model, which means that each text determines *n* × 7 inputs into the aggregation model, to obtain a single final expression classification, Fig. 4.1 right side. In the latter step, the classifiers used for the aggregation implementation were: (1) Voting; (2) Random Forest (RF) [30]; (3) AdaBoost [31]; and (4) Multi-layer Perceptron/Neutral Network (MLP/NN) [32]. These models were already presented in detail in previous chapters and in Novais *et al.* [46], for instance.

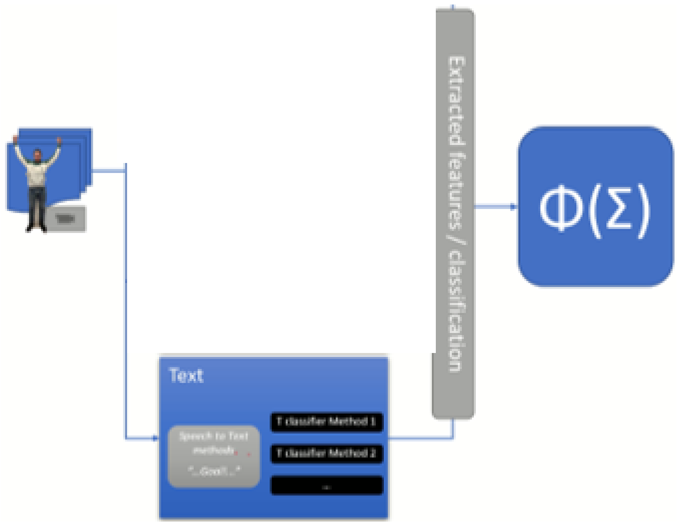


Figure 4.1 - Framework for text emotion classifier.

In the first step, (a), the following primary emotion classifiers were used: (1) End2End, with the code available in [76]; (2) MultiModal-Emotion-Recognition (MMER), a free implementation done by Ankur Bhatia and Ashwani Rathee that has its code available at [75]; and (3) GRU, which is a free implementation done by Lukas Garbas, with the code available in [78]. The main reasons to choose these initial methods as primary classifiers were the fact that they are the open-source methods and are focused on emotion classification, not on sentiment (most of the papers and code available about text expression classification focused on sentiment classification). As a note, the End2End model was trained using a LogisticRegression Pipeline from Sklearn [36] and the other models used the Keras library [81]. Furthermore, End2End does not need to prepare data to fit the model. On the other hand, models MMER and GRU need to create a tokenization of the data. The tokenization allows vectorizing a text corpus, by transforming each word into a sequence of integers, where each integer is the index of a token in a dictionary [82]. For instance, for the phrase: “Today is a beautiful day.”, the dictionary will have five values (punctuation does not count), one for each word of the phrase. When the sentence has only words that the model does not recognize, any of the primary models will return the neutral expression. After creating the tokenization, pre-trained word vectors are downloaded from *fasttext.cc*, containing 1 million words trained on Wikipedia 2017 and some sources of text data available online, including Gigaword dataset [83] [84]. These words would be compared with what exists on tokenization and, if the word does not exist then it would be added to the layer, which is called the embedding layer.

4.4 Test and Results

The three above-mentioned primary classifiers were trained with EMOTIONLINES dataset. The dataset was split into 70 % for train, 15 % for validation, and the last 15 % for testing. Training parameters can be seen in Tab. A.4 in Appendix A, and testing was conducted on a personal computer running Windows 10 over an AMD Ryzen 7 4800H @ 2.90 GHz with 16 GiB of RAM. The metrics used for the evaluation of the models were accuracy for the primary models (Tab. 4.2) and aggregator (Tab. 4.3).

Table 4.2 - Baseline results for the 3 primary methods.

	EMOTIONLINES		
	End2End	MMER	GRU
Accuracy	51.22 %	44.86 %	**53.72 %**

Table 4.3 - Accuracy for the ensemble methods.

		Without ranking			With ranking		
Num. methods	Voting	Random Forest	AdaBoost	MLP/ NN	Random Forest	AdaBoost	MLP/ NN
n = 3	**51.53%**	49.69%	47.20%	46.86%	45.86%	46.20%	48.50%
n = 2 (End2End & MMER)	44.32%	49.77%	49.23%	49.23%	***51.60%***	47.58%	51.45%

The best primary classifier is GRU (Tab. 4.2) with an accuracy of 53.72 %. The best result achieved with the aggregators is 49.69 % using the Random Forest without ranking and the best with ranking is MLP/NN, with 48.50 % of accuracy. The reasons for those worse results are not complete clear yet, but we can hypothesize on the influence of the low accuracy of the MMER method, which can influence negatively the aggregation results.

4.5 Conclusions

A first approach for a text emotion classification framework supported in an ensemble of several open-source codes for text emotion classification was presented in this chapter. Contrary to the facial (Chap. 2) and speech (Chap. 3) ensemble, the text classifier did not bring better results compared with the primary methods. The reasons for those worse results are not complete clear yet, but we can hypothesize on the influence of the low accuracy of the MMER method, which can influence negatively the aggregation results. Nevertheless, in general, the low accuracy is probably due to the dataset used to train the primary models and test the framework. This dataset has specific characteristics, namely it is a dialogue, but single sentences were sent to the classifiers. Finally, EMOTIONLINES is not a balanced dataset, once it has much more *joy* emotions samples than other emotions, and this also can help to decrease the quality of the results.

In future work, we intend to explore more datasets (balanced) as well as different types of datasets (not dialogue-based). Also, other methods (primary classifiers) can be tested to validate if a primary classifier with “very low” accuracy invalidates the aggregator result. Possible is also to explore different and more recent architectures, such as Bidirectional Encode Representations from Transformers (BERT) [85], once it presents in some situations as in [78] results of 0.83 in the F1 score.

5 EMOTION MULTI-SOURCE AGGREGATOR

Abstract. In the field of human-computer interaction (HCI), there is a subfield known as affective computing, that focus on helping the evolution of interaction between humans and robots. This evolution means that tasks that were normally performed by humans, can now be performed by robots. The interaction between humans and robots is more powerful when the robot is aware of the user's emotional state, becoming imperative if the machine is looking for answers depending on the user's mood. The emotion multi-source aggregator (EMsA) is a method for merging results from facial, speech, text, and other sources extracted from different primary emotions classifications. In this chapter, we describe the EMsA and present some results using the RAVDESS dataset, namely the 81.99 % accuracy achieved by EMsA using the combination of faces and speech.

keywords: Facial Emotion, Speech Emotion, Text Emotions, Classification, Ensembles, Machine Learning.

5.1 Introduction

Having the expression (emotions and sentiment) classification from audio and video can help many innovative technologies, from socially assistive robots to monitoring and diagnosis of "illnesses" without contact with a person [86]. Indeed, monitoring is becoming increasingly important for the treatment and management of chronic illnesses, neurological disorders, and mental health issues like diabetes, hypertension, asthma, autism spectrum, disorder, fatigue, depression etc. [86].

As several times mentioned, the final goal of this dissertation is to obtain from any source (e.g., video, sound, text) a classification for human emotions, which, for the time being, can based on facial, sound, or text individual analysis or, as explored in this chapter, a combined analysis of more of one of those sources. The next section presents a very brief state-of-the-art of similar methods and in the following section, our proposed method, the EMsA - emotion multi-source aggregator.

5.2 Related Work

Similar to the main goal of this dissertation, Ankur Bhatia [75] proposed a method which can get text, facial and sound emotions, and sentiment. The method uses the MELD (Multimodal EmotionLines Dataset) dataset [87] which is divided into three parts: text, audio, and video. In more detail, the Ankur Bhatia's method is divided into: (a) *Facial* – uses a deep neural network (DNN) called CAER-NET, which explores not only the facial emotion but also gets other context information. The network is composed by two sub-networks. The first sub-network includes two-stream encoding networks to extract feature of the face and context region, and the other sub-network is an adaptive fusion network to use features adaptively. (b) *Audio* – to extract some audio features, the authors used the OpenSmile Toolkit. The features used were MFCC, Chromagram-based, time spectral features, short-term energy, short-term entropy of energy, spectral centroid and spread, spectral entropy, spectral flux, and spectral roll-off. The method computes the feature in a 0.2 second time-window, moved in a 0.1 second step, and uses 16 kHz sample rate. (c) *Text* – uses a pre-trained GloVe embedding to convert text words into a vector, and a CNN and a LSTM network to extract features and for accurate classification. The extracted features are used as a tensor [88], [89] and then a model is created having its output as expression results and sentiment prediction [75].

Juan David Ortega [86] presented a cost-effective DNN architecture, which can learn a robust map between a subject's spontaneous and natural behaviors, from different sources of information and his emotional state. The authors used AVEC SEWA database [90] to learn the feature representation and to predict arousal, valence, and liking. Figure 5.1 presents the architecture of the proposed model, it has three different sources of information, namely, audio, video, and text, from which different features are extracted. In more detail: (a) *Facial* features are extracted for each video (at a frame step of 20 milliseconds), consisting of a normalized face orientation in degrees and pixel coordinates for 49 facial landmarks. (b) *Audio* features are extracted from acoustic low-level descriptors (LLDS): energy, spectral and cepstral, pitch, voice quality, and micro-prosodic. Finally, (c) *Text* features are made of a bag of words feature representation based on the transcription of the speech, the list contains 521 words.

In the first stage, the DNN [86] architecture processes each source (audio, video, and text) individually, using a pair of connected layers to generate correlations between features of the same type. The second stage merges the three outputs using the concatenation function. In the third stage, the layer receives the outputs (concatenation function) of the previous step in one block and feeds a fully connected layer. In the next stage, the DNN output is generated by a single linear neuron, used as a regression value. In the end, a scaling module is used to reduce the gap between the magnitudes of the predictions and the labels [86].

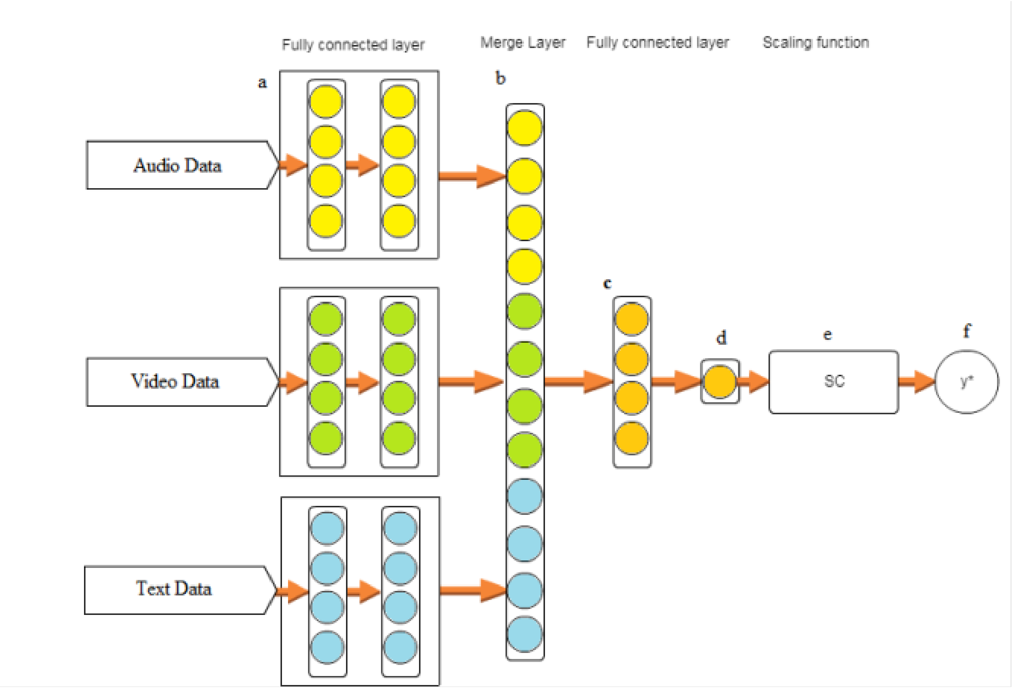


Figure 5.1 - Proposed DNN architecture for multi modal fusion: (a) independent layers; (b) merge layer; (c) fully-connected layer (adapted from [86]).

Siddiqui and Javaid [91] developed a framework for facial and speech emotion classification. The framework proposed consists of two layers of detection and three CNNs. In the first layer, two CNNs are trained using visible and infrared images individually and, after, the obtained features are fed to an SVM for classification method. For speech, the CNN was deployed to learn the emotions using features obtained by audio spectrograms. Figure 5.2 illustrates the authors' method, being a detailed description of it available in [91].

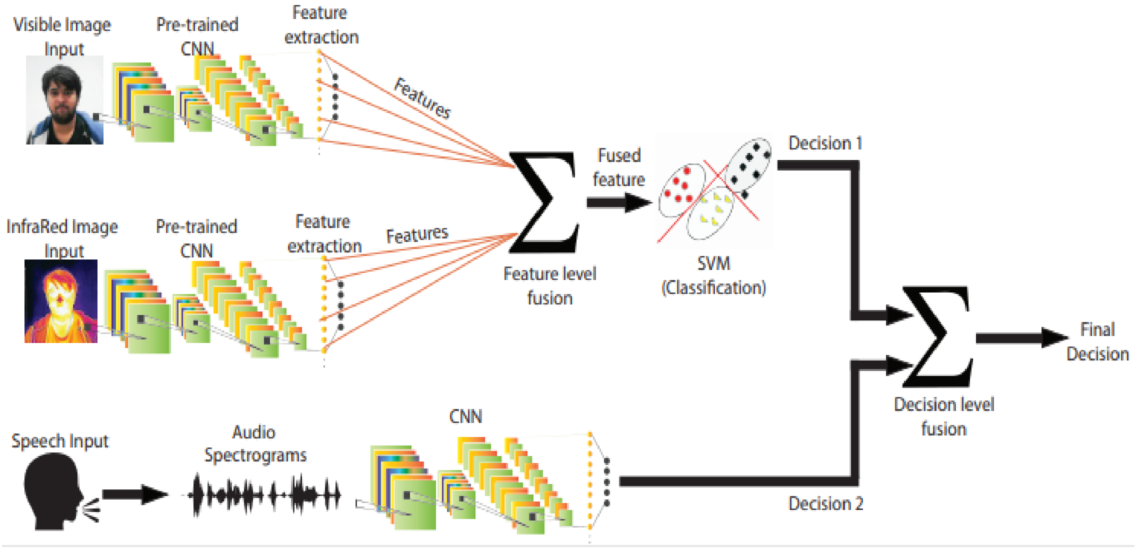


Figure 5.2 - Proposed model by Siddiqui and Javaid, adapted from [91].

Despite the presented methods having the same main goal of the dissertation, the way to achieve the goal are completely different. All the above authors focus in to develop a single model, i.e., teaching a new model from the scratch. Instead, we focus on using already developed open-source primary classifiers and composing (aggregating) those in a final classification method.

5.3 Multi-Source Aggregator and Sentiment Classifier

To demonstrate the final proof-of-concept, the EMsA was developed, which is an extension of the aggregators presented in previous chapters (see Chaps. 2 to 4). Figure 5.3 illustrates the full model. Now, the outputs from the primary classifiers for each source are sent to the aggregator, again implemented/tested using, for instance, Random Forest, AdaBoost, or MLP/NN as classifiers.

Each aggregator will receive 63 values ($7 \times n \times 3$), where 7 is the number of expressions (happy, sad etc.) *per* primary classifier, n is the number of classifiers per source (in our case we always use $n = 3$), and 3 is the number of sources (facial, speech, and text).

Finally, it is important to stress the sentiment classification. For the sentiment classification, we followed the work done by Roberts *et al.* [92] and later by Nahar *et al.* [93], where they mapped Ekman's six basic emotions, plus "love", to the sentiments of positive, negative, and unexpected. Table 5.1 illustrates this mapping, where the emotion “love” is not presented once it was not used in our studies. In addition, only for consistency, we mapped the “neutral” emotion to a sentiment “neutral”.

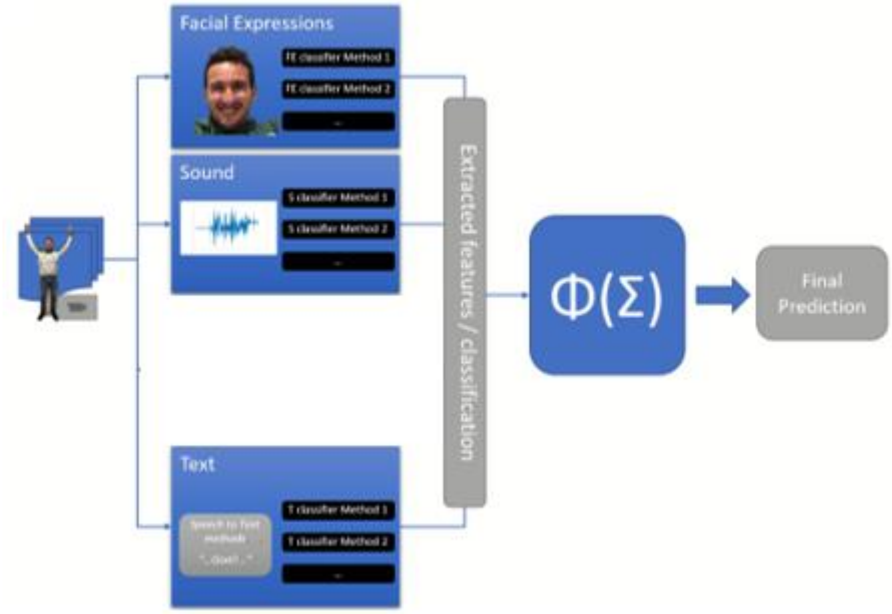


Figure 5.3 - Framework for emotion classifier – EMsA.

This mapping is used/necessary only once after the emotion is detected, as this is a direct procedure, i.e., every time the system returns an emotion it also automatically returns the corresponding sentiment. It is also important to mention that, in the case of *fear* the system returns two sentiments but, anyhow, this procedure does not require any additional learning of the system (once it is a direct mapping from emotion to sentiment).

Table 5.1 - Sentiment and emotions mapping.

	Emotions						
Sentiment	Anger	Happiness	Surprise	Disgust	Sadness	Fear	Neutral
Positive		×					
Negative	×			×	×	×	
Unexpected			×			×	
Neutral							×

Finally, it is important to point out that this is not the only solution. In fact, there are several procedures to relate emotions and sentiments, and also different ways to classify emotions (with a higher and different number of classes). A very recent (March 2022) and complete systematic review on affective computing, where detailed emotion models, databases, and sentiment models are analyzed, can be found in the work of Wang *et al.* [94].

To conclude and recall, the goal of the dissertation is not to build a new model for emotion classification (or sentiment classification), but to develop a framework for emotion classification based on facial, speech, and text emotion prediction, supported on an ensemble of open-source codes.

5.4 Tests and Results

As a first note, unfortunately, as far as we could see, it was not possible to find a dataset to test our framework with the 3 sources (image, sound, and text). So, the EMsA framework will only be tested with image and sound (speech) input, meaning that the aggregators will receive 7 × 3 × 2 = 42 inputs from the primary classifiers. As before, our tests will use as alternative aggregators the Random Forest, AdaBoost, and MLP/NN methods.

The main reason why EMsA was not tested with the MELD and SEWA datasets is compatibility problems between the datasets and the implemented framework. In particular, the MELD dataset, which is an extension of the EMOTIONLINES dataset, is a dialogue between 2 persons [87]. On the other hand, the SEWA dataset was not chosen because it does not have the six Ekman’s basic emotions, having, for instance, valence, arousal, and liking and disliking annotations [90].

So, to test EMsA, we use RAVDESS. As previously, the dataset was (stratified-) split into 70 % for training, 15 % for validation, and the last 15 % for testing. Training parameters can be seen in Tab. A.5 (Appendix A), and testing was conducted on a personal computer running Windows 10 over an AMD Ryzen 7 4800H @ 2.90 GHz with 16 GB of RAM. The metric used for the evaluation of the models was accuracy.

It is important to stress that, the three primary classifiers for face emotion classification were trained using FER dataset (see Chapter 2) and the three primary classifiers for speech were trained using RAVDESS (see Chapter 3). It is also important to stress that data used from RAVDESS to train the speech were not used to validate (test) the final multi-source aggregator. Nevertheless, the EMsA aggregator was trained using RAVDESS data (image and speech).

As a final note, we had to adapt the facial emotion classification previously developed and prepare it to deal with videos, instead of single static images, as RAVDESS is composed by movies clips. The process included the following steps. For each clip, the

(i) first 30 frames (1 second) and the (ii) last 30 frames (1 second) were discarded. Then, for the (iii) remaining frames it was applied the primary classifiers to each one, followed by a (iv) non-maximum suppression technique, i.e., over the remaining results of the clip a sliding neighborhood window with similar emotions are considered as candidate classes, which leads to several proposals. It was considered the proposal/emotion with the highest count.

Table 5.2 shows the baseline results for the aggregation methods, as explained above, and implemented as mentioned in Chap. 2 for the facial emotion classification – "face", and as mentioned in Chap. 3 for the speech emotion classification – "speech". In Chap. 3, the three versions of primary models were created. As the third version (***3Dataset***) had better accuracy, the three primary models of this version were selected to be used in this chapter. The speech results are lower compared to the results of Tab. 3.3 (***3Dataset***) because the files used for testing were .mp4 extension and had to be converted to .wav file, thus losing some sound quality, resulting in a loss of accuracy in the model.

Table 5.2 - Baseline results using RAVDESS dataset.

		Without ranking			With ranking		
Type of expression	Voting	Random Forest	AdaBoost	MLP/ NN	Random Forest	AdaBoost	MLP/ NN
Face	51.77%	**62.70%**	47.90%	58.52%	58.52%	55.95%	56.27%
Speech	68.49 %	72.02 %	67.84%	**74.28%**	69.77%	60.77%	71.70%

Finally, Tab. 5.3 summarizes the results for the EMsA. The best results were achieved using MLP/NN without ranking, with an accuracy of 81.99 %, more 19.29 % than the best result for facial emotion (it was 62.70 %) and more 7.71 % than the best result for speech emotion classification (it was 74.28 %).

These results clearly demonstrate that the contribution of an ensemble of open-source methods can easily be improved using an aggregator methodology like the one presented.

Table 5.3 - Results of the EMsA.

	Without ranking			With ranking		
Dataset	Random Forest	AdaBoost	MLP/ NN	Random Forest	AdaBoost	MLP/ NN
RAVDESS	80.38%	71.38%	**81.99%**	80.71%	68.49 %	79.42%

5.5 Conclusions

It has been presented a framework based on a multi-source aggregator (EMsA), which aggregates the results extracted from the primary emotion classifications from different sources, such as facial, speech, and text. The framework was tested using the RAVDESS dataset and validated the initial prove-of-concept. It was proved that it is possible to build a state-of-the-art system, supported on methods (primary classifiers) available as open-source, with a minimum training of an aggregator and improving the base accuracy. This allows, of course, minimize the carbon footprint of the algorithm (framework). Another advantage of this solution is to allow the speed-up in the development of an integrated solution for human emotion classifiers.

Also presented was an initial approach for sentiment classification, based on direct mapping of the classified emotion to the sentiment.

Future work should focus mainly, and more importantly, on the discovery of a dataset, or on the building of it, that will allow testing all the strands of the model. In addition, remaining modules mentioned in the dissertation (which at the time being we considered out of the focus of the dissertation) should be implemented to complete the presented framework. In those modules, body expression classification and the relation between the environment and human expression (emotion and sentiment) are included.

6 CONCLUSION AND FUTURE WORK

Abstract. We start this dissertation by stating that “Humans are prepared to comprehend each others’ emotions through subtle body movements, speech or facial expressions. From those, they change how they deliver messages when communicating between them. Machines, user interfaces, or robots need to empower this ability, in a way to change the interaction from the traditional “human-computer interaction” to a “human-machine cooperation”, where the machine provides the “right” information and functionality, at the “right” time, and in the “right” way.” In this dissertation, we give another small step to achieve this, by presenting a framework that allows the classification of emotion using several sources individually or in conjunction.

keywords: Facial Emotion, Speech Emotion, Text Emotions, Ensembles, Machine Learning

6.1 Final Conclusions

Since partial conclusions were presented in all chapters, this section focuses on the main and overall conclusion. In this dissertation, a framework based on a multi-source aggregator was presented, which aggregates the results extracted from the primary emotion classifications from different sources, such as facial, speech, and text. The framework was tested using the RAVDESS dataset and validated the proof-of-concept: it is possible, only using methods available as open-source (primary classifiers), with a minimum training of an aggregator, to achieve better accuracy than using the primary classifiers. This allows, of course, minimizing the carbon footprint of the algorithm (framework) and man-hours necessary to design and implement it. The EMsA best results, over face and sound data retrieved from video clips, are achieved using MLP/NN without ranking, with an accuracy of 81.99 %.

Another advantage, as already mentioned, is to allow the speed-up in the development of an integrated solution for human emotion classifiers. Also presented was an initial approach for sentiment classification, based on direct mapping of the classified emotion to the sentiment.

Additionally, presented for each different source, an initial aggregation of primary classifiers, for facial emotion classification, achieved an accuracy above 73 % in both FER2013 and RAF-DB datasets. For the speech emotion classification, it were used four datasets (namely, RAVDESS, TESS, CREMA-D, and SAVEE). For this latter source, the aggregation of primary classifiers achieved, in a combination of three of the mentioned datasets, results above 86 % of accuracy. The text emotion aggregation of primary classifiers was tested with one dataset called EMOTIONLINES, the classification of emotions achieved an accuracy above 53 %. This module clearly needs to be improved in the future.

In summary, the main goals of the dissertation were achieved, and the ensemble of off-the-shelf methods is a promising solution, as they provide a solution to improve the performance achieved by open-source primary methods.

6.2 Future Work

As partial conclusions were summarized in the Final Conclusions section, partial future work was also presented in all chapters and will be summarized here. In this context, we focus on three main items: (i) we will intend to explore different datasets, and to improve the aggregator's accuracy, by increasing the number of emotion classifiers and studying the influence of their characteristics, like the fact that they are over or under-fitted.

(ii) In addition, the remaining modules mentioned in the dissertation (but out of the focus of the dissertation) should be implemented in a way to complement the presented framework, namely, body expressions and the relation between the environment and human expression.

Finally, (iii) the first suggestion for this dissertation was to add Active Learning to the ensemble pipeline. However, this was quickly forgotten due to a lack of time to complete this task in the timeframe available for the development of the dissertation. Nevertheless, Active Learning can play an important role as not many completely suitable dataset seem to be available at the moment.

6.3 Publications

The dissertation resulted in two publications in international conferences, namely:

- Novais, R., Cardoso, P. J. S., Rodrigues, J. M. F. (2022). Emotion Classification from Speech by an Ensemble Strategy. 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, Lisbon, Portugal, 31 Aug. 31 - 2 Sept.
- Novais, R., Cardoso, P. J. S., Rodrigues, J. M. F. (2022). Facial Emotions Classification Supported in an Ensemble Strategy. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. Novel Design Approaches and Technologies. HCII 2022. Lecture Notes in Computer Science, vol 13308. Springer, Cham. https://doi.org/10.1007/978-3-031-05028-2_32

As a final note, one more publication that integrates all the work including the multi-source Aggregator is being prepared. The publication is expected to be submitted in December to the 25th International Conference on Human-Computer Interaction, to be held in Copenhagen, Denmark, 23-28 July 2023.

7 REFERENCES

[1] L. Schoneveld, A. Othmani, and H. Abdelkawy, “Leveraging recent advances in deep learning for audio-Visual emotion recognition,” *Pattern Recognit Lett*, vol. 146, pp. 1–7, Jun. 2021, doi: 10.1016/j.patrec.2021.03.007.

[2] A. Deguchi *et al.*, *Society 5.0*. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-15-2989-4.

[3] S. Rothfus, M. Worner, J. Inga, and S. Hohmann, “A Study on Human-Machine Cooperation on Decision Level,” in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2020, pp. 2291–2298. doi: 10.1109/SMC42975.2020.9282813.

[4] P. Kumar and A. Gupta, “Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey,” *J Comput Sci Technol*, vol. 35, no. 4, pp. 913–945, Jul. 2020, doi: 10.1007/s11390-020-9487-4.

[5] S. Ardabili, A. Mosavi, and A. R. Várkonyi-Kóczy, “Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods,” in *Lecture Notes in Networks and Systems*, vol. 101, 2020, pp. 215–227. doi: 10.1007/978-3-030-36841-8_21.

[6] F. Zhang, T. Zhang, Q. Mao, and C. Xu, “Joint Pose and Expression Modeling for Facial Expression Recognition,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 3359–3368. doi: 10.1109/CVPR.2018.00354.

[7] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, “Survey on Emotional Body Gesture Recognition,” *IEEE Trans Affect Comput*, vol. 12, no. 2, pp. 505–523, Oct. 2018.

[8] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, “Fusing Body Posture with Facial Expressions for Joint Recognition of Affect in Child-Robot Interaction,” *IEEE Robot Autom Lett*, vol. 4, no. 4, pp. 4011–4018, Jul. 2019, doi: 10.1109/LRA.2019.2930434.

[9] S. Leiva, L. Margulis, A. Micciulli, and A. Ferreres, “Dissociation between facial and bodily expressions in emotion recognition: A case study,” *Clin Neuropsychol*, vol. 33, no. 1, pp. 166–182, Jan. 2019, doi: 10.1080/13854046.2017.1418024.

[10] D. Canedo and A. Neves J. R., “Mood estimation based on facial expressions and postures,” in *Proceedings of the RECPAD*, 2020, pp. 49–50.

[11] T. Bänziger, M. Mortillaro, and K. R. Scherer, “Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception.,” *Emotion*, vol. 12, no. 5, pp. 1161–79, Oct. 2012, doi: 10.1037/a0025827.

[12] A. Kleinsmith and N. Bianchi-Berthouze, “Affective Body Expression Perception and Recognition: A Survey,” *IEEE Trans Affect Comput*, vol. 4, no. 1, pp. 15–33, Jan. 2013, doi: 10.1109/T-AFFC.2012.16.

[13] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, “Continuous body emotion recognition system during theater performances,” *Comput Animat Virtual Worlds*, vol. 27, no. 3–4, pp. 311–320, May 2016, doi: 10.1002/cav.1714.

[14] F. Ahmed, A. S. M. H. Bari, and M. L. Gavrilova, “Emotion Recognition From Body Movement,” *IEEE Access*, vol. 8, pp. 11761–11781, 2020, doi: 10.1109/ACCESS.2019.2963113.

[15] G. Liang, S. Wang, and C. Wang, “Pose-aware Adversarial Domain Adaptation for Personalized Facial Expression Recognition,” *ArXiv: 2007.05932*, Jul. 2020.

[16] P. Ekman and W. v. Friesen, “Constants across cultures in the face and emotion.,” *J Pers Soc Psychol*, vol. 17, no. 2, pp. 124–129, 1971, doi: 10.1037/h0030377.

[17] T. H. H. Zavaschi, A. L. Koerich, and L. E. S. Oliveira, “Facial expression recognition using ensemble of classifiers,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 1489–1492. doi: 10.1109/ICASSP.2011.5946775.

[18] A. Renda, M. Barsacchi, A. Bechini, and F. Marcelloni, “Comparing ensemble strategies for deep learning: An application to facial expression recognition,” *Expert Syst Appl*, vol. 136, pp. 1–11, Dec. 2019, doi: 10.1016/j.eswa.2019.06.025.

[19] G. Ali *et al.*, “Artificial Neural Network Based Ensemble Approach for Multicultural Facial Expressions Analysis,” *IEEE Access*, vol. 8, pp. 134950–134963, 2020, doi: 10.1109/ACCESS.2020.3009908.

[20] Z. Wang, F. Zeng, S. Liu, and B. Zeng, “OAENet: Oriented attention ensemble for accurate facial expression recognition,” *Pattern Recognit*, vol. 112, p. 107694, Apr. 2021, doi: 10.1016/j.patcog.2020.107694.

[21] N. K. Benamara *et al.*, “Real-time facial expression recognition using smoothed deep neural network ensemble,” *Integr Comput Aided Eng*, vol. 28, no. 1, pp. 97–111, Dec. 2020, doi: 10.3233/ICA-200643.

[22] R. Pecoraro, V. Basile, V. Bono, and S. Gallo, “Local Multi-Head Channel Self-Attention for Facial Expression Recognition,” *ArXiv:abs/2111.07224*, Nov. 2021.

[23] I. J. Goodfellow *et al.*, “Challenges in Representation Learning: A report on three machine learning contests,” *International conference on neural information processing. Springer, ArXiv:1307.0414*, pp. 117–124, Jul. 2013.

[24] J. H. Cheong, T. Xie, S. Byrne, and L. J. Chang, “Py-Feat: Python Facial Expression Analysis Toolbox,” *arXiv preprint ArXiv:2104.03509*, Apr. 2021.

[25] R. Banerjee, S. De, and S. Dey, “A Survey on Various Deep Learning Algorithms for an Efficient Facial Expression Recognition System,” *Int J Image Graph*, Dec. 2021, doi: 10.1142/S0219467822400058.

[26] I. M. Revina and W. R. S. Emmanuel, “A Survey on Human Face Expression Recognition Techniques,” *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, Jul. 2021, doi: 10.1016/j.jksuci.2018.09.002.

[27] “LHC-NET (2021). Local multi-head channel self-attention (code). https://github.com/bodhis4ttva/lhc_net, accessed 2021/12/28.”

[28] “Py-FEAT (2021) Python facial expression analysis toolbox (code). https://pythonrepo.com/repo/cosanlab-py-feat-python-deep-learning, accessed 2021/12/28.”

[29] “Shenk, J. (2021) Facial expression recognition (code). https://github.com/justinshenk/fer, accessed 2021/12/28.” doi: https://doi.org/10.5281/zenodo.5362356.

[30] L. Breiman, “Random forests. Machine learning,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[31] T. Hastie, S. Rosset, J. Zhu, and H. Zou, “Multi-class AdaBoost,” *Stat Interface*, vol. 2, no. 3, pp. 349–360, 2009, doi: 10.4310/SII.2009.v2.n3.a8.

[32] V. K. Ayyadevara, *Pro Machine Learning Algorithms*. Berkeley, CA: Apress, 2018. doi: 10.1007/978-1-4842-3564-5.

[33] “FER2013 (2021). Learn facial expressions from an image, https://www.kaggle.com/msambare/fer2013, accessed 2021/12/28.”

[34] “RAF-DB (2021). Real-world affective faces database, http://www.whdeng.cn/raf/model1.html, accessed 2021/12/28.”

[35] "OpenCV (2021). OpenCV: Cascade classifier – face detection, https://docs.opencv.org/4.5.5/db/d28/tutorial_cascade_classifier.html, accessed 2021/12/28."

[36] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *machine Learning research*, vol. 12, pp. 2825–2830, Jan. 2012.

[37] S. Li and W. Deng, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, Jan. 2019, doi: 10.1109/TIP.2018.2868382.

[38] P. Ekman, "Facial Expressions of Emotion: New Findings, New Questions," *Psychol Sci*, vol. 3, no. 1, pp. 34–38, Jan. 1992, doi: 10.1111/j.1467-9280.1992.tb00253.x.

[39] H. Kaur, V. Mangat, and Nidhi, "A survey of sentiment analysis techniques," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Feb. 2017, pp. 921–925. doi: 10.1109/I-SMAC.2017.8058315.

[40] J. Abdi, A. Al-Hindawi, T. Ng, and M. P. Vizcaychipi, "Scoping review on the use of socially assistive robot technology in elderly care," *BMJ Open*, vol. 8, no. 2, p. e018815, Feb. 2018, doi: 10.1136/bmjopen-2017-018815.

[41] M. Kyrarini *et al.*, "A Survey of Robots in Healthcare," *Technologies (Basel)*, vol. 9, no. 1, p. 8, Jan. 2021, doi: 10.3390/technologies9010008.

[42] C. Getson and G. Nejat, "Socially Assistive Robots Helping Older Adults through the Pandemic and Life after COVID-19," *Robotics*, vol. 10, no. 3, p. 106, Sep. 2021, doi: 10.3390/robotics10030106.

[43] J. Li, Z. Lin, P. Fu, Q. Si, and W. Wang, "A Hierarchical Transformer with Speaker Modeling for Emotion Recognition in Conversation," *ArXiv preprint ArXiv:2012.14781*, Dec. 2020.

[44] H. Abdollahi, M. Mahoor, R. Zandie, J. Sewierski, and S. Qualls, "Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study," *IEEE Trans Affect Comput*, 2022, doi: 10.1109/TAFFC.2022.3143803.

[45] A. Sorrentino, G. Mancioppi, L. Coviello, F. Cavallo, and L. Fiorini, "Feasibility Study on the Role of Personality, Emotion, and Engagement in Socially Assistive Robotics: A Cognitive Assessment Scenario," *Informatics*, vol. 8, no. 2, p. 23, Mar. 2021, doi: 10.3390/informatics8020023.

[46] R. Novais, P. J. S. Cardoso, and J. M. F. Rodrigues, “Facial Emotions Classification Supported in an Ensemble Strategy,” in *Lecture Notes in Computer Science,* vol. 13308, 2022, pp. 477–488. doi: 10.1007/978-3-031-05028-2_32.

[47] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.

[48] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset,” *IEEE Trans Affect Comput*, vol. 5, no. 4, pp. 377–390, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.

[49] P. Jackson, S. Haq, and J. Edge, “Audio-Visual Feature Selection and Reduction for Emotion Classification. In Proc. Int’l Conf. on Auditory-Visual Speech Processing, pp. 185-190, 2008.”

[50] K. Dupuis and M. K. Pichora-Fuller, “Toronto emotional speech set (TESS), 2020, Borealis, doi: https://doi.org/10.5683/SP2/E8H2MF.”

[51] A. S. Popova, A. G. Rassadin, and A. A. Ponomarenko, “Emotion Recognition in Sound,” in *Advances in Neural Computation, Machine Learning, and Cognitive Research* , vol. 736, 2018, pp. 117–124. doi: 10.1007/978-3-319-66604-4_18.

[52] M. Chen, X. He, J. Yang, and H. Zhang, “3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition,” *IEEE Signal Process Lett*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[53] K. Palanisamy, D. Singhania, and A. Yao, “Rethinking CNN Models for Audio Classification,” *ArXiv preprint ArXiv:2007.11154*, Jul. 2020.

[54] M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, “Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients,” in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, May 2020, pp. 1–5. doi: 10.1109/EAIS48028.2020.9122698.

[55] M. el Seknedy and S. Fawzi, “Speech Emotion Recognition System for Human Interaction Applications,” in *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS),* Dec. 2021, pp. 361–368. doi: 10.1109/ICICIS52592.2021.9694246.

[56] U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, “Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN,” *Int J Speech Technol*, vol. 24, no. 2, pp. 303–314, Jun. 2021, doi: 10.1007/s10772-020-09792-x.

[57] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models,” *Sensors*, vol. 21, no. 4, p. 1249, Feb. 2021, doi: 10.3390/s21041249.

[58] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, “A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism,” *Electronics (Basel)*, vol. 10, no. 10, p. 1163, May 2021, doi: 10.3390/electronics10101163.

[59] S. Prasanth, M. Roshni Thanka, E. Bijolin Edwin, and V. Nagaraj, “Speech emotion recognition based on machine learning tactics and algorithms,” *Mater Today Proc*, Feb. 2021, doi: 10.1016/j.matpr.2020.12.207.

[60] M. G. de Pinto, “Audio Emotion Classification from Multiple Datasets, https://github.com/marcogdepinto/emotion-classification-from-audio-files, accessed 2022/05/02.” 2020.

[61] S. Burnwal, “Speech Emotion Recognition, https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition/notebook. accessed 2022/05/02.” 2020.

[62] S. Tripathi, S. Tripathi, and H. Beigi, “Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning,” *IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, pp. 328–333, Apr. 2018.

[63] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, “Emotion detection from text and speech: a survey,” *Soc Netw Anal Min*, vol. 8, no. 1, p. 28, Dec. 2018, doi: 10.1007/s13278-018-0505-2.

[64] E. Batbaatar, M. Li, and K. H. Ryu, “Semantic-Emotion Neural Network for Emotion Recognition From Text,” *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.

[65] S. Joshi and D. Deshpande, “Twitter Sentiment Analysis System,” *Int J Comput Appl*, vol. 180, no. 47, pp. 35–39, Jun. 2018, doi: 10.5120/ijca2018917319.

[66] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, Ting-Hao, Huang, and L.-W. Ku, “EmotionLines: An Emotion Corpus of Multi-Party Conversations,” *ArXiv preprint ArXiv:1802.08379*, Feb. 2018.

[67] R. Novais, P. J. S. Cardoso, and J. M. F. Rodrigues, “Emotion Classification from Speech by an Ensemble Strategy,” *10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*, Aug. 2022.

[68] D. E. Cahyani and I. Patasik, “Performance comparison of TF-IDF and Word2Vec models for emotion text classification,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/eei.v10i5.3157.

[69] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, “A multimodal hierarchical approach to speech emotion recognition from audio and text,” *Knowl Based Syst*, vol. 229, p. 107316, Oct. 2021, doi: 10.1016/j.knosys.2021.107316.

[70] “ELMo, https://allenai.org/allennlp/software/elmo, accessed in 6/06/2022.”

[71] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, “Towards Improved Classification Accuracy on Highly Imbalanced Text Dataset Using Deep Neural Language Models,” *Applied Sciences*, vol. 11, no. 2, p. 869, Jan. 2021, doi: 10.3390/app11020869.

[72] J. Choudrie, S. Patil, K. Kotecha, N. Matta, and I. Pappas, “Applying and Understanding an Advanced, Novel Deep Learning Approach: A Covid 19, Text Based, Emotions Analysis Study,” *Information Systems Frontiers*, vol. 23, no. 6, pp. 1431–1465, Dec. 2021, doi: 10.1007/s10796-021-10152-6.

[73] S. E. Friedman, I. H. Magnusson, and S. M. Schmer-Galunder, “Extracting Qualitative Causal Structure with Transformer-Based NLP,” *ArXiv: 2108.13304*, Aug. 2021.

[74] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, “Transformer models for text-based emotion detection: a review of BERT-based approaches,” *Artif Intell Rev*, vol. 54, no. 8, pp. 5789–5829, Dec. 2021, doi: 10.1007/s10462-021-09958-2.

[75] A. Bhatia and A. Rathee, “Multimodal-Emotion-Recognition, https://github.com/ankurbhatia24/MULTIMODAL-EMOTION-RECOGNITION, accessed 2022/06/10”.

[76] Jc. Jesse, “End2End, https://github.com/Jcharis/end2end-nlp-project, accessed 2022/06/22.”

[77] “Neattext, https://pypi.org/project/neattext/, accessed in 2022/08/31.”

[78] L. Garbas, “NLP Text Emotion, https://github.com/lukasgarbas/nlp-text-emotion, accessed in 2022/06/06.”

[79] "Regex, https://regexr.com/, accessed in 2022/08/20."

[80] "Natural language Toolkit, https://www.nltk.org/, accessed in 2022/08/20."

[81] "Keras, https://keras.io/, accessed in 5/08/2022."

[82] Keras, "Keras PreProcessing, https://github.com/keras-team/keras-preprocessing/blob/master/keras_preprocessing/text.py, accessed in 2022/08/13."

[83] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "English Word Vectors, https://fasttext.cc/docs/en/english-vectors.html, accessed in 2022/08/13."

[84] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in Pre-Training Distributed Word Representations," *ArXiv: 1712.09405*, Dec. 2017.

[85] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[86] J. D. S. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition," *ArXiv: 1907.03196*, Jul. 2019.

[87] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations," in *Association for Computational Linguistics, ArXiv: 1810.02508*, Oct. 2018, pp. 527–536.

[88] A. LI, Y. LI, and X. LI, "TensorFlow and Keras-based Convolutional Neural Network in CAT Image Recognition," *DEStech Transactions on Computer Science and Engineering*, no. cmsam, Dec. 2017, doi: 10.12783/dtcse/cmsam2017/16428.

[89] R. Singhla, P. Singh, R. Madaan, and S. Panda, "Image Classification Using Tensor Flow," in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Mar. 2021, pp. 398–401. doi: 10.1109/ICAIS50930.2021.9395939.

[90] J. Kossaifi *et al.*, "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild," *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 3, pp. 1022–1040, Jan. 2019, doi: 10.1109/TPAMI.2019.2944808.

[91] M. F. H. Siddiqui and A. Y. Javaid, "A Multimodal Facial Emotion Recognition Framework through the Fusion of Speech with Visible and Infrared Images," *Multimodal Technologies and Interaction*, vol. 4, no. 3, p. 46, Aug. 2020, doi: 10.3390/mti4030046.

[92] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, “EmpaTweet: Annotating and Detecting Emotions on Twitter,” in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, 2012, pp. 3806–3813.

[93] L. Nahar, Z. Sultana, N. Iqbal, and A. Chowdhury, “Sentiment Analysis and Emotion Extraction: A Review of Research Paradigm,” in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, May 2019, pp. 1–8. doi: 10.1109/ICASERT.2019.8934654.

[94] Y. Wang *et al.*, “A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances,” *Information Fusion*, vol. 83–84, pp. 19–52, Mar. 2022.

APPENDIX A

This appendix summarizes the parameters used for Random Forest, AdaBoost, and MLP/Neural Network for the results presented during this dissertation, for each source and dataset.

Table A.1 - Grid search parameters (although the majority of the naming of the parameters is self-explicative, we suggest that the readers refer to the library's documentation, [36] for a more detailed explanation).

Random Forest	
Number of trees in the forest (n_estimators)	{25, 50, 100, 500}
Function to measure the quality of a split. (criterion)	{gini, entropy}
Maximum depth of the tree(max_depth)	{None, 2, 5, 10, 20}
Minimum number of samples required to split an internal node (min_samples_split)	{2, 5, 10}
Minimum number of samples required to be at a leaf node (min_samples_leaf)	{1, 2, 5, 10}
Number of features to consider when looking for the best split (max_features)	{1, 2, sqrt, log2}
Number of samples to draw from X to train each base estimator (max_samples)	{None, 0.1}
AdaBoost	
The maximum number of estimators at which boosting is terminated (n_estimators)	{25, 50, 100, 500}
Boosting algorithm (algorithm)	{SAMME, SAMME.R}
MLP/Neural Network	
The i-th element represents the number of neurons in the i-th hidden layer (hidden_layer_sizes)	{(10,), (100,), (10, 10), (100, 100), (10, 10, 10), (100, 100, 100)}
Learning rate schedule for weight updates (activation)	{identity, logistic, tanh, relu}
L2 penalty (regularization term) parameter (alpha)	{$10^{-3}, 10^{-2}, \ldots, 10^{3}$}
Learning rate schedule for weight updates (learning_rate)	{constant, invscaling, adaptive}

Table A.2 - Sets of parameters used to obtain the results for the different facial models (tuned using grid search stratified cross-validation).

		With ranking		**Without ranking**	
		FER2013	***RAF-DB***	***FER2013***	***RAF-DB***
Random Forest	n_estimators	100	500	50	100
	criterion	gini	entropy	gini	gini
	max_depth	10	None	None	20
	min_samples_split	10	5	2	10
	min_samples_leaf	10	1	10	1
	max_features	Sqrt	2	2	1
	max_samples	0.1	None	None	None
AdaBoost	n_estimators	500	50	500	500
	algorithm	SAMME.R	SAMME	SAMME.R	SAMME.R
MLP/NN	hidden_layer_sizes	(100, 100, 100)	(100,)	(10, 10)	(100,)
	activation	identity	relu	tanh	tanh
	alpha	0.01	0.1	0.1	1
	learning_rate	invscaling	constant	constant	constant

Table A.3 - Sets of parameters used to obtain the results for the different speech models (tuned using grid search stratified cross-validation).

RAVDESS		**With ranking**	**Without ranking**
		RAVDESS	***RAVDESS***
Random Forest	n_estimators	50	500
	criterion	entropy	gini
	max_depth	5	None
	min_samples_split	10	2
	min_samples_leaf	2	1
	max_features	1	1
	max_samples	0.1	None
AdaBoost	n_estimators	100	50
	algorithm	SAMME.R	SAMME
MLP/NN	hidden_layer_sizes	(100,)	(10,)
	activation	identity	identity
	alpha	1	0.01
	learning_rate	Invscaling	adaptive

RAVDESS + CREMA-D + SAVEE + TESS		**With ranking**	**Without ranking**
		All datasets	***All datasets***
Random Forest	n_estimators	500	500
	criterion	entropy	gini
	max_depth	20	None
	min_samples_split	5	10
	min_samples_leaf	10	1
	max_features	2	1
	max_samples	0.1	None
AdaBoost	n_estimators	100	50
	algorithm	SAMME	SAMME.R
MLP/NN	hidden_layer_sizes	(100, 100)	(10, 10)
	activation	logistic	relu
	alpha	0.01	0.01
	learning_rate	invscaling	constant

RAVDESS + SAVEE + TESS		**With ranking**	**Without ranking**
		3Datasets	***3Datasets***
Random Forest	n_estimators	500	100
	criterion	gini	gini
	max_depth	None	None
	min_samples_split	5	10
	min_samples_leaf	2	5
	max_features	2	1
	max_samples	0.1	None
AdaBoost	n_estimators	500	25
	algorithm	SAMME.R	SAMME
MLP/NN	hidden_layer_sizes	(100, 100, 100)	(10,)
	activation	identity	identity
	alpha	1	1
	learning_rate	Constant	constant

Table A.4 - Sets of parameters used to obtain the results for the different text models (tuned using grid search stratified cross-validation).

		With ranking	**Without ranking**
		EMOTIONLINES	***EMOTIONLINES***
Random Forest	n_estimators	50	100
	criterion	gini	gini
	max_depth	None	10
	min_samples_split	2	2
	min_samples_leaf	2	10
	max_features	2	10
	max_samples	0.1	None
AdaBoost	n_estimators	50	500
	algorithm	SAMME	SAMME
MLP/NN	hidden_layer_sizes	(100,100,100)	(100,)
	activation	identity	relu
	alpha	0.1	0.001
	learning_rate	adaptive	relu

Table A.5 - Sets of parameters used to obtain the results for the different EMsA models (tuned using grid search stratified cross-validation).

		With ranking	**Without ranking**
		RAVDESS	***RAVDESS***
Random Forest	n_estimators	50	100
	criterion	gini	Gini
	max_depth	20	10
	min_samples_split	10	10
	min_samples_leaf	2	1
	max_features	1	1
	max_samples	None	None
AdaBoost	n_estimators	500	25
	algorithm	SAMME.R	SAMME
MLP/NN	hidden_layer_sizes	(100,)	(100,100)
	activation	logistic	logistic
	alpha	0.001	0.001
	learning_rate	adaptive	invscaling