Multimodal Technologies and Interaction

MDPI

Article

Emotion Classification from Speech and Text in Videos Using a Multimodal Approach

Maria Chiara Caschera , Patrizia Grifoni * and Fernando Ferri

National Research Council, Institute of Research on Population and Social Policies (CNR-IRPPS), Via Palestro 32, 00185 Rome, Italy; mc.caschera@irpps.cnr.it (M.C.C.); fernando.ferri@irpps.cnr.it (F.F.)
* Correspondence: patrizia.grifoni@irpps.cnr.it

Abstract: Emotion classification is a research area in which there has been very intensive literature production concerning natural language processing, multimedia data, semantic knowledge discovery, social network mining, and text and multimedia data mining. This paper addresses the issue of emotion classification and proposes a method for classifying the emotions expressed in multimodal data extracted from videos. The proposed method models multimodal data as a sequence of features extracted from facial expressions, speech, gestures, and text, using a linguistic approach. Each sequence of multimodal data is correctly associated with the emotion by a method that models each emotion using a hidden Markov model. The trained model is evaluated on samples of multimodal sentences associated with seven basic emotions. The experimental results demonstrate a good classification rate for emotions.

Keywords: emotion classification; multimodal interaction; hidden Markov models

check for updates

Citation: Caschera, M.C.; Grifoni, P.; Ferri, F. Emotion Classification from Speech and Text in Videos Using a Multimodal Approach. *Multimodal Technol. Interact.* **2022**, *6*, 28. https://doi.org/10.3390/mti6040028

Academic Editor: Cristina Portalés Ricart

Received: 30 November 2021
Accepted: 7 April 2022
Published: 12 April 2022

1. Introduction

Due to the large amounts of multimedia data that are available, emotion classification has become a widely discussed topic that complements studies on sentiment analysis. Unlike sentiment analysis, which deals with evaluations (e.g., positive, negative, neutral) and aims to determine the attitude of a writer or a speaker towards some topic, or the overall contextual polarity of a document or a text, emotion classification aims to recognize the emotional state of a user (e.g., happy, angry, sad) during a conversation, and focuses on the cognitive and behavioral strategies that people use to influence their own emotional experience. The growing interest in extracting emotions from multimedia data, rather than from Natural Language (NL) data, stems from the fact that the tone of the speaker or a facial expression, for example, can improve the understanding of his or her actual emotional state.

Several works have addressed the issue of emotion classification. Most applications classify emotions into seven types: anger, disgust, fear, happiness, sadness, surprise, and neutral. An example is Ekman's classification [1], in which each emotion is considered a discrete category. Emotions have also been modelled using a wheel of eight emotions (joy, trust, fear, surprise, sadness, disgust, anger, and anticipation) by Plutchik [2], who also developed a theory in which emotions are divided into twenty-four "primary", "secondary", and "tertiary" dyads (i.e., feelings composed of two emotions) and triads (i.e., emotions formed from three primary emotions). Russell's classification model [3] divides emotions along two dimensions: valence, which ranges from unpleasant to pleasant, and arousal, which ranges from activation to deactivation. Emotions can therefore be placed in one of four quadrants: pleasant-active, which includes curiosity and interest; pleasant-inactive, which includes contentment and satisfaction; unpleasant-active, which includes emotions such as confusion and frustration; and unpleasant-inactive, which includes hopelessness and boredom. A comparison of the extant emotion models is provided in [4].

Emotions can be understood and categorized by analyzing facial expressions, vocal tones, gestures, and physiological signals. The joint use of more than one modality can offer significant advantages, as it allows for improved emotion mining and classification accuracy by helping to disambiguate linguistic meaning, introducing additional sentiment information, and enhancing the connection to real-world environments [5]. With respect to methods for extracting sentiment and opinions from NL, methods of emotion classification need to address not only the extraction of features from each modality considered (e.g., speech, handwriting, facial expressions, gestures) but also the fusion of these features (e.g., opinion words, prosody, coordinate features, distance features) [6]. In practice, multimedia content usually describes one concept in a redundant or complementary manner [7] and combines information from different modalities. One of the main sources of multimedia data on emotions is social media. For example, more than 500 h of video were uploaded to YouTube every minute in 2019. The pervasive use of social media has produced significant amounts of material on conversations, text, audio, and video; these represent an important source of data due to their huge sizes, the variety of their topics, and the dynamism of the language used. The availability of large amounts of this kind of information has sustained the growing interest in emotion classification as a very active research area that is related to semantic knowledge discovery, social network mining, and text and multimedia data mining. Emotion classification can have a deep impact on society, economy, policy, and any issue connected to opinions. It is also deeply related to computer vision applications and can be used, for example, in security, entertainment, automatic surveillance, robot motion, video indexing, and retrieval and monitoring systems.

The extraction of sentiments and emotions from text, images, audio, and videos requires the authors to model the complexity of the data coming from different channels. Hence, the main research questions addressed in this paper are as follows: How can multimodal data be modelled to extract emotions? How can the emotion classification process be modelled?

To answer these questions, the authors start with a discussion of the most relevant techniques applied for detecting sentiment and emotion in text, audio and video, and the advantages and main challenges encountered in extracting sentiments and emotions during conversations using multimodal interactions for syntactic and semantic information and modal features extracted from videos. In particular, as outlined in [8], the challenges concerning extracting sentiments and emotions involve noisy data, the presence of partial data records, difficulties in representing the complexity of human sentiments, and the ambiguity of human emotional signals.

The main contributions of this study are as follows:

- The authors represent the features of emotions related to multimodal data extracted from videos (e.g., prosodic features from audio, and postures, gestures, and expressions from video) in the form of a sequences of several types of modal information, using a linguistic approach.
- The authors formalize the emotion detection process as a multi-class classification problem and model each emotion detection process using a hidden Markov model (HMM). This model allows the authors to capture the discrete features (e.g., opinion words, prosody, facial expressions, gestures) that characterize emotions, and hence which features characterize the sequences of structured data of the multimodal sentence. The authors chose to use HMM because this model achieves good classification accuracy on multi-dimensional and discrete or categorical features.

The proposed method is used for multimodal emotion classification from facial expressions, speech prosody, gestures, and textual information. The experimental results demonstrate that our approach achieves a good classification rate for emotions.

The remaining sections of this paper are structured as follows. Section 2 sets out the motivation that encouraged the authors to perform this work. Section 3 presents an overview of methods for analyzing opinions and sentiment in text. Section 4 describes how emotions are extracted from audio features. Section 5 discusses the classification of emotions from multimodal data. Section 6 addresses the main challenges about emotion classification processes. Section 7 provides a method for detecting emotion from multimodal data, and Section 8 reports the related experimental results. Finally, Section 9 concludes the paper and suggests directions for future work.

2. Problem Definition

In this work, the authors focus on the process of classifying sentiments and emotions from conversations involving complex human sentiments and ambiguous human emotion signals.

In particular, considering the ability of human beings to be ironic, sarcastic, or liars (the authors recognize that human behavior and human communication process can be complex). This work addresses the complexity connected with the characteristics of communication and human behavior, which is also largely related to the ambiguities that characterize the communication process. Humans can transmit messages with inconsistent emotional features through different modalities [9], creating mismatches that generate ambiguities [10]. Irony, sarcasm, and, more generally, rhetorical expressions can sometimes be extracted from the tone of voice using prosodic, spectral, and contextual cues [11]; these allow the authors to reduce the number of possible interpretations, thus avoiding any ambiguity. For example, consider a user who sarcastically says, "I am so happy", with a sad facial expression. In this case, no ambiguity arises, as the sarcastic tone underlines that the meaning is different from the meaning of the spoken words.

A more complex case arises when a user is lying. The authors assume that in this case, there is no information about the tone. Although several efforts have been devoted to developing behavioral lie detection systems [12,13], the possibility of an ambiguous interpretation may persist. One example of a situation involving a lying person is during a police interrogation about the theft of some jewels. For example, during this interrogation, the thief speaks the following sentence:

"I am sad about the theft of this."

The thief has a sad tone, while he or she has a fearful facial expression and is indicating a picture of a jewel. Considering this example, the emotion expressed by the speech modality (sad) is different from the emotion indicated by the speaker's facial expression (fearful). In this case, ambiguity arises, as it is difficult for police to establish whether the person is telling the truth. Consequently, he or she may be innocent, even if he or she has a fearful facial expression.

Humans are driven by emotions that both influence and are influenced by their thoughts and actions, and human emotions can be conveyed via many types of information (e.g., speech, facial expressions, mental or physiological information). Emotion classification is particularly important in many situations for monitoring and identifying criticalities and suspicious behaviors, for example, when a high-level security alarm is raised. It is also particularly important within social networks, which are sometimes also used for terroristic and criminal purposes. Content involving criminal intent can be identified using emotion classification approaches. For example, in [14,15], the extraction of emotional features was performed to detect terroristic and aggressive behavior. Beyond security, emotion classification is important in computer vision applications used for video indexing and retrieval, robot motion, entertainment, monitoring of smart home systems [16,17], and neuro-physiological and psychological studies. For instance, emotion classification is important in monitoring the psychological and neuro-physiological condition of individuals with personality trait disorders [18,19], and to monitor and identify people with autism spectral disorders [20].

The examples cited above highlight the relevance of combining different modalities: some emotions are better recognized from audio (e.g., fear and sadness), while others are most effectively detected from video (e.g., happiness and anger) [21]. In addition, visual information can modify the perception of speech [22]. The study in [23] showed that sadness is best recognized from speech and is characterized by low pitch and energy. Busso et al. [24]

showed that in facial expression classification systems, anger can be confused with sadness, and happiness with neutral states. However, by using an acoustic method of emotion classification, anger and sadness can be identified with high accuracy, such as happiness and neutral states. In addition, a combination of different modalities may be important because, under some conditions, emotions can be not understood easily using one modality alone. For example, if a person has eyeglasses or a beard, the emotions indicated by the facial features may be detected with a high degree of error, and the limitations of the visual features can be overcome by audio features. In this vein, the authors of [25] underlined that the performance of an emotion classification system increases when two or more modalities are considered together, based onthe fact that two or more modalities often provide complementary and/or redundant information [7]. Hence, a multimodal approach may raise robustness and performance compared to a single-mode approach when the information is acquired in a noisy environment [26].

The proven need for a method of detecting emotions from multimodal information encouraged us to develop a process of identifying emotions from multimodal data, modelled using a linguistic approach. Before describing our approach, it is useful to provide an overview of existing methods used to identify sentiments and emotions extracted from textual, audio, and video data.

3. Textual Data and Emotion Classification

The extraction of sentiments from NL data from social media involves many challenges associated with the structure of micro-posts and tweets, which are often characterized by noise and implicit knowledge and are often short. Important matters that need to be addressed are parsing, the presence of sarcasm, the resolution of anaphora (i.e., what a noun phrase or pronoun refers to), and the abbreviations and poor spelling used in online social networks (e.g., Twitter, Facebook, You Tube). Information gathered from forums, content-sharing services, social networks, and blogs is one of the main data sources for retrieving opinions. This information is unstructured since it is for human consumption and cannot be processed by machines. In addition, these data may change over time. For example, reviewer data can be extracted from datasets on movie reviews and micro-blogging services, where users can express their opinions via status messages called tweets (e.g., on Twitter) that can change over time; an analysis of these allows us to evaluate a life cycle of a product and its weaknesses, and to predict the income from and profitability of an investment. To address these challenges, the predominant research fields [27] in sentiment analysis are sentiment classification, feature-based sentiment classification, and opinion summarization. Sentiment analysis has been widely investigated as a computational treatment of subjectivity, sentiment, and opinions in a text [28]. It is performed based on the polarity (i.e., positive, negative, or neutral) and intensity of the lexicon, and makes use of techniques in natural language processing (NLP), text analysis, and computational linguistics to identify and extract subjective information from the source material. Two main approaches have been used for sentiment analysis: lexicon-based (LB) techniques (i.e., corpus-based and dictionary-based approaches), which match data to a sentiment dictionary with opinion words to determine the polarity; and machine learning (ML) techniques, which apply a classification approach to classify the data extracted using NL [29].

LB techniques [30] do not need a training process for classifying data, and, during the classification process, the features of a given text involves are compared with sentiment lexicons whose sentiment values have been determined previously. LB methods mainly apply lexical relations [31], semantic similarity measures [32], and rules relating to parts of speech [33]. LB approaches also use clustering classifiers, such as exclusive clustering (e.g., the K-means clustering algorithm) [34], overlapping clustering [35], hierarchical clustering, agglomerative and divisive methods [36], and probabilistic clustering [37]. The advantage of clustering classifiers is their ability to obtain optimality measures for the classification of groups or classes. On the other hand, their main disadvantage arises from a lack of a learning dataset of labelled observations, unknown numbers of groups, and the fact that users implicitly choose the appropriate features and distance measures [38].

Relevant supervised ML approaches are decision tree classifiers [39], linear classifiers (e.g., support vector machine (SVM) [30] and neural networks (NNs) [40]), rule-based classifiers [41] and probabilistic classifiers [42] (e.g., naive Bayes (NB) [43] and Bayesian networks (BNs) [44]). Decision tree classifiers hierarchically divide data through constraints or predicates on the presence or absence of one or more words or on the similarity of documents for obtaining a set of terms that can be used to partition documents. Linear classifiers model normalized word frequencies of a document as a vector and represent this in the form of a hyperplane that separates the classes. NNs model the document word frequencies as a vector that acts as the input for the neurons (i.e., the basic units). NNs have been used to text data for predicting class labels [40]. Artificial neural networks (ANNs) [45] are extended forms of NNs that can be applied to decompose movie reviews and documents into a negative, positive, or fuzzy tone; however, they achieve high computational costs during the training process. ANNs achieve better results than SVM except in contexts with data imbalance [46]. As an advantage, SVM has low dependency on the dimensionality of the dataset, and it achieves good performance on experimental results. This approach can be efficiently applied to combine diverse information sources. The disadvantages of SVM include: the difficulty of interpreting the resulting model; to require annotated training data for the training process; the requirement for pre-processing of categorical or missing values. In addition, SVM has high computational costs at running time and it is sensitive to sparse and insufficient data. SVM is applied for the categorisation of text and movie reviews [30]. NB assumes that the features are independent. It requires labelled data for the learning process, and the trained model analyzes a text document giving the probabilities of categories by computing the word and categories join probabilities [47]. NB [43] has been used in reviews of web discourse predicting the most likely class for a new document [48]. When the feature space is larger, SVM performs better than NB, whereas, when the feature set is small, NB performs better than SVM [49]. Disadvantages of NB include the simplicity of the assumption of word independence since this assumption may not necessarily be valid [50]. Unlike NB, the assumption underlying BNs is that all the words are fully dependent. BNs have a very high computational complexity of is, and hence they are not frequently applied [44]. The accuracy of SVM is higher than the other algorithms.

A comparison of LB methods with ML techniques indicates that the former is more usable, as most domains lack a labelled training dataset, and supervised ML techniques fail when the amount of training data is insufficient [51]. In addition, providing a labelled training dataset for supervised ML techniques is very expensive. However, when trained, they give better performance than unsupervised LB methods that are conditioned by words included in the dictionary. In practice, for unsupervised LB methods, fewer words produce a decrease in performance, and the polarity of many words is domain- and context-dependent. The ML techniques are more accurate than LB techniques, but are less efficient and cannot be used in real-time.

Given these considerations, hybrid approaches have been investigated. These approaches (ML + LB) combine lexicons with learning, to achieve high accuracy from supervised learning algorithms and stability from LB approaches [52].

4. Audio Data and Emotion Classification

Unlike textual data, vocal data have multiple dimensions (e.g., maximum and minimum pitch contour, Mel-frequency, speech rate) and the vocal tokens may have several variations while representing exactly the same concept.

Those multiple dimensions imply a complex emotion classification process since the many different voice features are conditioned by the language used, the speakers, speaking style, and the type of sentences (e.g., declarative, interrogative, imperative, exclamative).

During the emotion classification process, the appropriate features are extracted from the available speech data to determine the emotions underlying the speech utterance.

The most commonly applied approaches for speech emotion classification are statistical classifiers, and these have been widely applied to many speech-based classification tasks [53]. The most typically used are HMM [54,55], the Gaussian mixture model (GMM) [56], ANNs [57] and K-nearest neighbors (KNN) [58].

HMMs have been typically used for isolated speech classification and segmentation [59] achieving a good classification accuracy [60]; for example, in [61], the authors demonstrated that the use of phoneme-based modelling allows HMM to achieve better discrimination between emotions. The use of HMM has the advantage of the physical relation between the HMM and the mechanism of production of speech signals, as HMM allows for the modelling of temporal information in the speech spectrum. The main disadvantage of HMM consists in the process of feature selection used during the building process of the classification model based on HMM.

When the number of available feature vectors is large, GMM [56] is well suited for developing a classification model of emotions. A GMM uses a multivariate Gaussian mixture density to model the probability density function of the observed data points and classifies the speech feature vectors into emotion categories considering the probability of the emotion category from the feature vectors of the specific model [62].

When relatively low numbers of training examples are available, an ANN [57] achieves better accuracy in classification than GMM and HMM. An ANN simulates the neural information processing in a human brain; this method, therefore, allows parallel processing of information, using a large number of neurons (i.e., processing elements), and uses large, interconnected networks of simple and nonlinear units [63]. ANN is efficient as a pattern classification method since it can process units and to learn system parameters for achieving local and parallel computation. The disadvantages of ANNs include long training times, complex optimization, and low robustness.

Among the ML algorithms for supervised statistical pattern classification, KNN is the simplest, and it considers that similar observations belong to similar classes of emotional states [58] assigning a target value to an item based on the distance to the nearest training case that has similar values to the predictor variables.

Classifiers for speech-based emotion analysis can be also divided into speaker-independent or speaker-dependent approaches. Unlike speaker-dependent techniques, speaker-independent systems do not need a training phase with data on users and are appropriate for many applications where it is difficult to perform training. Speaker-dependent approaches achieve better results in terms of accuracy than speaker-independent ones, as shown in [64]; however, they are not feasible for many applications that involve handling very large numbers of possible speakers. A speaker-independent approach was proposed in [65] to classify six emotions (anger, boredom, fear, happiness, sadness, neutral) using a GMM classifier, whereas in [66], emotions were classified using a Bayesian classifier and the class-conditional densities were modelled as unimodal Gaussians. Ayadi et al. [60] compared speaker-dependent and speaker-independent approaches and obtained accuracy rates of 76.12% for the speaker-dependent approaches and 64.77% for the speaker-independent ones for speech emotion classification using an HMM classifier. Finally, in noisy conditions, good performance has been obtained by combining sub-band spectral centroid weighted wavelet packet cepstral coefficients based on acoustic feature fusion with dynamic BNs for speech emotion classification [67].

5. Video Data and Emotion Classification

Alongside textual data, video data make up a large proportion of the content on social networks and contain more cues that can be used to identify emotions, since facial expressions are expressed by the visual data. In extracting emotions from facial expressions, the most important processes are human face detection and then extracting the features.

This extraction process has a wide range of applications, such as human face classification for surveillance, video conferencing, human-computer interaction, and so on.

The analysis of facial expressions allows us to understand the emotions being experienced by a human being. These are mainly divided into six basic emotions (surprise, joy, sadness, fear, disgust, and anger), and seven non-basic emotions (agreeing, curiosity, pain, fatigue, thinking, irritation, and engaged) [68].

To recognize emotions, the displacement of specific points and regions of the face (e.g., the eyes and eyebrows, the edge of the mouth, wrinkles, lips, and nasolabial furrow) are typically used [26].

The emotion recognition process has been addressed using methods based on the association between the movement of specific points or typical parts of the face and different emotional states. The FACS (Facial Action Coding System) [69] is a common standard used to categorize the physical expression of emotions. The FACS codes facial expressions as a set of facial action units (AUs), producing temporal profiles for each facial movement from a video. In particular, FACS focuses on facial parameterization, where the features are detected and encoded as a feature vector that is used to identify a particular emotion.

In [70], the authors analyzed the performance of several ML algorithms for extracting emotions from facial expressions, and they reported that HMMs outperformed KNN, ANNs, and BNs in terms of accuracy. In particular, HMMs achieved good classification accuracy on multi-dimensional, discrete, or categorical features; they, therefore, allow for dealing with the sequences of structured data of the multimodal sentences. In [71], a good level of robustness was achieved by combining five ML algorithms (RIPPER, multilayer perceptron, SVM, NB, and C4.5) for emotional classification from static facial images.

In addition, several works have been addressed to combine the visual and audio features from video data to provide emotion classification. Morency et al. [72] proved the potential of multimodal sentiment analysis by demonstrating that the joint use of multiple modalities achieved better results in terms of classification than classifiers that used only one modality at a time.

The benefits arising from a combined analysis of different features extracted from different modalities were demonstrated in [73], where a combination of the modulations in speech, textual clues, and facial expressions extracted from videos improved the identification of the level of tension from newscasts.

Facial expressions, followed closely by speech, are the best features for achieving high-precision affect recognition in emotion detection. This is because a combination of facial expressions and speech is the closest method to human–human interaction [74].

During the emotion detection process, facial features and prosodic features are combined to build joint feature vectors in a fusion process that takes place at the feature level [24]. In addition, the fusion process takes place at the decision level when audio-only and visual-only features are classified independently in terms of emotions and then combined. In [75], the authors applied fusion at the feature-level for recognizing the dimensional emotional labels of audio and visual signals rather than categorical emotional tags (e.g., anger and happiness). In addition, Poria et al. [76] applied feature-level fusion by concatenating the feature vectors of text, audio, and visual modalities to form a single, long feature vector. The fusion at the decision level was used in [77], where the audio and visual signals were analyzed in real time by applying a classifier based on dynamic BNs. In [78], speech and face data were used to model and describe the temporal dynamics of the emotion clues using HMMs. Finally, facial expressions, speech, and physiological signal features were combined to recognize emotions using KNNs [79].

In [24], Busso et al. compared these two fusion methods demonstrating that the emotion classification of data fused at the feature level achieves a similar precision concerning the emotion classification of data fused at the decision level. The best choice between these fusion techniques derives from the particular application. In addition, the decision-level fusion assumes that the modalities are conditional independent; this implies a loss of information on the mutual correlation between two modalities since information belonging to different modalities is often displayed by humans in a redundant and/or complementary manner.

The model-level fusion was proposed to overcome this limitation. This fusion combines feature-level and decision-level fusion in a hybrid manner. This type of fusion considers the correlation between information belonging to different modalities by modelling the modalities' correlation properties and relaxing the requirement for synchronization of this information [80]. This type of fusion was provided in [81] by using a multistream-fused HMM building order to model optimal connections among pitch and energy features from audio and facial features from video, based on maximum mutual information and the maximum entropy criterion. In [82], The hybrid fusion of body orientation, facial contours, lexical content of speech, and prosody was applied using NNs for recognizing emotions from facial expressions and speech. In [83], a hybrid fusion method that modelled the user's emotions over time was applied. The authors used recurrent NNs to interpret emotional transitions sequences of events from vocal, facial, and body expressions. A multimodal regression model was applied to infer emotions from short pieces of text and images on Twitter [84]. In [85], a framework for emotion recognition was presented based on the fusion of visible and infrared images with speech. The authors applied feature-level extraction to the image features using SVM and the speech features using CNNs and used decision-level fusion to combine the image and speech features.

Emotion extraction from videos has also been investigated by applying deep learning. In [86], CNNs were applied to extract emotions from videos and audio streams simultaneously. Emotions have also been extracted from videos by applying a convolutional deep belief network, which achieved better recognition accuracies than the SVM baselines in multimodal scenarios [74]. In [87], a new deep learning method was proposed for emotion classification in music videos by analyzing information from audio, video, and facial expressions. In [88,89], the authors investigated a combination of CNNs and RNNs and demonstrated that this approach performed better than other state-of-the-art methods for emotion recognition using popular emotion corpora. They also showed that these models generated robust multimodal features for emotion classification in an unsupervised manner. An overview of multimodal emotion recognition using deep learning is presented in [90].

6. Open Challenges

Studies of emotion detection have addressed several challenges arising from the fact that people often combine different emotions in the same sentence, which is easily understood by humans but difficult to parse with computers. In practice, humans can convey inconsistent emotional features in their messages through different modalities [9], creating mismatches that generate ambiguities. Irony, sarcasm, and rhetorical figures can sometimes be extracted from the tone of the voice using prosodic, spectral, and contextual cues [11]; these allow the number of possible interpretations to be reduced, thus avoiding any ambiguity.

A further challenge is connected with the automatic extraction of emotional information from a variety of data provided by different interaction modalities and from different domains. Salido Ortega et al. [91] addressed the classification of emotions associated with a particular context in which emotions are actually experienced, by using ML techniques to build models from contextual information. Their study involved young adults who were pursuing an engineering degree. However, they stated that the recognition of emotions from contextual information for individuals with different profiles and of different age groups remains an open challenge. Perifanos and Goutsos [92] combined text and image modalities to detect hate speech in social networks. A combination of text and speech was used since posters often use messages encoded in images to avoid NLP-based hate speech detection systems [92].

Since emotions are not only expressed in the form of text but also via images, audio, and video, this means that not only unstructured data but also data in several other forms are involved. More automatic techniques are required to extract sentiments and emotions from these data.

The works analyzed above indicate that the problem of multimodal emotion classification has mainly been addressed using ML techniques (e.g., BNs, HMMs, and NNs), which require a training phase.

The problem of emotion classification from multimodal data mainly derives from the management of multiple and combined forms of data (i.e., audio, video and text), and the fact that emotional expressiveness varies from one person to another. This poses challenges in terms of the differences in person-to-person communication patterns, as some people express themselves more visually and others more vocally. ML techniques, therefore, pose challenges associated with the need for training on heterogeneous features (i.e., the intensity of lexicons and polarity from text, prosodic features from audio, and postures, expressions, and gestures from video and connected noise).

Further challenges related to emotion classification derive from the large amounts of noisy data, the presence of partial data, and the difficulty of representing the complexity of human emotions. Another challenge involves the ambiguity of human emotional signals. An ambiguity in the emotion classification process can arise when different emotional states (e.g., lowered eyebrows that may indicate anger or concentration) are identified by the analysis of similar configurations of features (e.g., vocal and facial features) [82]. Moreover, when people shout, this may signify anger or may simply be necessary for communicating in a noisy environment.

As introduced in Section 2, a complex case is when a person is lying. Although several efforts have been made to develop behavioral lie detection systems [12,13], the possibility of an ambiguous interpretation may persist. An example of a situation involving a lying person is a police interrogation about a jewel theft, as discussed in Section 2.

7. Multimodal Emotion Classification

As described in the previous sections, data extracted from text, audio, and video are characterized by different features with different metrics, dynamic structures, and time scales, and, therefore, are heterogeneous. However, these heterogeneous data are connected by semantic relations. For this reason, video, audio, and text need to be transformed into features taking into consideration the differences between these data and the semantic and temporal relations among them.

To address this issue, the authors propose a language-based approach for managing not only the heterogeneity of these data but also their relationships. In particular, the authors use a linguistic method that is able to formalize different modal information and their correlations in a combined structure [93].

In this section, the authors describe the multimodal features combined in our approach to classifying the emotions in interaction processes shown in the videos.

7.1. Dataset Construction

When building the dataset, the authors took into consideration the fact that some emotions are better identified from audio (e.g., sadness and fear) while others are best detected from video (e.g., anger and happiness) [21]. A further hypothesis was that, in emotion classification from audio, anger and sadness can be distinguished with high accuracy, as happiness and neutral states, while in facial expression classification, anger may be mistaken with sadness, and happiness with a neutral state [24]. Hence, for each type of emotion, there is an optimum modality for expressing it.

As stated in [94], emotional databases can be categorized in spontaneous, invoked, and acted or simulated emotions. This categorization considers a data source: spontaneous emotions refer to data obtained by recording in an undisturbed environment (e.g., talk shows, reality shows, or various types of live coverage); invoked emotions refer to data obtained by recording an emotional reaction provoked by staged situations or aids (e.g., videos, images, or computer simulations); acted or simulated emotions refer to acted-out emotional samples. A spontaneous dataset might be composed of material that is questionable due to background noise, artifacts, and overlapping voice. Instead, the main disadvantage of the invoked dataset is the lack of results' repeatability, as people might differently react to the same stimuli. For this study, the authors needed high-quality recordings, with clear undisturbed emotional expression. Therefore, the authors decided to build an acted or simulated dataset composed of data extracted from recording unqualified volunteers who acted out emotional samples. Six people were involved in building the dataset, with post-graduated and working ages ranging from 27 to 60. During the data acquisition process, these people were placed in front of a camera with their faces visible, and background music and sound were not present. Each participant was asked to express ten different multimodal sentences connected to each emotion (six people × ten multimodal sentences × seven emotions).

To annotate the videos, the authors used the method proposed in [95], in which syntactic and semantic information was extracted from multimodal dialogues. The elements of the unimodal sentences and their properties (i.e., actual value, syntactic role, modality, and kinds of cooperation between modalities [7]) were extracted and combined to generate linear sequences of elements, called multimodal linearized sentences. Grammatical rules were then applied to these multimodal sentences to parse them, and the multimodal sentences were annotated with the correct interpretation and the correct syntactic roles of the element of the sentence.

In this way, the annotated multimodal sentences were associated with the relevant emotions extracted from the features of facial expressions, emotional speech, and text. To extract these emotions from video sequences and texts, the authors used existing tools. To extract emotion from facial expressions, the authors used CrowdEmotion [96], which provides an online demo [97] for analyzing facial points from the real-time video, and yields a time series for the six universal emotions (happiness, surprise, anger, disgust, fear and sadness) as defined in [69]. Emotion was extracted from texts using the TheySayPreCeive REST API Demo [98], an engine created by TheySay [99] for monitoring, understanding, and measuring opinions and emotions expressed in text. This tool applies grammatical and semantic analysis using a proprietary parser, in which meaning is identified from text including sentiment, intent, and other subjective dimensions across multiple levels, including documents, sentences, entities, topics, and relations. Finally, the authors extracted the pitch, energy, and other features from emotional speech using openSMILE [100], which extracts the features of emotional speech, music, and sounds by recognizing the prosodic features and the audio component of the spectrum in the sound modulations of audio signals from videos and audio streams.

The data sample distribution per emotion in the final dataset is shown in Table 1.

Table 1. Data samples distribution per emotion in the dataset.

Emotions	Number of Samples
Anger	60
Disgust	60
Fear	60
Happiness	60
Neutral	60
Sadness	60
Surprise	60

In total, the resulting dataset contained 420 multimodal sentences corresponding to seven emotions (60 for each of the emotions of anger, disgust, fear, happiness, neutral, sadness, and surprise). The training and testing sets consisted of 60% and 40% of the dataset, respectively.

7.2. Representation

Using this language-based method, the authors converted the information extracted from the videos into a sequence of sentences consisting of several modes of information. Each sentence was composed of a sequence of elements (a string of symbols), where each element was associated with the audio or visual modality, or another modality used in the video. The features extracted from each modality were syntactic (e.g., syntactic role) and semantic (e.g., concept, representation). To extract the syntactic and semantic features, the authors used the method put forward in [95]. Sentences were segmented into elements to extract information about the modality used and the representation in that modality. Following this, each element was localized in a list of elements to associate it with the temporal interval. In addition, the syntactic features were extracted using the Stanford Parser [101], and the semantic features were extracted based on the conceptual structure of the context [93]. The emotions and their features (e.g., facial points in the real-time video; sentiment, intent, and opinion words in the text; the audio component of the spectrum in sound modulations; prosody in audio signals), were extracted using specific tools for each modality as described in Section 7.1.

The proposed method is based on the representation of each multimodal sentence MS_i as a sequence of elements $\{E^i\}_{i=1}^{n}$, where each E^i represents a meaningful feature of the language used. The multimodal sentence MS_i [102] is a fundamental concept and forms a grammatical unit composed of a set of terminal elements E^i that are the elementary parts of a multimodal language. Each terminal element E^i can be characterised by meaningful features. These meaningful features are modelled to represent the modality used for representing the element, the representation of the element in the used modality, the temporal features of the element, the syntax extracted by the Stanford Parser [101] (e.g., noun, verb, adjective, adverb, pronoun, preposition, etc.), the semantic meaning of the element considering the conceptual structure of the context [93], and the cooperation between modalities [7] when more than one modality is used to define the sentence (e.g., complementarity and redundancy) [7]. Note that the feature involving cooperation is applied to the elements of the multimodal sentence that are in a close-by relationship [7]; otherwise, this feature has the value "noncooperation".

As defined in [102], each terminal element E^i is identified by a set of meaningful features, as follows: ${E^i}_{mod}$ corresponds to the modality (e.g., speech, facial expression, gesture) used to create the element E^i; ${E^i}_{repr}$ indicates how the element E^i is represented by the modality; ${E^i}_{time}$ measures the time interval (based on the start and end time values) over which the element E^i was created; ${E^i}_{role}$ corresponds to the syntactic role that the element E^i plays in the multimodal sentence, according to the Penn Treebank Tag set [103] (e.g., noun, verb, adjective, adverb, pronoun, preposition, etc.); and ${E^i}_{concept}$ gives the semantic meaning of the element considering the conceptual structure of the context [104]. Given two elements E^i and E^j, where E^j has a close-by relationship with E^j [7], ${E^i}_{coop}$ is set to the same value as E^j and specifies the type of cooperation [7] between the elements E^i and E^j. Finally, the field representation (${E^i}_{repr}$) includes the features (e.g., characteristic points from a face image, prosody from speech, opinion word from text) conveyed by the modality (e.g., speech, text, facial expression, gesture) used to create the element E^i.

When detecting emotions, the authors need to consider features that allow for discrimination between different emotions. These features are unique and differ depending on the modality used to convey emotion. For example, features of speech that allow us to discriminate between emotions include the pitch, rate, tone, and articulation, while the main visual features include the facial features of the eyes, nose, and mouth. Murray and Arnott [105] presented a review that summarizes the relationships between speech features and emotional states and showed that sadness corresponded to a slightly lower average pitch, lower intensity, and resonant voice quality.

Since there are features that allow us to identify each emotion depending on the modality used, the authors need to enrich the formalism defined in [102] by introducing into the representation field $E^i{}_{repr}$, the particular features of the emotion with respect to the modality.

For the sake of clarity, the authors consider the example of deception introduced in Section 2, where a person speaks the words:

"I am sad about the theft of this."

while showing a fearful facial expression and gesturing towards a picture of a jewel, as illustrated in the timeline in Figure 1. The authors start from the hypothesis that all the multimodal elements described above are extracted using tools for gesture classification, facial expression, and handwriting classification. Our method is applied downstream of these tools. All the elements defined through the interaction modalities (in this example, the authors have speech, facial expressions, and gestures) are combined in the multimodal sentence. The authors can use the definitions of complementarity and redundancy [7] to show that the speech element "🔊 sad" and the facial expression element "😨" (fearful) are redundant, whereas the speech element "this" and the gesture element indicate that the jewel in the picture are complementary.

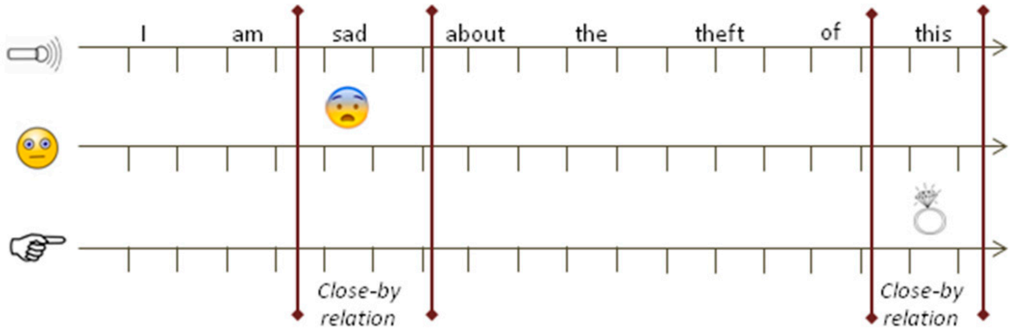


Figure 1. Example of an ambiguous multimodal sentence.

Information about the interaction modalities, the temporal intervals, and the element representations is extracted during the process of interaction between the user and the system, while the concepts are extracted using an ontology defined according to the context of the interaction. An explanation of how this information is extracted is beyond the scope of this paper.

The syntactic roles and dependencies between the elements of the multimodal sentence are extracted using the Stanford Parser [101], which parses the natural language sentence associated with the multimodal sentence by applying a linearization process [93]. Knowledge of the syntactic roles and dependencies allows us to build a syntax graph to represent the syntactic structure of the sentence (Figure 2).

Note that the emotion conveyed by the multimodal sentence shown in Figure 1 is not univocally defined, since although the speech element "🔊 sad" and the facial expression element "😨" have a close relationship [7] and are combined into the same syntactic role (see Figure 2), they express two different emotions (sadness and fear). To unambiguously identify the emotion conveyed, the two modal elements need to refer to the same emotion. In this case, there are two possible interpretations of the multimodal sentence ("I am sad about the theft of this jewel" and "I am fearful about the theft of this jewel").

The multimodal sentence in this example is modelled as a sequence of 10 elements, of which 8involve speech, one relates to facial expression, and one to a gesture, as shown in Figure 3. Each element E^i (for $i = 1, \ldots, 10$) is characterized by the set of features: $E^i{}_{mod}$, $E^i{}_{repr}$, $E^i{}_{time}$, $E^i{}_{role}$, $E^i{}_{concept}$, and $E^i{}_{coop}$. The sequences of features that characterize these elements are composed in a feature's vector f^t at time t:

$$f^t = [(E^i{}_k)]^t \text{ with } i = 1 \ldots 10 \text{ and } k \,\epsilon\, K \text{ and } K = \{mod, repr, time, role, concept, coop\} \quad (1)$$

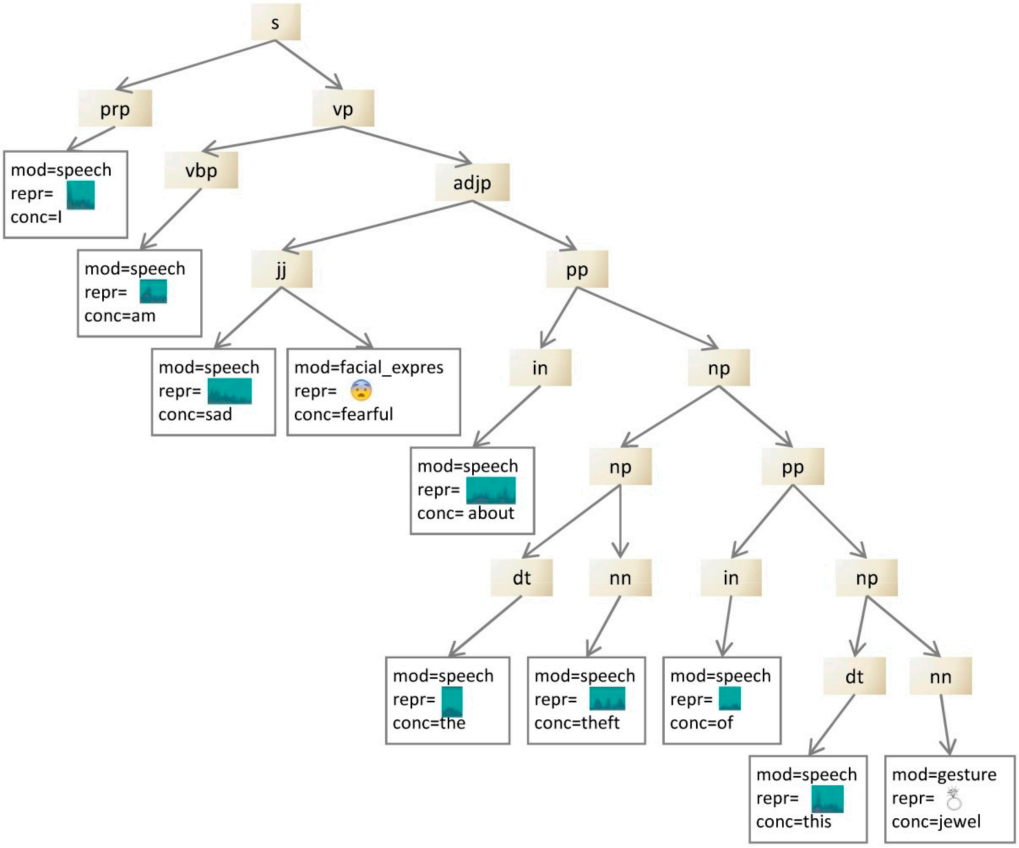
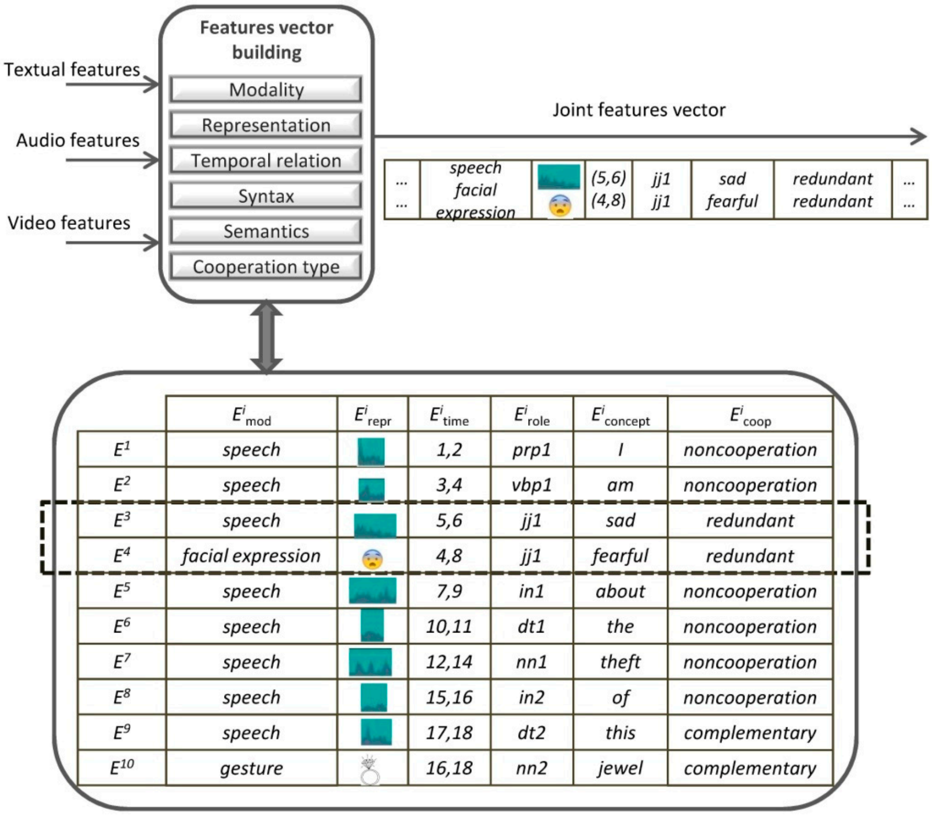


Figure 2. Syntax-graph of the multimodal sentence in Figure 1.



...	speech		(5,6)	jj1	sad	redundant	...
...	facial expression	😨	(4,8)	jj1	fearful	redundant	...

	E^i_{mod}	E^i_{repr}	E^i_{time}	E^i_{role}	$E^i_{concept}$	E^i_{coop}
E^1	speech		1,2	prp1	I	noncooperation
E^2	speech		3,4	vbp1	am	noncooperation
E^3	speech		5,6	jj1	sad	redundant
E^4	facial expression	😨	4,8	jj1	fearful	redundant
E^5	speech		7,9	in1	about	noncooperation
E^6	speech		10,11	dt1	the	noncooperation
E^7	speech		12,14	nn1	theft	noncooperation
E^8	speech		15,16	in2	of	noncooperation
E^9	speech		17,18	dt2	this	complementary
E^{10}	gesture	💍	16,18	nn2	jewel	complementary

Figure 3. Process of building the joint features vector.

To construct the joint features vector, the multimodal elements are arranged to support the correct classification of emotions. The features are extracted from each modality and modelled as the elements of the multimodal sentence and are then combined into a joint feature vector by the *Features vector building* module, as illustrated in Figure 3.

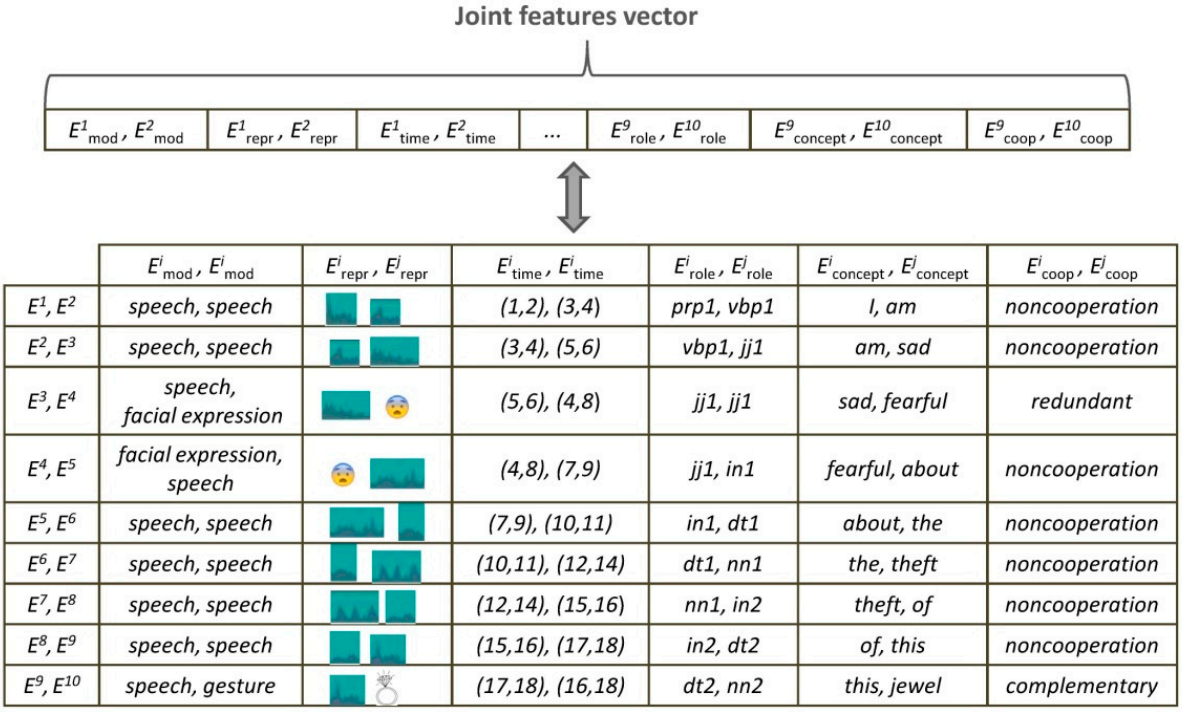
The joint features vector contains the pairs of elements that make up the multimodal sentence. The pairs of elements contained in the features vector f_t of the consecutive elements making up the multimodal sentence($E^{i}{}_{mod}E^{i+1}{}_{mod}$, $E^{i}{}_{repr}E^{i+1}{}_{repr}$, $E^{i}{}_{repr}E^{i+1}{}_{repr}$, $E^{i}{}_{time}E^{i+1}{}_{time}$, $E^{i}{}_{role}E^{i+1}{}_{role}$, $E^{i}{}_{concept}E^{i+1}{}_{concept}$, $E^{i}{}_{coop}E^{i+1}{}_{coop}$) are concatenated to form a single, long feature vector v^t, represented by the joint features vector:

$$v^t = [(\,E^{i}{}_{k},\, E^{i+1}{}_{k})]^t \text{ with } i = 1, 10 \text{ and } k \in K \text{ and } K = \{mod, repr, time, role, concept, coop\} \quad (2)$$

As shown in Figure 3, the joint features vector for the multimodal sentence illustrated in Figure 1 is:

$$v^t = [(speech, speech), (\;,\;), (1\text{–}2, 3\text{–}4), (prp1, vbp1), (I, am), (noncooperation, noncooperation) \ldots (speech, gesture), (\;, 💍), (17\text{–}18, 16\text{–}18), (dt2, nn2), (this, jewel), (complementary, complementary)] \quad (3)$$

Figure 4 shows the pairs of features corresponding to the multimodal elements making up the joint feature vector, which were extracted from the video of the theft interrogation described above.



Joint features vector
$E^1{}_{mod}$, $E^2{}_{mod}$
$E^1{}_{repr}$, $E^2{}_{repr}$
$E^1{}_{time}$, $E^2{}_{time}$
…
$E^9{}_{role}$, $E^{10}{}_{role}$
$E^9{}_{concept}$, $E^{10}{}_{concept}$
$E^9{}_{coop}$, $E^{10}{}_{coop}$

	$E^i{}_{mod}$, $E^j{}_{mod}$	$E^i{}_{repr}$, $E^j{}_{repr}$	$E^i{}_{time}$, $E^j{}_{time}$	$E^i{}_{role}$, $E^j{}_{role}$	$E^i{}_{concept}$, $E^j{}_{concept}$	$E^i{}_{coop}$, $E^j{}_{coop}$
E^1, E^2	speech, speech		(1,2), (3,4)	prp1, vbp1	I, am	noncooperation
E^2, E^3	speech, speech		(3,4), (5,6)	vbp1, jj1	am, sad	noncooperation
E^3, E^4	speech, facial expression	😨	(5,6), (4,8)	jj1, jj1	sad, fearful	redundant
E^4, E^5	facial expression, speech	😨	(4,8), (7,9)	jj1, in1	fearful, about	noncooperation
E^5, E^6	speech, speech		(7,9), (10,11)	in1, dt1	about, the	noncooperation
E^6, E^7	speech, speech		(10,11), (12,14)	dt1, nn1	the, theft	noncooperation
E^7, E^8	speech, speech		(12,14), (15,16)	nn1, in2	theft, of	noncooperation
E^8, E^9	speech, speech		(15,16), (17,18)	in2, dt2	of, this	noncooperation
E^9, E^{10}	speech, gesture	💍	(17,18), (16,18)	dt2, nn2	this, jewel	complementary

Figure 4. Example showing the pairs of elements of the joint features vector.

When the modal features have been modelled as the joint features vector, emotion classification essentially becomes a classification process. Based on the findings of the previously analyzed studies and the works in [10,106–108], HMMs appear to be appropriate to extract information from multimodal data, and hence for extracting emotions from multimodal data formalized using a multimodal language. An HMM allows us to classify the language sequence data since this approach can be applied in an analogous way to classify text sequence data and proteins [109].

The concepts and notions described above are used in the following sections in a discussion of our classification method for emotions based on HMMs.

7.3. Proposed Model

According to the theory introduced in [1], which asserted that any emotion can be considered as a composition of primary emotions (i.e., happiness, surprise, fear, anger, neutral, disgust and sadness), there are only seven basic emotions.

In the proposed model, each of these basic emotions is modelled by a specific HMM, based on a parametric model that is used to statistically describe a time series under the Markov assumption. Each of the seven HMMs can be visualized as a spatio-temporal model describing a multimodal sentence in the form of a chain of joint features vectors in which high-level features are combined from multiple channels into a single vector. The authors apply a strict left-to-right model in which each state is transferred to the next. The authors use HMMs as they can represent the differences in the whole structure of multimodal sentences, manage multimodal features, and incorporate temporal frequent pattern analysis for baseball event classification, as set out in [108]. This method was also selected due to its proven effectiveness in the extraction and classification process [107].

An HMM models the joint probability of a time series X_i as a chain of observations x_i^t and corresponding discrete (unobserved) hidden states z_i^t.

Each HMM is made up of five components: the hidden states, the observation symbols for each state, the probability distribution for state transitions, the probability distribution for the observation symbols in each state, and the probability distribution of initial states [59].

Each hidden state in an HMM is essentially an abstraction of a joint feature vector and describes a pair of high-level features that make up the elements of the multimodal sentence (observations). The joint feature vectors are used as observation sequences for the HMMs for different emotions. The observation sequence for a multimodal sentence consisting of m elements E^i is:

$$\begin{gathered} x_i{}^t \equiv v^t = [(E^i{}_k, E^{i+1}{}_k)]^t \\ \text{for } i = 1, \ldots, m \text{ and } k \,\epsilon\, K \text{ and } K = \{mod, repr, time, role, concept, coop\} \end{gathered} \tag{4}$$

The probability value $p(x_t/\lambda_k)$ is computed between each emotion model λ_k and the analyzed joint features vector (x_t), to return the most probable emotion (associated with λ_i) associated with the analyzed joint features vector.

Figure 5 illustrates the proposed model for emotion classification. The authors extract the multimodal features from videos, which may contain images, audio signals, and text. The data from each of these modalities were processed. The visual feature extraction module extracts the facial expression features from the frames using CrowdEmotion [96], while the audio feature extraction module extracts emotional speech features from the audio stream using openSMILE [100], and the textual features extraction module extracts the features from text using TheySayPreCeive REST API Demo [98], as described in Section 7.1.

The extracted modal information is then used by the Feature vector building module to build the joint feature vector (see Figure 5). This step models the relevant information conveyed by the different modalities and their correlations using a linearized language-based representation that provides the syntax and semantics for the multimodal sentences, as described in Section 7.2.

To classify the emotions, the authors train one HMM for each emotion class using the joint features vector associated with the specific emotion, as illustrated in the Training phase in Figure 5. At the test stage, the authors identify the most probable class, as shown in the Testing phase in Figure 5. For this reason, the proposed emotion classification model consists of two phases: the training phase and the testing phase.

Since our aim is to identify emotions from multimodal features, the authors considered the sets of joint feature vectors with their assigned categories ('neutral' or one of the six emotions) as our training dataset for the observation sequences. The training phase allows the key parameters to be captured (i.e., the concept or the syntactic role) and the correct emotion to be associated with the multimodal sentence. The training data from a multimodal sentence dataset D, with N samples, is represented as $D = \{MS_i, EM_i\}_{i=1}^{N}$, where MS_i is a multimodal sentence, $EM_i \in$ EM is its emotion label, and EM = {anger, disgust, fear, happiness, neutral, sadness, and surprise}. The model was trained on 252 multimodal sentences and 7emotions (giving 42 multimodal sentences for each emotion).

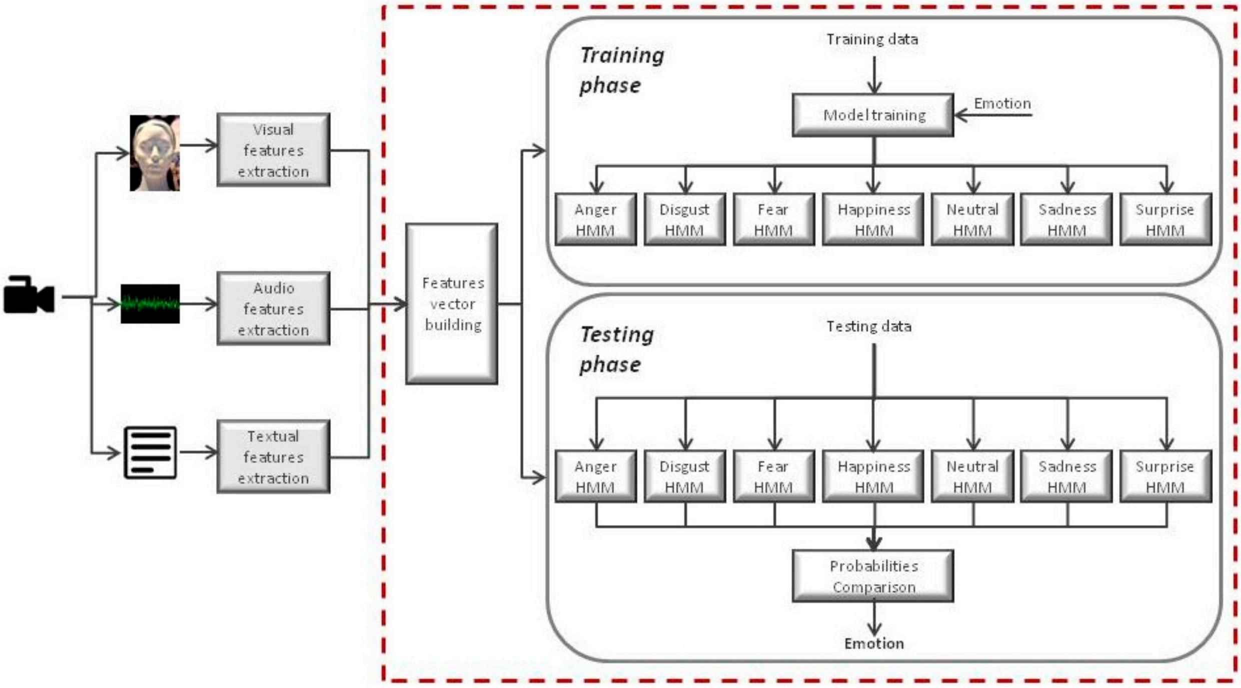


Figure 5. Overview of the proposed emotions classification model.

Each of the emotion models (the anger, disgust, fear, happiness, neutral, sadness, and surprise HMMs) was trained on the joint feature vectors associated with the related emotion. The sequence of joint feature vectors is the observation sequence of our model that contain multimodal features. The association between any pair of joint feature vectors and emotions is modelled in the hidden states.

The training set consisted of 60% of the full dataset and contained 420 multimodal sentences representing emotions. Table 2 shows the number of samples used for the emotion model.

Table 2. Numbers of emotion samples used in the training phase.

Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
36	36	36	36	36	36	36

In the testing phase, the joint feature vectors extracted from the test data were passed as input to all the trained models to identify the emotion in the multimodal sentence. For the testing set, the authors used 40% of the full dataset, consisting of 420 multimodal sentences containing emotions, as shown in Table 3.

Table 3. Numbers of emotion samples used for the HMMs in the testing phase.

Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
24	24	24	24	24	24	24

As described in Figure 5, the Probability comparison module picks out the HMM with the highest probability value. In this way, the videos, audio streams, and/or text are classified using the appropriate model.

For example, if the happiness model is the one that gives the highest probability value for a particular multimodal sample, the Probabilities Comparison module determines that the emotion predicted for this multimodal sample is happiness.

8. Evaluation

After training, the model was tested. An evaluation process was performed on a testing dataset containing 168 multimodal sentences and 7emotions (with 24 multimodal sentences for each emotion), as described in Section 7.1.

The confusion matrix for this case shows the number of multimodal sentences in the testing dataset, classified based on the true emotion label and predicted by the model as anger, disgust, fear, happiness, neutral, sadness, and surprise.

Table 4 shows the confusion matrix generated for the emotion classification model based on multimodal sentences. The rows represent the number of predicted classifications made by the model for anger (ang), disgust (dis), fear (fea), happiness (hap), neutral (neu), sadness (sad), and surprise (sur). The columns represent the true classifications of the test data.

Table 4. Numbers of emotion samples used in the training phase.

		Predicted						
		Anger	**Disgust**	**Fear**	**Happiness**	**Neutral**	**Sadness**	**Surprise**
True	**Anger**	0.58	0.17	0.08	0.00	0.04	0.13	0.00
	Disgust	0.08	0.46	0.17	0.00	0.13	0.08	0.08
	Fear	0.08	0.21	0.54	0.00	0.04	0.08	0.04
	Happiness	0.00	0.00	0.00	0.71	0.13	0.00	0.17
	Neutral	0.17	0.04	0.04	0.13	0.46	0.13	0.04
	Sadness	0.04	0.08	0.08	0.04	0.13	0.54	0.08
	Surprise	0.04	0.04	0.08	0.13	0.08	0.04	0.58

In the testing phase, the authors used three performance evaluation measures for each of the trained emotion models, as follows:

- Precision (P_i): This gives a score for each emotion class and is defined as the ratio of the multimodal sentences that are correctly classified by the model as belonging to the given emotion class to the total number of multimodal sentences classified by the model as belonging to the given emotion class.
- Recall (R_i): This gives a score for a particular class and is defined as the ratio of the number of multimodal sentences correctly classified by the model as belonging to the given emotion class to the total number of multimodal sentences actually belonging to the given emotion class.
- Specificity (S_i): This measures the proportion of no true emotion classes that are correctly identified as false.

Precision, recall, and specificity are all important measures of relevance for the classification model. High precision means that the model returns more relevant instances than irrelevant ones, while high recall means that the model returns most of the relevant instances. Specificity quantifies the avoidance of no true classes that are classified as true, and hence high specificity means a low type I error rate [110].

For an HMM$_\mathrm{i}$ trained to classify emotions i (where i = *ang*, *dis*, *fea*, *hap*, *neu*, *sad*, *sur*), these measures are defined as follows [111]:

$$P_i = \frac{\sum_{j=ang}^{sur} x_{jj}}{\sum_{j=ang}^{sur} x_{jj} + \sum_{\substack{j\,=\,an \\ j \neq i}}^{sur} x_{ji}} \tag{5}$$

$$R_i = \frac{\sum_{j=ang}^{sur} x_{jj}}{\sum_{j=ang}^{sur} x_{jj} + \sum_{\substack{j = ang \\ j \neq i}}^{sur} x_{ij}} \tag{6}$$

$$S_i = \frac{\sum_{i=ang}^{sur} \sum_{\substack{j = ang \\ j \neq i}}^{sur} \sum_{\substack{k = ang \\ k \neq i}}^{sur} x_{jk}}{\sum_{i=ang}^{sur} \sum_{\substack{j = ang \\ j \neq i}}^{sur} \sum_{\substack{k = ang \\ k \neq i}}^{sur} x_{jk} + \sum_{\substack{j = an \\ j \neq i}}^{su} x_{ji}} \tag{7}$$

for j = *ang*, *dis*, *fea*, *hap*, *neu*, *sad*, *sur* and k = *ang*, *dis*, *fea*, *hap*, *neu*, *sad*, *sur*.

Table 4 presents a summary of the experiments and gives the normalized multi-class confusion matrix for the emotion classification model when applied to the 168 multimodal sentences associated with seven emotions (24 multimodal sentences for each emotion). The main source of confusion was between fear and disgust, due to the similarity between the multimodal features that characterized these emotions.

Figure 6 displays the results of the evaluation parameters for all the emotion models and the values for the specificity, recall, and precision, thus enabling a comparative analysis of all the different models (the anger, disgust, fear, happiness, neutral, sadness, and surprise HMMs).

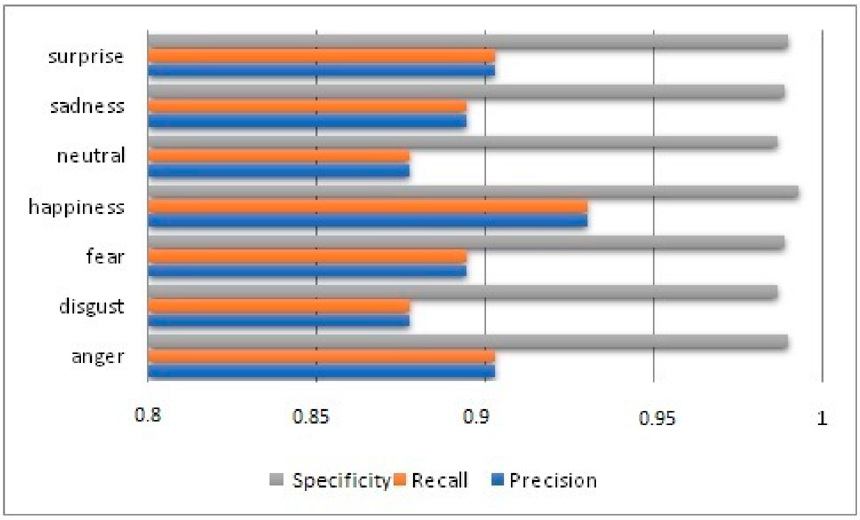


Figure 6. Emotion classification rates.

Figure 6 shows that the best emotion classification result was achieved for happiness, as the happiness HMM yielded the highest values for the three performance evaluation measures.

9. Conclusions and Future Work

In this paper, the authors have presented a method for classifying emotions from multimodal data extracted from video in the form of sequences of different modal information using a linguistic approach. Our method classifies emotions by analyzing multimodal sentences consisting of features extracted from facial expressions, speech prosody, gestures, and textual information.

The authors modelled each emotion as an HMM, and these were trained and tested on samples of multimodal sentences containing seven basic emotions (anger, disgust, fear, happiness, neutral, sadness, and surprise). The experimental results showed good emotion classification rates, and the best results were achieved for happiness.

In terms of the influence of each modality on the process of emotion identification, the results from our model support those of other studies in the literature. In particular, the correct identification of sadness and fear is influenced by the use of speech in multimodal sentences, while the correct identification of anger and happiness is influenced by the use of facial expressions and gestures. Therefore, the combination of different modalities allows for improving the correct identification of emotions.

The authors have developed a process of emotion classification based on linguistic features and other features that characterize the interaction modalities (e.g., speech, facial expressions, and gestures). In future work, the authors will investigate how contextual, cultural-related, and gender-related features can improve the performance of the proposed method. The need for this research was suggested in studies conducted by Mesquita et al. [112], who demonstrated that culture shapes and constitutes individuals' emotions and presented evidence that gender causes differences concerning emotions [113].

In future work, the authors will overcome the limitations of this work arising from the relatively small sizes of the samples used and will include more participants and collect further data in an attempt to improve the classification rate. In addition, the authors will try to understand the impact of age and gender differences on emotional data.

Author Contributions: Conceptualization, M.C.C., P.G. and F.F.; Data curation, M.C.C.; Formal analysis, M.C.C.; Investigation, M.C.C., P.G. and F.F.; Methodology, M.C.C., P.G. and F.F.; Supervision, F.F.; Validation, M.C.C.; Writing—original draft, M.C.C., P.G. and F.F.; Writing—review & editing, P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

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

References

1. Ekman, P. Basic emotions. In *Handbook of Cognition and Emotion*; Dalgleish, T., Power, T., Eds.; John Wiley & Sons: New York, NY, USA, 1999.
2. Plutchik, R. The Nature of Emotions; American Scientist Vol. 89, No. 4 (JULY-AUGUST 2001); Sigma Xi, The Scientific Research Honor Society. 2001, pp. 344–350. Available online: https://www.jstor.org/stable/27857503 (accessed on 22 February 2017).
3. Russell, J.A. Core affect and the psychological construction of emotion. *Psychol. Rev.* **2003**, *110*, 145. [CrossRef] [PubMed]
4. Rubin, D.C.; Talarico, J.M. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory* **2009**, *17*, 802–808. [CrossRef] [PubMed]
5. Chen, D.; Mooney, R. Panning for gold: Finding relevant semantic content for grounded language learning. In Proceedings of the Symposium Machine Learning in Speech and Language Processing, Bellevue, WA, USA, 27 June 2011; Available online: www.cs.utexas.edu/~{}ml/papers/chen.mlslp11.pdf (accessed on 18 February 2016).
6. Paleari, M.; Chellali, R.; Huet, B. Features for multimodal emotion recognition: An extensive study. In Proceedings of the 2010 IEEE Conference on Cybernetics and Intelligent Systems (CIS), Berks, UK, 1–2 September 2010; pp. 90–95. [CrossRef]
7. Caschera, M.C.; Ferri, F.; Grifoni, P. Multimodal interaction systems: Information and time features. *Int. J. Web Grid Serv.* **2007**, *3*, 82–99. [CrossRef]
8. Caschera, M.C.; Ferri, F.; Grifoni, P. Sentiment analysis from textual to multimodal features in digital environments. In Proceedings of the 8th International Conference on Management of Digital EcoSystems (MEDES), ACM, New York, NY, USA, 1–4 November 2016; pp. 137–144. [CrossRef]
9. Lee, S.; Narayanan, S. Audio-visual emotion recognition using Gaussian mixture models for face and voice. In Proceedings of the IEEE International Symposium on Multimedia, Berkeley, CA, USA, 15–17 December 2008.
10. Caschera, M.C. Interpretation methods and ambiguity management in multimodal systems. In *Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services: Evolutionary Techniques for Improving Accessibility*; Grifoni, P., Ed.; IGI Global: Hershey, PA, USA, 2009; pp. 87–102.
11. Tepperman, J.; Traum, D.; Narayanan, S. Yeah right: Sarcasm recognition for spoken dialogue systems. In Proceedings of the InterSpeech-ICSLP, Pittsburgh, PA, USA, 17–21 September 2006.
12. Frank, M.G.; O'Sullivan, M.; Menasco, M.A. Human behavior and deception detection. In *Handbook of Science and Technology for Homeland Security*; Voeller, J.G., Ed.; John Wiley & Sons: New York, NY, USA, 2009.
13. Abouelenien, M.; Perez-Rosas, V.; Mihalcea, R.; Burzo, M. Deception detection using a multimodal approach. In Proceedings of the 16th ACM International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, 12–16 November 2014.
14. Ma, M.D. Methods of detecting potential terrorists at airports. In *Security Dimensions and Socio-Legal Studies*; CEEOL: Frankfurt am Main, Germany, 2012; pp. 33–46.
15. Butalia, M.A.; Ingle, M.; Kulkarni, P. Facial expression recognition for security. *Int. J. Mod. Eng. Res. (IJMER)* **2012**, *2*, 1449–1453.

16. Lim, T.B.; Husin, M.H.; Zaaba, Z.F.; Osman, M.A. Implementation of an automated smart home control for detecting human emotions via facial detection. In Proceedings of the 5th International Conference on Computing and Informatics, ICOCI 2015, Istanbul, Turkey, 11–13 August 2015; pp. 39–45.
17. Bollen, J.; Mao, H.; Zeng, X. Twitter mood predicts the stock market. *J. Comput. Sci.* **2011**, *2*, 1–8. [CrossRef]
18. Williamson, J.R.; Quatieri, T.F.; Helfer, B.S.; Ciccarelli, G.; Mehta, D.D. Vocal and facial biomarkers of depression based on motor in coordination and timing. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 7 November 2014; ACM: New York, NY, USA, 2014; pp. 65–72.
19. Yang, Y.; Fairbairn, C.; Cohn, J.F. Detecting depression severity from vocal prosody. *IEEE Trans. Affect. Comput.* **2013**, *4*, 142–150. [CrossRef] [PubMed]
20. Sivasangari, A.; Ajitha, P.; Rajkumar, I.; Poonguzhali, S. Emotion recognition system for autism disordered people. *J. Ambient Intell. Humaniz. Comput.* **2019**, 1–7. [CrossRef]
21. De Silva, L.C.; Miyasato, T.; Nakatsu, R. Facial emotion recognition using multimodal information. In Proceedings of the IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97), Singapore, 12 September 1997; pp. 397–401.
22. Massaro, D.W. Illusions and issues in bimodal speech perception. In Proceedings of the Auditory Visual Speech Perception'98, Sydney, Australia, 4–7 December 1998; pp. 21–26.
23. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Taylor, J.G. Emotion recognition in human computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [CrossRef]
24. Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.M.; Kazemzadeh, A.; Lee, S.; Neumann, U.; Narayanan, S. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04), State College, PA, USA, 14–15 October 2004; ACM: New York, NY, USA, 2004; pp. 205–211.
25. Chen, L.S.; Huang, T.S.; Miyasato, T.; Nakatsu, R. Multimodal human emotion/expression recognition. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998.
26. Pantic, M.; Rothkrantz, L.J.M. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* **2003**, *91*, 1370–1390. [CrossRef]
27. Vinodhini, G.; Chandrasekaran, R.M. Sentiment analysis and opinion mining: A survey. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **2012**, *2*, 282–292.
28. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
29. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* **2014**, *5*, 1093–1113. [CrossRef]
30. Rustamov, S.; Mustafayev, E.; Clements, M.A. Sentiment analysis using neuro-fuzzy and hidden Markov models of text. In Proceedings of the IEEE Southeastcon 2013, Jacksonville, FL, USA, 4–7 April 2013; pp. 1–6.
31. Kamps, J.; Marx, M.; Mokken, R.; Rijke, M. Using WordNet to measure semantic orientations of adjectives. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004; pp. 1115–1118.
32. Wu, C.; Shen, L.; Wang, X. A new method of using contextual information to infer the semantic orientations of context dependent opinions. In Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence, Shanghai, China, 7–8 November 2009.
33. Peng, T.C.; Shih, C.C. An unsupervised snippet-based sentiment classification method for Chinese unknown phrases without using reference word pairs. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, Canada, 31 August–3 September 2010.
34. Li, G.; Liu, F. A clustering-based approach on sentiment analysis. In Proceedings of the IEEE International Conference on Intelligent System and Knowledge Engineering, Hangzhou, China, 15–16 November 2010; pp. 331–337.
35. Adam, A.; Blockeel, H. Dealing with overlapping clustering: A constraint-based approach to algorithm selection. In Proceedings of the 2015 International Conference on Meta-Learning and Algorithm Selection (MetaSel'15), Porto, Portugal, 7 September 2015; Volume 1455, pp. 43–54.
36. Shetty, P.; Singh, S. Hierarchical clustering: A Survey. *Int. J. Appl. Res.* **2021**, *7*, 178–181. [CrossRef]
37. Maddah, M.; Wells, W.M.; Warfield, S.K.; Westin, C.F.; Grimson, W.E. Probabilistic clustering and quantitative analysis of white matter fiber tracts. In Proceedings of the 2007 Conference on Information Processing in Medical Imaging, Kerkrade, The Netherlands, 2–6 July 2007; Volume 20, pp. 372–383. [CrossRef]
38. Rodriguez, M.Z.; Comin, C.H.; Casanova, D.; Bruno, O.M.; Amancio, D.R.; Costa, L.D.F.; Rodrigues, F. Clustering algorithms: A comparative approach. *PLoS ONE* **2019**, *14*, e0210236. [CrossRef] [PubMed]
39. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* **1991**, *21*, 660–674. [CrossRef]
40. Ruiz, M.; Srinivasan, P. Hierarchical neural networks for text categorization. In Proceedings of the ACM SIGIR Conference 1999, Berkeley, CA, USA, 15–19 August 1999.
41. Tung, A.K.H. Rule-based classification. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer: Boston, MA, USA, 2009. [CrossRef]
42. Garg, A.; Roth, D. Understanding probabilistic classifiers. In *Machine Learning: ECML 2001*; De Raedt, L., Flach, P., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2001; Volume 2167. [CrossRef]

43. Melville, P.; Gryc, W. Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the KDD'09, Paris, France, 28 June–1 July 2009; ACM 978-1-60558-495-9/09/06.
44. Aggarwal, C.C.; Zhai, C.X. *Mining Text Data*; Springer Science + Business Media: New York, NY, USA; Dordrecht, The Netherlands; Heidelberg, Germany; London, UK, 2012.
45. Jian, Z.; Chen, X.; Wu, H.-S. Sentiment classification using the theory of ANNs. *J. China Univ. Posts Telecommun.* **2010**, *17*, 58–62.
46. Moraes, R.; Valiati, J.F.; Neto, W.P.G. Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Syst. Appl.* **2013**, *40*, 621–633. [CrossRef]
47. Kang, H.; Yoo, S.J.; Han, D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. *Expert Syst. Appl.* **2012**, *39*, 6000–6010. [CrossRef]
48. Zhang, Z.; Ye, Q.; Zhang, Z.; Li, Y. Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Syst. Appl.* **2011**, *38*, 7674–7682. [CrossRef]
49. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, USA, 6–7 July 2002; Volume 10, pp. 79–86.
50. Singh, P.K.; Husain, M.S. Methodological study of opinion mining and sentiment analysis techniques. *Int. J. Soft Comput. (IJSC)* **2014**, *5*, 11–21. [CrossRef]
51. Patil, P.; Yalagi, P. Sentiment analysis levels and techniques: A survey. *Int. J. Innov. Eng. Technol. (IJIET)* **2016**, *6*, 523–528.
52. Stalidis, P.; Giatsoglou, M.; Diamantarasa, K.; Sarigiannidis, G.; Chatzisavvas, K.C. Machine learning sentiment prediction based on hybrid document representation. *arXiv* **2015**, arXiv:1511.09107v1.
53. Prakash, C.; Gaikwad, V.B.; Singh, R.R.; Prakash, O. Analysis of emotion recognition system through speech signal using KNN and GMM classifier. *IOSR J. Electron. Commun. Eng. (IOSR-JECE)* **2015**, *10*, 55–61.
54. Schuller, B.; Rigoll, G.; Lang, M. Hidden Markov model-based speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2003, Hong Kong, China, 6–10 April 2003; Volume 2.
55. Nwe, T.; Foo, S.; De Silva, L. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623. [CrossRef]
56. Hu, H.; Xu, M.; Wu, W. GMM supervector based SVM with spectral features for speech emotion recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Honolulu, HI, USA, 15–20 April 2007.
57. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 57–366. [CrossRef]
58. Pao, T.; Chen, Y.; Yeh, J. Emotion recognition from Mandarin speech signals. In Proceedings of the International Symposium on Chinese Spoken Language Processing, Hong Kong, China, 15–18 December 2004.
59. Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. In *Readings in Speech Recognition*; Waibel, A., Lee, K.-F., Eds.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1990; pp. 267–296.
60. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
61. Lee, C.; Yildrim, S.; Bulut, M.; Kazemzadeh, A.; Busso, C.; Deng, Z.; Lee, S.; Narayanan, S. Emotion recognition based on phoneme classes. In Proceedings of the ICSLP 2004, Jeju Island, Korea, 4–8 October 2004; pp. 2193–2196.
62. Reshma, M.; Singh, A. Speech emotion recognition by Gaussian mixture model. *Int. J. Comput. Sci. Inf. Technol.* **2015**, *6*, 2969–2971.
63. Hendy, N.A.; Farag, H. Emotion recognition using neural network: A comparative study. *World Acad. Sci. Eng. Technol.* **2013**, *7*, 433–439.
64. Navas, E.; Hernáez, I.; Luengo, I. An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 1117–1127. [CrossRef]
65. Atassi, H.; Esposito, A. A speaker independent approach to the classification of emotional vocal expressions. In Proceedings of the Twentieth International Conference on Tools with Artificial Intelligence, ICTAI 2008, Dayton, OH, USA, 3–5 November 2008; IEEE Computer Society: Washington, DC, USA, 2008; pp. 147–152, ISBN 978-0-7695-3440-4.
66. Lugger, M.; Yang, B. The relevance of voice quality features in speaker independent emotion recognition. In Proceedings of the ICASSP 2007, Honolulu, HI, USA, 15–20 April 2007.
67. Huang, Y.; Tian, K.; Wu, A.; Zhang, G. Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition. *J. Ambient Intell. Hum. Comput.* **2019**, *10*, 1787–1798. [CrossRef]
68. Sikandar, M. A survey for multimodal sentiment analysis methods. *Int. J. Comput. Technol. Appl.* **2014**, *5*, 1470–1476.
69. Ekman, P.; Oster, H. Facial expressions of emotion. *Ann. Rev. Psychol.* **1979**, *30*, 527–554. [CrossRef]
70. Poria, S.; Cambria, E.; Hussain, A.; Huang, G.-B. Towards an intelligent framework for multimodal affective data analysis. *Neural Netw.* **2015**, *63*, 104–116. [CrossRef] [PubMed]
71. Cerezo, E.; Hupont, I.; Baldassarri, S.; Ballano, S. Emotional facial sensing and multimodal fusion in a continuous 2D affective space. *J. Ambient Intell. Hum. Comput.* **2012**, *3*, 31–46. [CrossRef]
72. Morency, L.-P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multimodal Interfaces, ICMI 2011, Alicante, Spain, 14–18 November 2011; pp. 169–176.

73. Ramos Pereira, M.H.; CardealPádua, F.L.; Machado Pereira, A.C.; Benevenuto, F.; Dalip, D.H. Fusing audio, textual, and visual features for sentiment analysis of news videos. In Proceedings of the ICWSM 2016, Cologne, Germany, 17–20 May 2016; pp. 659–662.
74. Kahou, S.E.; Bouthillier, X.; Lamblin, P.; Gulcehre, C.; Michalski, V.; Konda, K.; Jean, S.; Froumenty, P.; Dauphin, Y.; Boulanger-Lewandowski, N.; et al. Emonets: Multimodaldeeplearningapproachesforemotionrecognitioninvideo. *J. Multimodal User Interfaces* **2016**, *10*, 99–111. [CrossRef]
75. Wollmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; Narayanan, S.S. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSMT modeling. In Proceedings of the Interspeech, Makuhari, Japan, 26–30 September 2010; pp. 2362–2365.
76. Poria, S.; Cambria, E.; Howard, N.; Huang, G.-B.; Hussain, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* **2016**, *174*, 50–59. [CrossRef]
77. Cid, F.; Manso, L.J.; Núñez, P. A novel multimodal emotion recognition approach for affective human robot interaction. In Proceedings of the FinE-R 2015 IROS Workshop, Hamburg, Germany, 28 September–3 October 2015; pp. 1–9.
78. Datcu, D.; Rothkrantz, L. Multimodal recognition of emotions in car environments. In Proceedings of the Second Driver Car Interaction & Interface Conference (DCI&I-2009), Praag, Czech Republic, 2–3 November 2009.
79. Meftah, I.T.; Le Thanh, N.; Ben Amar, C. Multimodal approach for emotion recognition using a formal computational model. *Int. J. Appl. Evol. Comput. (IJAEC)* **2013**, *4*, 11–25. [CrossRef]
80. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *PAMI* **2009**, *31*, 39–58. [CrossRef]
81. Zeng, Z.; Tu, J.; Pianfetti, B.M.; Huang, T.S. Audio–visual affective expression recognition through multistream fused HMM. *Trans. Multimed.* **2008**, *10*, 570–577. [CrossRef]
82. Fragopanagos, N.; Taylor, J.G. Emotion recognition in human–computer interaction. *Neural Netw.* **2005**, *18*, 389–405. [CrossRef]
83. Caridakis, G.; Malatesta, L.; Kessous, L.; Amir, N.; Paouzaiou, A.; Karpouzis, K. Modeling naturalistic affective states via facial and vocal expressions recognition. In Proceedings of the 8th International Conference on Multimodal Interfaces (ICMI '06), Banff, AB, Canada, 2–4 November 2006; pp. 146–154.
84. You, Q.; Luo, J.; Jin, H.; Yang, J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM'16), San Francisco, CA, USA, 22–25 February 2016; ACM: New York, NY, USA, 2016; pp. 13–22.
85. Siddiqui, M.F.H.; Javaid, A.Y. A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images. *Multimodal Technol. Interact.* **2020**, *4*, 46. [CrossRef]
86. Zhou, W.; Cheng, J.; Lei, X.; Benes, B.; Adamo, N. *Deep Learning-Based Emotion Recognition from Real-Time Videos*; HCI: Jacksonville, FL, USA, 2020.
87. Pandeya, Y.R.; Bhattarai, B.; Lee, J. Deep-learning-based multimodal emotion classification for music videos. *Sensors* **2021**, *21*, 4927. [CrossRef] [PubMed]
88. Khorrami, P.; Le Paine, T.; Brady, K.; Dagli, C.; Huang, T.S. How deep neural networks can improve emotion recognition on video data. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 619–623. [CrossRef]
89. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Multimodal emotion recognition using deep learning architectures. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9. [CrossRef]
90. Abdullah, S.M.A.; Ameen, S.Y.A.; Sadeeq, M.A.M.; Zeebaree, S. Multimodal emotion recognition using deep learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 52–58. [CrossRef]
91. Salido Ortega, M.; Rodríguez, L.; Gutierrez-Garcia, J.O. Towards emotion recognition from contextual information using machine learning. *J. Ambient Intell. Human Comput.* **2019**, *11*, 3187–3207. [CrossRef]
92. Perifanos, K.; Goutsos, D. Multimodal hate speech detection in Greek social media. *Multimodal Technol. Interact.* **2021**, *5*, 34. [CrossRef]
93. Caschera, M.C.; Ferri, F.; Grifoni, P. InteSe: An integrated model for resolvingambiguities in multimodalsentences. *IEEE Trans. Syst. Man Cybern. Syst.* **2013**, *43*, 911–931. [CrossRef]
94. Sapiński, T.; Kamińska, D.; Pelikant, A.; Ozcinar, C.; Avots, E.; Anbarjafari, G. Multimodal Database of Emotional Speech, Video and Gestures. *World Acad. Sci. Eng. Technol. Int. J. Comput. Inf. Eng.* **2018**, *12*, 809–814.
95. Caschera, M.C.; D'Ulizia, A.; Ferri, F.; Grifoni, P. MCBF: Multimodal Corpora Building Framework. In *Human Language Technology: Challenges for Computer Science and Linguistics*; Volume 9561 of the Series Lecture Notes in Computer Science; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 177–190.
96. Available online: https://cdn.crowdemotion.co.uk (accessed on 22 February 2017).
97. crowdemotion api. Available online: https://cdn.crowdemotion.co.uk/demos/api-demo/index.html (accessed on 22 February 2017).
98. Available online: http://apidemo.theysay.io/ (accessed on 22 February 2017).
99. Criptodivisas en Pruebas. Available online: http://www.theysay.io/ (accessed on 22 February 2017).

100. Eyben, F.; Weninger, F.; Groß, F.; Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In Proceedings of the ACMMM'13, Barcelona, Spain, 21–25 October 2013; pp. 835–838.
101. Software of the Stanford Natural Language Processing Group. Available online: http://Nlp.stanford.edu/software/ (accessed on 8 March 2017).
102. Caschera, M.C.; Ferri, F.; Grifoni, P. An approach for managing ambiguities in multimodal interaction. In *OTM-WS 2007, Part I: LNCS*; Meersman, R., Tari, Z., Eds.; Springer: Heidelberg, Germany, 2007; Volume 4805, pp. 387–397.
103. Marcus, M.P.; Santorini, B.; Marcinkiewicz, M.A. Building a large annotated corpus of English: The Penn Treebank. *Comput. Linguist.* **1994**, *19*, 313–330.
104. Caschera, M.C.; Ferri, F.; Grifoni, P. Ambiguity detection in multimodal systems. In *Advanced Visual Interfaces 2008*; ACM Press: New York, NY, USA, 2008; pp. 331–334.
105. Murray, I.R.; Arnott, J.L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Am.* **1993**, *93*, 1097–1108. [CrossRef]
106. Caschera, M.C.; Ferri, F.; Grifoni, P. From modal to multimodal ambiguities: A classification approach. *JNIT* **2013**, *4*, 87–109.
107. Grifoni, P.; Caschera, M.C.; Ferri, F. Evaluation of a dynamic classification method for multimodal ambiguities based on Hidden markov models. *Evol. Syst.* **2021**, *12*, 377–395. [CrossRef]
108. Grifoni, P.; Caschera, M.C.; Ferri, F. DAMA: A dynamic classification of multimodal ambiguities. *Int. J. Comput. Intell. Syst.* **2020**, *13*, 178–192. [CrossRef]
109. Yakhnenko, O.; Silvescu, A.; Honavar, V. Discriminatively trained Markov model for sequence classification. In Proceedings of the ICDM'05: Fifth IEEE International Conference on Data Mining, Houston, TX, USA, 27–30 November 2005; pp. 498–505.
110. Doan, A.E. Type I and Type II Error. In *Encyclopedia of Social Measurement*; Kempf-Leonard, K., Ed.; Elsevier: Amsterdam, The Netherlands, 2005; pp. 883–888. ISBN 9780123693983. [CrossRef]
111. Manliguez, C. Generalized Confusion Matrix for Multiple Classes. 2016. Available online: https://www.researchgate.net/publication/310799885_Generalized_Confusion_Matrix_for_Multiple_Classes (accessed on 22 February 2017). [CrossRef]
112. Mesquita, B.; Boiger, M.; De Leersnyder, J. Doing emotions: The role of culture in everyday emotions. *Eur. Rev. Soc. Psychol.* **2017**, *28*, 95–133. [CrossRef]
113. Martin-Krumm, C.; Fenouillet, F.; Csillik, A.; Kern, L.; Besancon, M.; Heutte, J.; Paquet, Y.; Delas, Y.; Trousselard, M.; Lecorre, B.; et al. Changes in emotions from childhood to young adulthood. *Child Indic. Res.* **2018**, *11*, 541–561. [CrossRef]