Emotion Mining: from Unimodal to Multimodal Approaches

Chiara Zucco, Barbara Calabrese, and Mario Cannataro^(✉)

Data Analytics Research Center, Department of Medical and Surgical Sciences,
University “Magna Græcia” of Catanzaro, Catanzaro, Italy
{chiara.zucco,calabreseb,cannataro}@unicz.it

Abstract. In the last decade, Sentiment Analysis and Affective Computing have found applications in different domains. In particular, the interest of extracting emotions in healthcare is demonstrated by the various applications which encompass patient monitoring and adverse events prediction. Thanks to the availability of large datasets, most of which are extracted from social media platforms, several techniques for extracting emotion and opinion from different modalities have been proposed, using both unimodal and multimodal approaches. After introducing the basic concepts related to emotion theories, mainly borrowed from social sciences, the present work reviews three basic modalities used in emotion recognition, i.e. textual, audio and video, presenting for each of these i) some basic methodologies, ii) some among the widely used datasets for the training of supervised algorithms and iii) briefly discussing some deep Learning architectures. Furthermore, the paper outlines the challenges and existing resources to perform a multimodal emotion recognition which may improve performances by combining at least two unimodal approaches. architecture to perform multimodal emotion recognition.

Keywords: Sentiment analysis · Affective computing · Text mining · Data mining · Neuroscience · Hardware accelerators

1 Introduction

Affect is a term widely used in psychology to describe the psychophysiological response to stimuli, as for example emotions, while the word sentiment refers to a more organized and highly socialized feeling, which can be summarized with the expression “an opinion colored by emotion” [1].

Several are the studies showing how affective phenomena may influence different human cognitive processes such as the mechanisms of attention and information processing, as well as the processes of judgment and decision making [2] and communication. On the other hand, it is well known how many pathological conditions, as for example mood disorders, are characterized by a distorted and inconsistent emotional state [3,4].

In humans the ability to be aware, to express and to recognize their own and other's emotions through communicative processes is more developed than in other animals. This outstanding capacity reflects the additional hierarchical levels of processing. Levels that allow learning, inference and simulation [5] and which may constitute a crucial aspect in emulating human intelligence [6]. The development of computational models capable of mimic the natural ability to identify emotional states from speech, facial expressions and written messages is a challenging topic which is gaining particular interest in recent years. The main areas for the computational study related to the recognition of emotions are Affective Computing and Sentiment Analysis. Historically, Affective Computing relates to the area of Artificial Intelligence (AI) that aims at the development of systems capable of recognizing, interpreting and simulating emotions understood in the meaning of affect, and therefore the detection of affective human states is more focused on different biosignals, while Sentiment Analysis aims at extracting opinions and emotions in the sense of sentiment, and was mainly focused on textual sources.

As a result of the popularity of social platforms, the availability of heterogeneous content related to opinions and emotions is increasingly growing, offering the possibility to collect and merge the multimodal information for knowledge extraction. Since communication among human is a mix of verbal and nonverbal content, a system able to measure the emotional state of a person would take advantage from a multimodal approach. For this reason, lately different efforts in performing multimodal emotion recognition have been made.

Even though significant AI advancements, the topic continues to pose numerous open challenges both in the field of research and in large industrial sectors, due to the significant impact on marketing strategies [7], recommender systems [8] and, recently, also in the medical and psychological field for the development of diagnostic and therapeutic clinical decision support systems [9,10].

The present work aims to provide a general overview of the technologies and methodologies involved in recognition of emotional states and to give an insight into the opportunity and challenges for developing an emotion recognition system that merges different modalities. The paper is organized as follows: in Sect. 2 the fundamental and most accepted approaches for the scientific study of emotions in psychological, cognitive and social science are presented. Section 3 is devoted to discuss the basic unimodal approaches used in emotion recognition, and also to present some existing datasets and used tools. Section 4 discusses the use of Deep Learning (DL) in emotion recognition. In Sect. 5 the opportunities and challenges which arises in developing a multimodal emotion recognition system are outlined. Finally, Sect. 6 concludes the paper.

2 Emotion Theories

The difficulty of defining emotions in scientific terms is a well-known problem in the history of psychology. Currently, there is no scientific consensus on the definition of emotion, but there are several heuristic theories which can be grouped into

two different viewpoints: the “basic emotions” and the “appraisal” approaches[2]. Despite the variability of the emotional responses, which may be linked to subjective, cognitive and cultural differences, basic emotions models presuppose the existence of universal and easily recognizable psychophysiological states.

The main assumption underlying the basic emotions approach [11, 12], is that emotions belonging to the same class may vary in intensity or other dimensions, but share not only comparable causes and stimuli of responses but also have a biological analogy, as for example similar behavioral patterns, bodily activations and facial expressions. This strands tends to identify this set of similar emotions with the most prototypical one (for example “joy” contains “happiness”, “enjoyment”, “pleasure”, “joyfulness”, “ecstasy”, “thrill” etc.) [13].

On the other hand, appraisal approaches [14–16] assume that emotions are triggered by a physiological response which, unlike the basic emotions approach, derives from an interpretation of the specific situation through personal criteria [13]. With respect to the identification and classification of emotions, there is a tendency to differentiate between discrete and dimensional theories of emotions. A short description of these models will be given below.

2.1 Discrete Theories of Emotions

The models of discrete emotions are generally the most used in the field of Affective Computing, especially as regards the recognition of emotions from facial expressions. Numerous models of discrete basic human emotions exist, differing among each other in the number and the type of identified emotions. Some widespread criteria to identify basic emotions from non-basic ones are: 1) a universally recognizable facial expression, 2) a rapid spontaneous and automatic recognition, 3) a unique feeling [11]. To date, the most accredited models of discrete emotions in the scientific community refer to the theories of Arnold, Ekman, Izard, Oatley and Johnson-Laird, Plutchik and Tomkins. Table 1 summarizes the list of basic emotions for each of these theories.

Table 1. Emotions definitions.

Categorical emotion theory	Identified emotions
Arnold [17]	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman [11]	Anger, disgust, fear, joy, sadness, surprise
Izard [12]	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Oatley and Johnson-Laird [18]	Anger, disgust, anxiety, happiness, sadness
Plutchik [19]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Tomkins [20]	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise

Among the theories summarized in the Table 1, the one proposed by Oatley and Johnson-Laird [18] has the difference that the existence of basic emotions not only has biological foundations but also a semantic component. The authors address basic emotions as “semantic primitives”, which means that humans know they are feeling a particular emotion X but they don’t know how to define it.

In [21], an interesting work related to the evaluation of emotion theories for computational purposes is carried out. In particular, starting from a corpus of over 21,000 tweets, six basic emotions theories were analyzed through an iterative clustering algorithm based on a variant of Latent Semantic Analysis to discern which one has the most semantically distinct set of emotions. Results showed that Ekman’s model, which is the most popular in Affective Computing, is the one in which emotions are more semantically distinct. Then, Bann et al. [21] also considered 21 emotions given by joining all the six different models and extracted the optimal semantically separated basic emotion set which was proposed as a new model of basic emotions consisting of eight emotions: Accepting, Ashamed, Contempt, Interested, Joyful, Pleased, Sleepy, Stressed.

2.2 Dimensional Emotional Models

Although the most accredited paradigm in the field of neuroscience research states that emotions can be divided into discrete and independent categories, numerous dimensional models were proposed in the literature. In dimensional theories take as an assumption that all the affective states derive from common neurophysiological systems. Consequently, every emotion can be expressed as a combination of these systems, also addressed as dimensions. Only a few are the dimensional theories widely accepted by the scientific community, described below.

The complex model of emotions of Russell [22] is one of the first and most well-known dimensional models. The complex model identifies two independent dimensions: valence and arousal, represented as the two dimensions of a plane in which 28 emotions are mapped in a circle. The arousal-nonarousal scale measures the intensity of emotion and constitutes the vertical axis of the representation system. The points belonging to the upper semicircle are characterized by high arousal, while the points belonging to the inferior semicircle are characterized by low arousal. Valence is measured by a pleasure-displeasure scale, which measures the pleasure of an emotion. On the left semicircle, unpleasant emotions are represented, while on the right semicircle pleasant emotions are shown [19].

Furthermore there are hybrid models, for example the Plutchik model is an example of a model that merges the categorical and dimensional approaches. In fact the Plutchik basic emotions represented in Table 1 only refer to the primary Plutchik’s emotions. In fact, in his model, named “Wheel of emotions”, affective states are represented in a structural and circocentric way. The proximity to the center represents a greater intensity, while the eight dimensions identified are visually represented as eight sectors inscribed in the circocentric structure and arranged as four pairs of opposites: Joy-Sadness, Fear-Anger, Anticipation-Surprise, Disgust-Trust.

3 Basic Unimodal Emotion Recognition Approaches

The focus of this Section is to present the three most used modalities for the recognition of emotions: text, images and audio. For each of the modality, the main tasks, the existing approaches, the available datasets and the functionalities of some existing tools in the literature will be presented. We term those methods as “unimodal” because each one uses only a type of input data to detect emotions.

3.1 Emotion Recognition from Textual Sources

As stated in Sect. 1, in general extracting opinions and emotions from textual content refers to the area known as Sentiment Analysis which may be seen as the intersection of statistical methods, Machine Learning, Information Retrieval and Natural Language Processing.

In the foundational works of Liu, a computational definition of emotion and sentiment is presented [23] and to date it is widely accepted among the Sentiment Analysis research community.

According to his definition, an opinion is a quintuple (entity, entity’s feature, sentiment, opinion holder, time) and the most basic Sentiment Analysis task is polarity detection, whose main goal is to detect whether a text unit contains a positive, negative, or neutral opinion, and/or also considering a “valence” score, indicating how strongly T is positive or negative. Valence scores can be expressed as a nominal (strong negative, weak negative, weak positive, strong positive), or as a continuous variable, frequently belonging to a specific range (for example $[-1, 1]$). Similarly, an emotion can be seen as a quintuple in which sentiment is replaced with an emotion type. More in details, a categorical emotion type may be expressed as the couple (emotion class, emotion intensity), where:

- the emotion class indicates the class to which the specific emotion belongs w.r.t. a given system of basic emotions representation;
- the emotion intensity represents a “valence” score, indicating how strongly the emotion is expressed in the given text unit.

When considering a discrete theory of emotions, a basic emotion detection task in SA can be seen as the multiclass classification problem that, given an input text T and a list $E = [E_1, \dots, E_k]$ of basic emotions classes, aims at detecting whether the text T contains one emotion E_i for $i = 1 \dots k$ and, eventually extracting the respective valence score v_i . However, as will be shown in the following Subsection, existing annotated emotion datasets do not contain only basic emotions and therefore the classification problem usually is designed as a multiclass and multilabel problem. If instead a dimensional theory of emotions is taken into account, a basic emotion detection task may be seen as the regression problem of assigning valence or arousal values to a text T , on the basis of its content.

A SA process generally follows a standard Text Mining process. Input data are extracted from social media, or other sources converted in plain text format and pre-processed. Pre-processing methods include standard NLP and text mining techniques such as stemming, tokenization, part of speech tagging, entity extraction and relation extraction. For an online text, specific data pre-processing methods include cleaning, like removing URLs, HTML tags, abbreviation expansion, emoji, and repeated characters handling. The intermediate step of a Sentiment Analysis process is represented by the analysis module that can follow three types of approaches, that is, supervised, unsupervised and hybrid approach and it is based on three possible levels of analysis, i.e., document-based, sentence-based and aspect-based [24–26].

In lexicon based approaches, the starting point is a set of words in which for every term a given emotion or relative polarity is associated, with or without a score. This set of words can be manually expanded through the use of synonyms or antonyms following a dictionary-based approach. A major issue of lexicon based approaches is to not take into account the specific application domain, with a resulting low text-contextualization capability. Also statistical and semantic methods have been used to enrich the set of annotated words, following a so called corpus-based approach. Despite having the advantage that the performance does not depend on the size of the dataset, as in machine learning approaches, a significant drawback in lexicon-based approach is that it is not suitable for the rapid change to which the language of the web is subject. To exploit the advantages of the two previous approaches, hybrid methodologies that combine machine learning techniques with lexicon-based approaches have been developed.

3.1.1 Available Textual Datasets

To identify the polarity/emotions of an input text, supervised methods need a set of annotated data on which the model is trained. One of the challenges in sentiment analysis is that polarity and emotions expressed in the text strongly depend on used language, context and domain.

Most of the datasets used within the SA are manually annotated. One of the problems related to manual annotation is that the human evaluation of “sentiment” is strongly conditioned by personal experiences, thoughts and beliefs. It is estimated that different people who read a text agree on the generic “sentiment” contained in it only in the 60–65% of cases on average. It is therefore clear how difficult it might be to obtain high quality datasets, reaching high values of “inter-annotation agreement” levels, especially in the case of recognition of emotions. However, the classification of polarity is the textual categorization task that currently holds the largest number of well-noted datasets. For this reason, in this section most of the datasets reported are mainly annotated to perform polarity detection. Few are the datasets that can be used as a benchmark for the evaluation of the performance of the classification algorithms.

- IMDb is a dataset of movie reviews collected from “Internet Movie Database” (IMDb). There are several versions of the dataset that have been collected and annotated. The most used versions are those noted at the document level. In particular, in the version of Pang and Lee [27], known as “Movie Review Dataset”, there are 2000 reviews, 1000 categorized as positive and 1000 as negative. A second dataset (called “IMDb dataset”) was annotated by Maas et al. [28] and consists of 50,000 movie reviews. Positive polarity is associated with the text of the review if the movie has been evaluated with more than six stars and negative polarity otherwise.
- Stanford Twitter Sentiment (STS), also known as Sentiment 140¹, is an annotated dataset introduced by Go et al. [29]. The training set contains 1.6 million tweets containing emoticons. The annotation was performed automatically by assigning a positive label to the tweet containing positive emoticons (:), :-), :), :D, or =) and negative if it contains negative emoticons :(, :-(, o: (. However, since the emoticons may not reflect the actual sentiment of the tweet, the dataset has been extensively used for subjectivity classification tasks as well as a dataset for sentiment analysis.
- Sentiment Strength Twitter Dataset (SS-Tweet). Proposed by Thelwall et al. [30] for the evaluation of the SentiStrength tool, the dataset contains 4242 manually recorded tweets. Unlike the datasets described so far, the annotation is ordinal, in a range -5 (extremely negative) to 5 (extremely positive).
- SemEval Datasets: SemEval (Semantic Evaluation) is a series of computational competitions of semantic analysis systems taking place annually. The sentiment analysis task was introduced for the first time in SemEval-2013. The dataset has been annotated with the use of Amazon Mechanical Turk², for a total of 15196 tweets annotated for SA task at document and aspect level. The datasets from SemEval-2014 to SemEval-2016 are extensions of SemEval2013. In SemEval2016 the dataset has been extended to include other tasks, such as the quantification of tweets with the aim of estimating the distribution of tweets between classes compared to individual tweets. Finally, the SemEval-2017 and SemEval-2018 competitions were more focused on the categorization of affect in the Tweets. They are emotional datasets, scoring for a single emotion, rating for a single emotion, classification among 9 emotions and in addition the neutral class, scoring and valence rating (agreement of terms of positive or negative sentiment) [31].

3.2 Affective Computing Methodologies

A generic Affective Computing process starts with one or more biosignal acquisition. Typically these include measures related to physical aspects and physiological signals. Only the first category, whose standard modalities are facial or body expressions (such as gestures and movements) and speech will be discussed.

After this first step, a pre-processing of these signals is needed to remove or decrease the noise, that can be given, for example, by artifacts and, consequently,

¹ <http://help.sentiment140.com/for-students>.

² <http://mturk.com>.

increase the Ratio between Signal and Noise (SNR). Other pre-processing tasks are filtering and segmentation, concerning events or stimuli. A feature selection task can be applied to perform the analysis only on a reduced feature set.

Depending on the type of analysis to be performed, different features can be considered, for example, time (e.g., statistical analysis), frequency (e.g., Fourier analysis), time-frequency (e.g., wavelets), or power domain (e.g., periodogram and autoregression).

3.3 Emotion Recognition from Facial Expression

The foundational study in [32] reported that the 55% of the communication is visual. Therefore, expressions and body gestures are considered the most obvious and significant channels to infer affect. Ekman’s theory of basic emotion [11] is the dominant emotion theory to classify facial expressions.

Each of the basic emotions is characterized by a series of muscular movements, formalized by what is called the Facial Action Coding System (FACS) and reported in Table 2. The Facial Action Code System (FACS) was published by Paul Ekman and Wallace Friesen in 1978 and, subsequently, updated in 1992 and again in 2002 [33]. The FACS was widely used for experiments on the recognition of emotions by the computer, from human facial expressions. This system objectively measures the frequency and intensity of facial expressions and deduces what is called an action unit (AU).

In order to provide a specific index for each type of movement and expression, the FACS takes into consideration 44 fundamental units named by Ekman and Friesen “Action Units AU” which can give rise to more than 7000 possible combinations. The total number of classified movements or characteristics is 58, some of which are typically associated with a specific emotion, while others are not associated with any other specific emotion.

Table 2. Description of facial expressions in relation to the six Ekman’s basic emotions theory.

Emotion	Description of facial expression
Disgust	Curled nose, raised cheeks, raised upper lip, upper eyelid and lowered eyebrows, raised lower eyelid
Anger	Lowered eyebrows tending to join in the center, tensions in the upper and lower eyelids, pressed lips
Joy	Wrinkles around the eyes, upper lower eyelids and raised cheeks, corners of the mouth stretched upwards
Sadness	Inner angles of the raised eyebrows, inner angles of the upper eyelids raised, corners of the mouth downwards
Surprise	Eyebrows raised and bent upwards, eyes wide open, mouth open
Fear	Raised eyebrows that tend to unite, upper eyelid raised, lower eyelid stretched, mouth open and lips stretched outward

Some of the most important techniques for facial expression recognition are briefly described below:

- Active Appearance Models (AAM) [34]: they are well-known algorithms for modeling deformable objects. The models decouple the shape and texture of objects, using a gradient-based model adaptation approach. The most popular applications of AAM include recognition, tracking, segmentation and synthesis.
- Active Shape Models (ASM) [35]: these are statistical models that adapt to the data or object of an image in a manner consistent with the training data provided. These models are mainly used to improve the automatic analysis of images in noisy or messy environments.
- Muscle-based models [36]: these are models that consist of characteristic facial points corresponding to the facial muscles, for detecting the movement of facial components, such as the eyebrows, the eyes and the mouth, thus recognizing facial expressions.
- Constrained local 3D model (CLM-Z) [37] is a non-rigid face tracking model used to trace facial features in various poses, including both depth and intensity information. Non-rigid face tracking refers to points of interest in an image, such as the tip of the nose, the corners of the eyes and lips. The CLM-Z model can be described by the parameters $p = [s, R, q, t]$, where s is a scale factor, R is the rotation of the object, t represents the 2D translation and q is the vector that describes the non-rigid variation of the q .
- GAVAM (Generalized Adaptive View-Based Appearance Model) [38] is a probabilistic structure that combines dynamic or movement-based approaches to track the position and orientation of the head through video sequences and employs user-independent static approaches to detect the head position from an image. GAVAM is considered a real-time high-precision, user-independent algorithm for tracking the position of the head in real time.

3.4 Emotion Recognition from Speech

The analysis of expressive language consists in examining the paralinguistic characteristics, that is, the aspects of verbal and non-verbal communication, such as the tone of the voice and its intensity. The analysis can be conducted from different points of view, including signal processing, linguistics, psychoacoustics and speech recognition.

A first type of analysis is based on the construction of voice production models that try to model speech (speech) considering the breathing mechanisms and the structure of the phonatory apparatus (primarily vocal cords, mouth and nose). A second approach involves the study of speech from the point of view of perception that analyzes how speech is perceived and processed by the ear and the brain. The first approach, specifically, seeks to model the production of the item using mathematical models of the vocal tract. For the formulation of mathematical models, the vocal tract is studied by analyzing images of the vocal part obtained by ultrasonography, digital radiography and magnetic resonance.

Variations in the breathing pattern, specific vocal cord shape factors can determine variations in prosodic parameters, such as duration, intensity, fundamental frequency and spectral content of speech. Specifically, the fundamental frequency is the vibration speed of the vocal cords and depends on the size and tension of the vocal cords at a given instant of time. It can change in relation to stress, emotion and level of intonation.

In [39] and [40] the most relevant features for the recognition of emotions like the pitch contour, the energy of speech signals, and features related to spectral content are described. At the linguistic level, the analysis for the recognition of emotions involves the identification of the intonation of sentences, analysis of effort and accent in the pronunciation of words and sentences.

3.4.1 Existing Databases of Emotional Speech

- EMODB³: The Berlin Database of Emotional Speech (EMODB) is a public German speech database that incorporates audio files with seven emotions: happiness, sadness, anger, fear, disgust, boredom, and neutral [41].
- SAVEE⁴: The Surrey Audio-Visual Expressed Emotion (SAVEE) is a public British English speech database that has audio files with seven emotion labels: happiness, sadness, anger, fear, disgust, surprise, and neutral [42].
- EMOVO⁵: is a public Italian speech database that includes audio files with seven emotion labels: happiness, sadness, anger, fear, disgust, surprise, and neutral [43].

4 Deep Learning Algorithms for Emotion Detection

The performance of the emotion extraction approaches presented so far is strongly related to how data are represented. In this sense, an essential step is feature engineering, i.e., the process that uses domain knowledge to design a good representation of data in terms of suitable features. By taking advantage of the large training dataset, Deep Learning Algorithm seeks to learn data representation along with the mapping that associates each input representation to its output. Moreover, Deep Neural Networks (DNNs) are Artificial Neural Networks designed to have different levels of non-linear and nested operations, that have shown to improve non-linear model tasks. The previous statements are some of the key points explaining the popularity of Deep Neural Networks (DNNs) and why they have increasingly been implemented also to face the problem of emotion recognition, by achieving state-of-the-art results for a wide range of tasks [44, 45]. Several DNNs architectures have been proposed both in Sentiment Analysis and Affective Computing, for example, Convolutional Neural Networks (CNN) [46, 47], Recurrent Neural Networks (RNNs) with or without attention mechanism, Autoencoders [48] and also Deep Belief Networks (DBN) [48–50].

³ <http://emodb.bilderbar.info/start.html>.

⁴ <http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/>.

⁵ <http://voice.fub.it/activities/corpora/emovo/index.html>.

Both in Sentiment Analysis and Affective Computing, a key role is played in capturing long-term dependencies intended for example as extracting relations among distant words in a sentence but also as capturing temporal variations in facial or vocal expressions. To address this problem, a particular set of RNNs model is typically used in a practical application and are called gated recurrent RNNs, in particular, Long Short Term Memory (LSTM) network [51]. For what concerns textual emotion recognition, LSTMs models are the state-of-the-art algorithm in time series predictions features, for example monitoring systems [52]. In Face emotion recognition, Gated RNNs have been combined with CNN to improve sequential images modelling [53].

5 Challenges and Tools for Multimodal Emotion Recognition

Recently, the scientific community has been increasing efforts for the joint application of Sentiment Analysis and Affective Computing techniques to create multi-modal systems, especially for monitoring and preventing mental health.

For example, in [54], a system that combines Sentiment Analysis and Affective Computing techniques to assess a subject's mental health is presented. In particular, the authors proposed the use of embedded sensors in mobile devices (such as laptops and smartphones) to trace head and eye movements, facial expressions as well as heartbeat. Among the features useful to verify interactions among users, the speed of typing, the number of clicks and mouse movements, etc. were considered, starting from the assumption that a positive or negative mood has effects on the different degrees of activity of the user. Lastly, the monitoring of sentiment associated with the text, related to user posts on Twitter, was performed using a free tool for sentiment analysis, i.e., *Sentiment 140*⁶ and also a prediction algorithm for images posted along with tweets was used.

The major challenges for developing an integrated system, especially in combining data from integrated daily devices, can be identified in the following:

1. input data should be appropriate to the type of analysis to be made. For example, even if smartphones allow videos with good quality, the facial expression recognition process requires high-resolution images;
2. the choice of appropriate pre-processing, feature extraction and analysis techniques to achieve good performance is mandatory;
3. the selection of the most suitable approach for integrating information extracted from multiple sources in the system is crucial.

Concerning the last point, commonly used approaches are the following:

- Fusion at feature-level: after a first phase of pre-processing of data extracted from different sources, all features are considered as different components of a joined feature vector, and then classification is performed accordingly.

⁶ <http://www.sentiment140.com/>.

- Fusion at decision level: instead of combining features in a single vector features as in feature-level fusion, a separate classifier for each modality is used. The output of each classifier was treated as a classification score.

In [55], both the previous approaches were tested for integrating facial expression, speech, and textual data to build a multi-modal sentiment analysis framework. The experimental results show that the accuracy of fusion at the feature level is higher than the accuracy of fusion at decision-level. Accurately, the authors reported a precision value of 78.2% and a recall value of 77.1% for tests relative to feature-level fusion, whereas they referred a precision value of 75.2% and a recall value of 73.4% for decision level fusion. Another remarkable point is that, regardless of the fusion techniques, the results show how the simultaneous use of video, text and audio modalities allows achieving better accuracy than when only pairs of the three patterns are considered. Considering the approach based on fusion at features level, the precision values of experiments are: (i) 72.45% by using only visual and text-based features, (ii) 73.21% by using visual and audio-based features and (iii) 71.15% by using audio and text-based features.

5.1 Existing Multimodal Dataset for Emotion Recognition

- SEMAINE Database. This dataset was developed in 2007 by McKeown et al. [56]. It is a large audiovisual database created for building agents capable of involving a person in a prolonged and emotional conversation using a Sensitive Artificial Listener (SAL) [57] paradigm. SAL is an interaction that involves two parts: a ‘man’ and an ‘operator’ (a machine or a person who simulates a machine). There were 150 participants, 959 conversations, each lasting 5 min. For the recordings, participants were asked to speak in turn to four emotionally stereotyped characters. The characters are Prudence, which is balanced and sensitive; Poppy, who is happy and outgoing; Spike, who is angry and in conflict; and Obadiah, who is sad and depressive.
- Interactive emotional dyadic acquisition database (IEMOCAP). The IEMOCAP dataset was developed in 2008 by Busso et al. [58]. 10 actors were asked to record their facial expressions in front of the cameras. In particular, the dataset contains a total of 10 h of recording, each of which expresses one of the following emotions: happiness, anger, sadness, frustration and a neutral state.
- eINTERFACE. This dataset was developed in 2006 by Martin et al. [59] and contains audio and video for the evaluation of algorithms for the recognition of emotions from audio and video. The emotions labeled are: happiness, sadness, surprise, anger, disgust and fear.
- CK++ dataset: the Cohn Kanade dataset contains facial images of 210 adults. The participants are 18–50 years old, 81% Americans, 13% Afro Americans and 6% of other ethnic groups; 69% females. Participants are asked to perform 23 facial expressions.

- Belfast Database. This data set was developed in 2000 by Douglas-Cowie et al. [57]. The database consists of audiovisual data of people discussing emotional issues and are taken from television chat programs and religious programs. Includes 100 speakers and 239 clips, with 1 neutral clip and 1 emotional clip for each speaker. Two types of descriptors were provided for each clip: dimensional and categorical, according to the different emotion approaches.

6 Conclusions

This article presented an overview of the existing approaches for extracting sentiment and emotion from different input modalities through the use of Sentiment Analysis and Affective Computing techniques. In particular, audio, video and textual data were considered and, for each input modality a pipeline of analysis, existing datasets and tools were presented. Deep Learning approaches were also considered and discussed. Subsequently, recent efforts and challenges to combine these different unimodal approaches toward multimodal systems were reported.

References

1. Hockenbur, D., Hockenbur, S.: *Discovering Psychology*. Macmillan, London (2010)
2. Jeon, M.: Emotions and affect in human factors and human-computer interaction: Taxonomy, theories, approaches, and methods. In: *Emotions and Affect in Human Factors and Human-Computer Interaction*, pp. 3–26. Elsevier (2017)
3. Valstar, M.: Automatic behaviour understanding in medicine. In: *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research including Business Opportunities and Challenges*, pp. 57–60. ACM (2014)
4. Zucco, C., Calabrese, B., Cannataro, M.: Sentiment analysis and affective computing for depression monitoring. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1988–1995. IEEE (2017)
5. Smith, R., Lane, R.D., Steklis, N., Weihs, K., Steklis, H.D.: The evolution and development of the uniquely human capacity for emotional awareness: a synthesis of comparative anatomical, cognitive, neurocomputational, and evolutionary psychological perspectives (2019)
6. Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A.: *A Practical Guide to Sentiment Analysis*. Springer International Publishing, Cham (2017). <https://doi.org/10.1007/978-3-319-55394-8>
7. Atzeni, M., Dridi, A., Reforgiato Recupero, D.: Fine-grained sentiment analysis on financial microblogs and news headlines. In: Dragoni, M., Solanki, M., Blomqvist, E. (eds.) *SemWebEval 2017*. CCIS, vol. 769, pp. 124–128. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69146-6_11
8. Jayashree, R., Kulkarni, D.: Recommendation system with sentiment analysis as feedback component. In: Deep, K., et al. (eds.) *Proceedings of Sixth International Conference on Soft Computing for Problem Solving*. AISC, vol. 547, pp. 359–367. Springer, Singapore (2017). https://doi.org/10.1007/978-981-10-3325-4_36
9. Rosenquist, J.N., Fowler, J.H., Christakis, N.A.: Social network determinants of depression. *Mol. Psychiatry* **16**(3), 273–281 (2011)

10. Coviello, L.L., et al.: Detecting emotional contagion in massive social networks. *PloS one* **9**(3), e90315 (2014)
11. Ekman, P., Wallace, V.: *Unmasking the Face*. Malor Book, Cambridge (2003)
12. Izard, C.E.: *The Face of Emotion*. Appleton-Century-Crofts, New York (1971)
13. Gendron, M., Barrett, L.F.: Reconstructing the past: a century of ideas about emotion in psychology. *Emot. Rev.* **1**(4), 316–339 (2009)
14. Frijda, N.H.: *The Emotions*. Cambridge University Press, Cambridge (1986)
15. Lazarus, R.S., Lazarus, R.S.: *Emotion and adaptation*. Oxford University Press on Demand (1991)
16. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* **17**(3), 715–734 (2005)
17. Arnold, M.B.: *Emotion and Personality*. Columbia University Press, New York (1960)
18. Oatley, K., Johnson-Laird, P.N.: Towards a cognitive theory of emotions. *Cogn. Emot.* **1**(1), 29–50 (1987)
19. Plutchik, R.: *Emotion Theory, Research and Experiences*. Academic Press, London (1980)
20. Tomkins, S.S.: Affect theory. In: *Approaches to emotion*, vol. 163, pp. 163–195 (1984)
21. Bann, E.Y., Bryson, J.J.: The conceptualisation of emotion qualia: semantic clustering of emotional tweets. In: *Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop*, pp. 249–263. World Scientific (2014)
22. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161 (1980)
23. Liu, B.: *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge (2015)
24. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng. J.* **5**(4), 1093–1113 (2014)
25. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowl.-Based Syst.* **89**, 14–46 (2015)
26. Zucco, C., Calabrese, B., Agapito, G., Guzzi, P.H., Cannataro, M.: Sentiment analysis for mining texts and social networks data: Methods and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. e1333 (2019)
27. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: *EMNLP 2002 Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, pp. 79–86 (2002)
28. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)
29. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. *CS224N Project Report*, Stanford, pp. 1–12 (2009)
30. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. *J. Am. Soc. Inf. Sci. Technol.* **61**(12), 2544–2558 (2010)
31. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: Semeval-2018 task 1: affect in tweets. In: *Proceedings of the 12th International Workshop on Semantic Evaluation*, pp. 1–17 (2018)

32. Mehrabian, A., Ferris, S.R.: Inference of attitudes from nonverbal communication in two channels. *J. Consult. Psychol.* **31**(3), 248 (1967)
33. Friesen, W.V., Ekman, P.: *Emfacs-7: Emotional facial action coding system*. Unpublished manuscript, University of California at San Francisco
34. Lin, C., Xue, G.-R., Zeng, H.-J., Yu, Y.: Using probabilistic latent semantic analysis for personalized web search. In: Zhang, Y., Tanaka, K., Yu, J.X., Wang, S., Li, M. (eds.) *APWeb 2005. LNCS*, vol. 3399, pp. 707–717. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31849-1_68
35. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. *Comput. Vis. Image Underst.* **61**(1), 38–59 (1995)
36. Ohta, H., Saji, H., Nakatani, H.: Recognition of facial expressions using muscle-based feature models. *Syst. Comput. Jpn.* **31**(10), 78–88 (2000)
37. Baltrušaitis, T., Robinson, P., Morency, L.-P.: 3D constrained local model for rigid and non-rigid facial tracking. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2610–2617. IEEE (2012)
38. Morency, L.-P., Whitehill, J., Movellan, J.: Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In: 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–8. IEEE (2008)
39. Mozziconacci, S.J.L.: Modeling emotion and attitude in speech by means of perceptually based parameter values. *User Modeling and User-Adapted Interaction* **11**(4), 297–326 (2001)
40. 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.* **93**(2), 1097–1108 (1993)
41. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
42. Haq, S., Jackson, P.J.B., Edge, J.: Audio-visual feature selection and reduction for emotion classification. In: *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2008)*, Tangalooma, Australia, September 2008
43. Costantini, G., Iaderola, I., Paoloni, A., Todisco, M.: Emovo corpus: an Italian emotional speech database. In: *International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 3501–3504. European Language Resources Association (ELRA) (2014)
44. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1253 (2018)
45. Li, S., Deng, W.: Deep facial expression recognition: A survey. *arXiv preprint [arXiv:1804.08348](https://arxiv.org/abs/1804.08348)* (2018)
46. dos Santos, C., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 69–78 (2014)
47. Kim, Y.: Convolutional neural networks for sentence classification. *arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882)* (2014)
48. Zhou, X., Guo, J., Bie, R.: Deep learning based affective model for speech emotion recognition. In: 2016 International IEEE Conferences on Ubiquitous Intelligence and Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCOM/IoP/SmartWorld), pp. 841–846. IEEE (2016)

49. Ruangkanokmas, P., Achalakul, T., Akkarajitsakul, K.: Deep belief networks with feature selection for sentiment classification. In: 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 9–14. IEEE (2016)
50. Jin, Y.: Deep Belief Networks for Sentiment Analysis. Ph.D. thesis, University of New Brunswick (2017)
51. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint [arXiv:1503.00075](https://arxiv.org/abs/1503.00075) (2015)
52. Suhara, Y., Xu, Y., Pentland, A.: Deepmood: forecasting depressed mood based on self-reported histories via recurrent neural networks. In: Proceedings of the 26th International Conference on World Wide Web, pp. 715–724. International World Wide Web Conferences Steering Committee (2017)
53. Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., Zareapoor, M.: Hybrid deep neural networks for face emotion recognition. *Pattern Recogn. Lett.* **115**, 101–106 (2018)
54. Zhou, D., et al.: Tackling mental health by integrating unobtrusive multimodal sensing. In: AAAI, pp. 1401–1409 (2015)
55. Poria, S., Cambria, E., Howard, N., Huang, G.B., Hussain, A.: Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* **174**, 50–59 (2016)
56. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The Semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.* **3**(1), 5–17 (2011)
57. Douglas-Cowie, E., Cowie, R., Cox, C., Amir, N., Heylen, D.: The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In: LREC Workshop on Corpora for Research on Emotion and Affect, pp. 1–4. ELRA (2008)
58. Busso, C., et al.: IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**(4), 335 (2008). <https://doi.org/10.1007/s10579-008-9076-6>
59. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The enterface'05 audio-visual emotion database. In: 22nd International Conference on Data Engineering Workshops (ICDEW 2006), p. 8. IEEE (2006)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.