Neurocomputing 174 (2016) 50–59

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Fusing audio, visual and textual clues for sentiment analysis from multimodal content

Soujanya Poria [a], Erik Cambria [b,*], Newton Howard [c], Guang-Bin Huang [d], Amir Hussain [a]

[a] Department of Computing Science and Mathematics, University of Stirling, UK
[b] School of Computer Engineering, Nanyang Technological University, Singapore
[c] Media Laboratory, Massachusetts Institute of Technology, USA
[d] School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form
31 December 2014
Accepted 2 January 2015
Available online 17 August 2015

Keywords:
Multimodal fusion
Big social data analysis
Opinion mining
Multimodal sentiment analysis
Sentic computing

ABSTRACT

A huge number of videos are posted every day on social media platforms such as Facebook and YouTube. This makes the Internet an unlimited source of information. In the coming decades, coping with such information and mining useful knowledge from it will be an increasingly difficult task. In this paper, we propose a novel methodology for multimodal sentiment analysis, which consists in harvesting sentiments from Web videos by demonstrating a model that uses audio, visual and textual modalities as sources of information. We used both feature- and decision-level fusion methods to merge affective information extracted from multiple modalities. A thorough comparison with existing works in this area is carried out throughout the paper, which demonstrates the novelty of our approach. Preliminary comparative experiments with the YouTube dataset show that the proposed multimodal system achieves an accuracy of nearly 80%, outperforming all state-of-the-art systems by more than 20%.

1. Introduction

Subjectivity and sentiment analysis are the automatic identification of private states of the human mind (i.e., opinions, emotions, sentiments, behaviors and beliefs). Further, subjectivity detection focuses on identifying whether data is subjective or objective. Wherein, sentiment analysis classifies data into positive, negative and neutral categories and, hence, determines the sentiment polarity of the data.

To date, most of the works in sentiment analysis have been carried out on natural language processing. Available dataset and resources for sentiment analysis are restricted to text-based sentiment analysis only. With the advent of social media, people are now extensively using the social media platform to express their opinions. People are increasingly making use of videos (e.g., YouTube, Vimeo, VideoLectures), images (e.g., Flickr, Picasa, Facebook) and audios (e.g., podcasts) to air their opinions on social media platforms. Thus, it is highly crucial to mine opinions and identify sentiments from the diverse modalities.

So far the field of multimodal sentiment analysis has not received much attention [1,2], and no prior work has specifically addressed extraction of features and fusion of information extracted from different modalities. In this paper, we discuss the feature extraction process from different modalities as well as the way we use them to build a novel multimodal sentiment analysis framework. For experiments, we have used datasets from YouTube originally developed by [1]. We have employed several supervised machine-learning-based classifiers for the sentiment classification task. The best performance has been obtained with the extreme learning machine (ELM) [3–5], an emerging learning technique that provides efficient unified solutions to generalized feed-forward networks including (but not limited to) single-/multi-hidden-layer neural networks, radial basis function networks, and kernel learning. ELMs offer significant advantages such as fast learning speed, ease of implementation, and minimal human intervention. They thus offer strong potential as a viable alternative technique for large-scale computing and machine learning in many different application fields, including image [6], text [7], and speech [8] processing, as well as multimodal data analysis [9].

The rest of the paper is organized as follows: Section 2 presents motivations behind the proposed work; Section 3 covers related work on emotion and sentiment recognition from different modalities; Section 4 describes the datasets used and proposes an overview of the experiment; next, Sections 5, 6 and 7 explain how visual, audio and textual data are processed, respectively; Section 8 illustrates the methodology adopted for fusing different modalities; Section 9 proposes a proof of concept of real-time multimodal sentiment analysis avatar; Section 10 presents experimental

[*] Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.01.095
0925-2312/© 2015 Elsevier B.V. All rights reserved.

results; finally, Section 11 concludes the paper and outlines future work.

2. Motivations

Research in this field is rapidly growing and attracting the attention of both academia and industry alike. This combined with advances in signal processing and AI has led to the development of advanced intelligent systems that intend to detect and process affective information contained in multimodal sources. The majority of such state-of-the-art frameworks however, rely on processing a single modality, i.e., text, audio, or video. Further, all of these systems are known to exhibit limitations in terms of meeting robustness, accuracy, and overall performance requirements, which, in turn, greatly restrict the usefulness of such systems in real-world applications.

The aim of multi-sensor data fusion is to increase the accuracy and reliability of estimates [10]. Many applications, e.g., navigation tools, have already demonstrated the potential of data fusion. This depicts the importance and feasibility of developing a multimodal framework that could cope with all three sensing modalities: text, audio, and video in human-centric environments. The way humans communicate and express their emotions and sentiments can be expressed as multimodal. The textual, audio, and visual modalities are concurrently and cognitively exploited to enable effective extraction of the semantic and affective information conveyed during communication.

With significant increase in the popularity of social media like Facebook and YouTube, many users tend to upload their opinions on products in video format. On the contrary, people wanting to buy the same product, browse through on-line reviews and make their decisions. Hence, the market is more interested in mining opinions from video data rather than text data. Video data may contain more cues to identify sentiments of the opinion holder relating to the product. Audio data within a video expresses the tone of the speaker, and visual data conveys the facial expressions, which in turn help to understand the affective state of the users. The video data can be a good source for sentiment analysis but there are major challenges that need to be overcome. For example, expressiveness of opinions vary from person to person [1]. A person may express his or her opinions more vocally while others may express them more visually.

Hence, when a person expresses his opinions with more vocal modulation, the audio data may contain most of the clues for opinion mining. However, when a person is communicative through facial expressions, then most of the data required for opinion mining, would have been found in facial expressions. So, a generic model needs to be developed which can adapt itself for any user and can give a consistent result. Our multimodal sentiment classification model is trained on robust data, and the data contains the opinions of many users. In this paper, we show that the ensemble application of feature extraction from different types of data and modalities enhances the performance of our proposed multimodal sentiment system.

3. Related work

Sentiment analysis and emotion analysis both represent the private state of the mind and to-date, there are only two well-known state-of-the-art methods [1,2] in multimodal sentiment analysis. In this section, we describe the research done so far in both sentiment and emotion detection using visual and textual modality. Both feature extraction and feature fusion are crucial for the development of a multimodal sentiment analysis system. Existing research on multimodal sentiment analysis can be categorized into two broad categories: those devoted to feature extraction from each individual modality, and those developing techniques for the fusion of features coming from different modalities.

3.1. Video: emotion and sentiment analysis from facial expressions

In 1970, Ekman et al. [11] carried out extensive studies on facial expressions. Their research showed that universal facial expressions provide sufficient clues to detect emotions. They used anger, sadness, surprise, fear, disgust, and joy as six basic emotion classes. Such basic affective categories are sufficient to describe most of the emotions expressed by facial expressions. However, this list does not include the emotion expressed through facial expression by a person when he or she shows disrespect to someone; thus, a seventh basic emotion, contempt, was introduced by Matsumoto [12]. Ekman et al. [13] developed a facial expression coding system (FACS) to code facial expressions by deconstructing a facial expression into a set of action units (AU). AUs are defined via specific facial muscle movements. An AU consists of three basic parts: AU number, FACS name, and muscular basis. For example, for AU number 1, the FACS name is inner brow raiser and it is explicated via frontalis pars medialis muscle movements. In consideration to emotions, Friesen and Ekman [14] proposed the emotional facial action coding system (EFACS). EFACS defines the sets of AUs that participate in the construction of facial expressions expressing specific emotions.

The Active Appearance Model [15,16] and Optical Flow-based techniques [17] are common approaches that use FACS to understand expressed facial expressions. Exploiting AUs as features like *k*-nearest-neighbors, Bayesian networks, hidden Markov models (HMM), and artificial neural networks (ANN) [18] has helped many researchers to infer emotions from facial expressions. All such systems, however, use different, manually crafted corpora, which makes it impossible to perform a comparative evaluation of their performance.

3.2. Audio: emotion and sentiment recognition from speech

Recent studies on speech-based emotion analysis [16,20–23] have focused on identifying several acoustic features such as fundamental frequency (pitch), intensity of utterance [19], bandwidth, and duration. The speaker-dependent approach gives much better results than the speaker-independent approach, as shown by the excellent results of Navas et al. [24], where about 98% accuracy was achieved by using the Gaussian mixture model (GMM) as a classifier, with prosodic, voice quality as well as Mel frequency cepstral coefficients (MFCC) employed as speech features. However, the speaker-dependent approach is not feasible in many applications that deal with a very large number of possible users (speakers).

To our knowledge, for speaker-independent applications, the best classification accuracy achieved so far is 81% [25], obtained on

Table 1
Sample of SenticNet data.

Concept	Polarity
A lot	+0.258
A lot sex	+0.858
A little	+0.032
Abandon	−0.566
Abase	−0.153
Abash	−0.174
Abashed	−0.174
Abashment	−0.186
Abhor	−0.391
Abhorrence	−0.391

the Berlin Database of Emotional Speech (BDES) [26] using a two-step classification approach and a unique set of spectral, prosodic, and voice features, selected with the Sequential Floating Forward Selection (SFFS) algorithm [27]. As per the analysis of Scherer et al. [28], the human ability to recognize emotions from speech audio is about 60%. Their study shows that sadness and anger are detected more easily from speech, while the recognition of joy and fear is less reliable. Caridakis et al. [29] obtained 93.30% and 76.67% accuracy in identifying anger and sadness, respectively, from speech, using 377 features based on intensity, pitch, Mel frequency cepstral coefficients (MFCC), Bark spectral bands, voiced segment characteristics, and pause length.

3.3. *Text: emotion and sentiment recognition from textual data*

Affective content recognition in text is a rapidly developing area of natural language processing, which has garnered the attention of both research communities and industries in recent years. Sentiment analysis tools have numerous applications. For example, it helps companies to comprehend customer sentiments about products and, political parties to understand what voters feel about party's actions and proposals. Significant studies have been done to identify positive, negative, or neutral sentiment associated with words [30,31], multi-words [32], phrases [33], sentences [34], and documents [35]. The task of automatically identifying fine grained emotions, such as anger, joy, surprise, fear, disgust, and sadness, explicitly or implicitly expressed in a text has been addressed by several researchers [36,37]. So far, approaches to text-based emotion and sentiment detection rely mainly on rule-based techniques, bag of words modeling using a large sentiment or emotion lexicon [38], or statistical approaches that assume the availability of a large dataset annotated with polarity or emotion labels [39].

Several supervised and unsupervised classifiers have been built to recognize emotional content in texts [40]. The SNoW architecture [41] is one of the most useful frameworks for text-based emotion detection. In the last decade, researchers have been focusing on sentiment extraction from texts of different genres, such as news [42], blogs [43], Twitter messages [44], and customer reviews [45] to name a few. Sentiment extraction from social media helps to predict the popularity of a product release, results of election poll, etc. To accomplish this, several knowledge-based sentiment [42] and emotion [46] lexicons have been developed for word- and phrase-level sentiment and emotion analysis.

3.4. *Studies on multimodal fusion*

The ability to perform multimodal fusion is an important prerequisite to the successful implementation of agent–user interaction. One of the primary obstacles to multimodal fusion is the development and specification of a methodology to integrate cognitive and affective information from different sources on different time scales and measurement values. There are two main fusion strategies: feature-level fusion and decision-level fusion.

Feature-level fusion [47] combines the characteristics extracted from each input channel in a 'joint vector' before any classification operations are performed. Some variations of such an approach exist, e.g., Mansoorizadeh et al. [48] proposed asynchronous feature-level fusion. Modality fusion at feature-level presents the problem of integrating highly disparate input features, suggesting that the problem of synchronizing multiple inputs while re-teaching the modality's classification system is a nontrivial task.

In decision-level fusion, each modality is modeled and classified independently. The unimodal results are combined at the end of the process by choosing suitable metrics, such as expert rules and simple operators including majority votes, sums, products and statistical weighting. A number of studies favor decision-level fusion as the preferred method of data fusion because errors from different classifiers tend to be uncorrelated and the methodology is feature-independent [49]. Bimodal fusion methods have been proposed in numerous instances [50,51], but optimal information fusion configurations remain elusive.

Cambria et al. [2] proposed a novel methodology termed *Sentic Blending* to fuse the modalities in order to grasp emotion associated with the multimodal content. Unlike other approaches, they fused facial expressions with natural language text, and also tracked the sentiment change over time. For the experiment they used FGNET [52] and MMI [53] datasets. Paleari et al. [54] carried out both decision- and feature-level fusion. They experimented with the eNTERFACE dataset and showed that decision-level fusion outperformed feature-level fusion. Many multimodal methodologies have ad hoc workarounds for the purpose of fusing information from multiple modalities, but the entire system must be retrained before new modalities can be included. Also, they are not adaptive to quality changes in input. Thus, in order to better adapt to data trends it is preferable not to perform long-term adjustments.

4. Datasets employed

4.1. *YouTube dataset*

This is the only available dataset developed by [1]. Forty-seven videos were collected from the social media web site YouTube. Videos in the dataset were about different topics (for instance politics, electronics product reviews, etc.). The videos were found using the following keywords: opinion, review, product review, best perfume, toothpaste, war, job, business, cosmetics review, camera review, baby product review, I hate, I like [1]. The final video set had 20 female and 27 male speakers randomly selected from YouTube, with their age ranging approximately from 14 to 60 years. Although they belonged to different ethnic backgrounds (e.g., Caucasian, African-American, Hispanic, Asian), all speakers expressed themselves in English.

The videos were converted to mp4 format with a standard size of 360×480. The length of the videos varied from 2 to 5 min. All videos were pre-processed to avoid the issues of introductory titles and multiple topics. Many videos on YouTube contained an introductory sequence where a title was shown, sometimes accompanied by a visual animation. To address this issue first 30 s was removed from each video. Morency et al. [1] provided transcriptions with the videos. Each video was segmented and each segment was labeled by a sentiment, thanks to [1]. Because of this annotation scheme of the dataset, textual data was available for our experiment. We used YouTube dataset in our experiment to build the multimodal sentiment analysis system as well as to evaluate the system's performance. Section 10 presents this process in detail.

4.2. *SenticNet*

As an a priori polarity lexicon of concepts, we used SenticNet 3.0 [32], a lexical resource that contains 30,000 concepts along with their polarity scores in the range from −1.0 to +1.0. SenticNet 3.0 also contains all WordNet Affect (WNA) [36] concepts. The first 10 SenticNet concepts in lexicographic order along with the corresponding polarities are shown in Table 1.

4.3. EmoSenticNet

We also used EmoSenticNet [55], an extension of SenticNet containing about 13,741 common-sense knowledge concepts, including those concepts that exist in the WNA list, along with their affective labels in the set anger, joy, disgust, sadness, surprise, fear. In order to build a suitable knowledge base for emotive reasoning, we applied the so-called blending technique to ConceptNet and EmoSenticNet. Blending is a technique that performs inference over multiple sources of data simultaneously, taking advantage of the overlap between them.

It linearly combines two sparse matrices into a single matrix, in which the information between two initial sources is shared. Before performing blending, we represent EmoSenticNet as a directed graph similar to ConceptNet. For example, the concept birthday party was assigned an emotion joy. We considered these as two nodes, and added the assertion HasProperty on the edge directed from the node birthday party to the node joy. Next, we converted the graphs to sparse matrices in order to blend them. After blending the two matrices, we performed Truncated Singular Value Decomposition (TSVD) on the resulting matrix to discard the components that represented relatively small variations in the data. Only 100 significant components of the blended matrix were retained in order to produce a good approximation of the original matrix. The number 100 was selected empirically: the original matrix was found to be best approximated using 100 components.

4.4. Overview of the experiment

First, we present an empirical method used for extracting the key features from visual and textual data for sentiment analysis. Then, we describe a fusion method employed to fuse the extracted features for automatically identifying the overall sentiment expressed by a video:

- In YouTube dataset each video was segmented into several parts. According to the frame rate of the video, we first converted each video segment into images. Then, for each video segment we extracted the facial features from all images and took the average to compute the final feature vector. Similarly, the audio and textual features were also extracted from each segment of the audio signal and text transcription of the video clip, respectively.
- Next, we fused the audio, visual and textual feature vectors to form a final feature vector which contained the information of both audio, visual and textual data. Later, a supervised classifier was employed on the fused feature vector to identify the overall polarity of each segment of the video clip. On the other hand, we also carried out an experiment on decision-level fusion, which took the sentiment classification result from 3 individual modalities as inputs and produced the final sentiment label as an output.

5. Extracting features from visual data

Humans are known to express emotions in a number of ways, including, to a large extent, through the face. Facial expressions play a significant role in the identification of emotions in a multimodal stream. A facial expression analyzer automatically identifies emotional clues associated with facial expressions, and classifies facial expressions in order to define sentiment categories and to discriminate between them. We use positive, negative and neutral as sentiment classes in the classification problem. In the annotations provided with the YouTube dataset, each video was segmented into some parts and each of the sub segments was of few seconds duration. Every segment was annotated as either 1, 0, or -1 denoting positive, neutral or negative sentiment.

Using a MATLAB code, we converted all videos in the dataset to image frames. Subsequently, we extracted facial features from each image frame. To extract facial characteristic points (FCPs) from the images, we used the facial recognition software Luxand FSDK 1.7. From each image, we extracted 66 FCPs; see examples in Table 2. The FCPs were used to construct facial features, which are defined as distances between FCPs; see examples in Table 3.

GAVAM [56] was also used to extract facial expression features from the face. Table 4 shows the extracted features from facial images. In our experiment we used the features extracted by FSDK 1.7 along with the features extracted using GAVAM. If a segment of a video has n number of images, then we extracted features from each image and take average of those feature values in order to compute the final facial expression feature vector for a segment. We used an ELM classifier to build the sentiment analysis model from the facial expressions. 10-fold cross validation was carried out on the dataset producing 68.60% accuracy.

Table 2
Some relevant facial characteristic points (out of the 66 facial characteristic points detected by Luxand).

Features	Description
0	Left eye
1	Right eye
24	Left eye inner corner
23	Left eye outer corner
38	Left eye lower line
35	Left eye upper line
29	Left eye left iris corner
30	Left eye right iris corner
25	Right eye inner corner
26	Right eye outer corner
41	Right eye lower line
40	Right eye upper line
33	Right eye left iris corner
34	Right eye right iris corner
13	Left eyebrow inner corner
16	Left eyebrow middle
12	Left eyebrow outer corner
14	Right eyebrow inner corner
17	Right eyebrow middle
54	Mouth top
55	Mouth bottom

Table 3
Some important facial features used for the experiment.

Features
Distance between right eye and left eye
Distance between the inner and the outer corner of the left eye
Distance between the upper and the lower line of the left eye
Distance between the left iris corner and the right iris corner of the left eye
Distance between the inner and the outer corner of the right eye
Distance between the upper and the lower line of the right eye
Distance between the left iris corner and the right iris corner of the right eye
Distance between the left eyebrow inner and the outer corner
Distance between the right eyebrow inner and the outer corner
Distance between top of the mouth and bottom of the mouth

Table 4
Features extracted using GAVAM from the facial features.

Features
The time of occurrence of the particular frame in milliseconds
The displacement of the face w.r.t. *X*-axis. It is measured by the displacement of the normal to the frontal view of the face in the X-direction
The displacement of the face w.r.t. *Y*-axis
The displacement of the face w.r.t. *Z*-axis
The angular displacement of the face w.r.t. *X*-axis. It is measured by the angular displacement of the normal to the frontal view of the face with the *X*-axis
The angular displacement of the face w.r.t. *Y*-axis
The angular displacement of the face w.r.t. *Z*-axis

6. Extracting features from audio data

We automatically extracted audio features from each annotated segment of the videos. Audio features were also extracted using a 30 Hz frame-rate and a sliding window of 100 ms. To compute the features, we used the open source software OpenEAR [57].

Specifically, this toolkit automatically extracts pitch and voice intensity. Z-standardization was used to perform voice normalization. The voice intensity was thresholded to identify samples with and without voice. Using openEAR we extracted 6373 features. These features includes several statistical measures, e.g., max and min values, standard deviation, and variance, of some key feature groups. Some of the useful key features extracted by openEAR are described below:

- *Mel frequency cepstral coefficients*: MFCC were calculated based on the short time Fourier transform (STFT). First, log-amplitude of the magnitude spectrum was taken, followed by grouping and smoothing the fast Fourier transform (FFT) bins according to the perceptually motivated Mel-frequency scaling. The Jaudio tool provided the first five of 13 coefficients, which were found to produce the best classification result.
- *Spectral centroid*: Spectral Centroid is the center of gravity of the magnitude spectrum of the STFT. Here, $M_i[n]$ denotes the magnitude of the Fourier transform at frequency bin n and frame i. The centroid is used to measure the spectral shape. A higher value of the centroid indicates brighter textures with greater frequency. The spectral centroid is calculated as follows:

$$C_i = \frac{\sum_{i=0}^{n} nM_i[n]}{\sum_{i=0}^{n} M_i[n]}$$

- *Spectral flux*: Spectral flux is defined as the squared difference between the normalized magnitudes of successive windows: $F_i = \sum_{n=1}^{n} (N_t[n]-N_{t-1}[n])^2$ where $N_t[n]$ and $N_{t--1}[n]$ are the normalized magnitudes of the Fourier transform at the current frame t and the previous frame $t-1$, respectively. The spectral flux represents the amount of local spectral change.
- *Beat histogram*: It is a histogram showing the relative strength of different rhythmic periodicities in a signal, and is calculated as the auto-correlation of the RMS.
- *Beat sum*: This feature is measured as the sum of all entries in the beat histogram. It is a very good measure of the importance of regular beats in a signal.
- *Strongest beat*: It is defined as the strongest beat in a signal, in beats per minute and is found by finding the strongest bin in the beat histogram.
- *Pause duration*: Pause direction is the percentage of time the speaker is silent in the audio segment.
- *Pitch*: This is computed using the standard deviation of the pitch level for a spoken segment.
- *Voice quality*: Harmonics to noise ratio in the audio signal.
- *PLP*: The Perceptual Linear Predictive Coefficients of the audio segment were calculated using the openEAR toolkit.

7. Sentiment analysis of textual data

Identifying sentiments in text is a challenging task, mainly because of the ambiguity of words in text, complexity of meaning, and the interplay of various factors such as irony, politeness, writing style, as well as the variability of language from person to person and from culture to culture. In this work, we followed the sentic computing paradigm [58], which considers text as expressing both semantics and sentics, i.e., denotative and connotative information commonly associated with real-world objects, actions, events, and people. As we conducted concept-level sentiment analysis, concept extraction from text was the fundamental step of the experiment. Below, we first describe the concept extraction algorithm from text [59], followed by a description of feature extraction methods based on the extracted concepts for concept-level sentiment analysis.

7.1. Subject noun rule

Trigger: when the active token was found to be the syntactic subject of a verb.
Behavior: if a word h was in a subject noun relationship with a word t then the concept $t-h$ was extracted.
Example: in (1), *movie* was in a subject relation with *boring*.

(1) The movie is boring.

Here the concept `(boring-movie)` was extracted.

7.2. Joint subject noun and adjective complement rule

Trigger: when the active token was found to be the syntactic subject of a verb and the verb was in an adjective complement relation with an adverb.
Behavior: if a word h was in a subject noun relationship with a word t, and the word t was in an adjective complement relationship with a word w, then the concept $w-h$ was extracted.
Example: in (2), *flower* was in a subject relation with *smells* and *smells* was in an adjective complement relationship with *bad*.

(2) The flower smells bad.

Here the concept `(bad-flower)` was extracted.

7.3. Direct nominal objects

This complex rule dealt with direct nominal objects of a verb.

Trigger: when the active token was head verb of a direct object dependency relation.
Behavior: if a word *h* was in a direct nominal object relationship with a word *t*, then the concept $h-t$ was extracted.
Example: in (3) the system extracted the concept `(see, movie)`.

(3) Paul saw the movie in 3D.

`(see,in,3D)` was not treated at this stage since it will later be treated using the standard rule for prepositional attachment.

7.4. Adjective and clausal complements Rules

These rules dealt with verbs which had as complements, either an adjective or a closed clause (i.e., a clause, usually finite, with its own subject).

Trigger: when the active token was head verb of one of the complement relations.
Behavior: if a word *h* was in a direct nominal object relationship with a word *t*, then the concept $h-t$ was extracted.
Example: in (4), *smells* was the head of a clausal complement dependency relation with *bad* as the dependent.

(4) This meal smells bad.

In this example, the concept `(smell,bad)` was extracted.

7.5. Negation

Negation is also a crucial component of natural language text which usually flips the meaning of a sentence. This rule was used to identify whether a word was negated.

Trigger: when in text, a word was negated.
Behavior: if a word *h* was negation by a *negation marker t*, then the concept $t-h$ was extracted.
Example: in (5), *like* was the head of the negation dependency relation with *not* as the dependent. Here, *like* was negated by the negation marker *not*.

(5) I do not like the movie.

According to the rule described above, the concept `(not, like)` was extracted.

7.6. Open clausal complements

Open clausal complements are clausal complements of a verb that do not have their own subject, meaning they (usually) share their subjects with that of the matrix clause. The corresponding rule was complex in the same way as the one for direct objects.

Trigger: when the active token was the head of the relation.
Behavior: as for the case of direct objects, the algorithm tried to determine the structure of the dependent of the head verb. Here the dependent was itself a verb, therefore, the system tried to establish whether the dependent verb had a direct object or a clausal complement of its own. In a nutshell, the system was dealing with three elements: the head verb(h), the dependent verb(d), and the (optional) complement of the dependent verb (t). Once these elements had all been identified, the concept (h,d,t) was extracted.
Example: in (6), *like* was the head of the *open clausal complements* dependency relation with *praise* as the dependent. The complement of the dependent verb *praise* was *movie*.

(6) Paul likes to praise good movies.

So, in this example, the concept `(like,praise,movie)` was extracted.

7.7. Modifiers

7.7.1. Adjectival, adverbial and participial modification

The rules for items modified by adjectives, adverbs or participles, all share the same format.

Trigger: these rules were activated when the active token was modified by an adjective, an adverb or a participle.
Behavior: if a word *w* was modified by a word *t* then the concept *(t,w)* was extracted.
Example: in (7) the concept *bad, loser* was extracted.

(7) Paul is a bad loser.

7.7.2. Prepositional phrases

Although prepositional phrases do not always act as modifiers, we introduced them in this section as the distinction does not affect their treatment.

Trigger: the rule was activated when the active token was recognized as typing a prepositional dependency relation. In this case, the head of the relation was the element to which the PP attached, and the dependent was the head of the phrase embedded in the PP.
Behavior: instead of looking for the complex concept formed by the head and dependent of the relation, the system used the preposition to build a ternary concept.
Example: in (8), the parser yields a dependency relation typed `prep_with` between the verb *hit* and the noun *hammer* (=the head of the phrase embedded in the PP).

(8) Bob hit Marie with a hammer.

Therefore, the system extracted the complex concept `(hit, with, hammer)`.

7.7.3. Adverbial clause modifier

This kind of dependency concerned full clauses that act as modifiers of a verb. Standard examples involved temporal clauses and conditional structures.

Trigger: the rule was activated when the active token was a verb modified by an adverbial clause. The dependent was the head of the modifying clause.
Behavior: if a word *t* was an adverbial clause modifier of a word *w*, then the concept $(t-w)$ was extracted.
Example: in (9), the complex concept `(play,slow)` was extracted.

(9) The machine slows down when the best games are playing.

7.7.4. Noun compound modifier

Trigger: the rule was activated when it finds a noun composed of several nouns. A noun compound modifier of an NP was any noun that served to modify the head noun.
Behavior: if a noun-word w was modified by another noun-word t then the complex concept $(t-h)$ was extracted.
Example: in (10), the complex concept `(birthday,party)` was extracted.

(10) Erik threw the birthday party for his girlfriend.

8. Fusion

This section discusses the feature-level fusion method for using information from textual, audio and visual modalities.

8.1. Feature-level fusion

Multimodal fusion is the heart of any multimodal sentiment analysis engine. As discussed in Section 3, there are two main fusion techniques: feature-level fusion and decision-level fusion. We implemented feature-level fusion by concatenating the feature vectors of all three modalities, to form a single long feature vector. This trivial method had the advantage of relative simplicity, yet was shown to produce significantly high accuracy. We concatenated the feature vector of each modality into a single feature vector stream. This feature vector was then used for classifying each video segment into sentiment classes. To estimate the accuracy, we used 10-fold cross validation.

8.2. Decision-level fusion

In decision-level fusion, we obtained feature vectors from the above-mentioned methods but instead of concatenating the feature vectors as in feature-level fusion, we used a separate classifier for each modality. The output of each classifier was treated as a classification score. In particular, we obtained a probability score for each sentiment class, from each classifier. In our case, as there are 3 sentiment classes, we obtained 3 probability scores from each modality. We then calculated the final label of the classification using a rule-based approach given below:

$$l' = \text{argmax}_i(q_1 s_i^a + q_2 s_i^v + q_3 s_i^t), \quad i = 1, 2, 3, \dots, C$$

where q_1, q_2 and q_3 represent weights for the three modalities. We adopted an equal-weighted scheme, so in our case $q_1 = q_2 = q_3 = 0.33$. C is the number of sentiment classes, and s_i^a, s_i^v and s_i^t denote the scores from audio, visual and textual modality, respectively.

9. Proof of concept

We developed a real-time multimodal sentiment analysis avatar based on the methods described above. The avatar allows a user to express his or her opinions in front of a camera. Later, it splits the video into several parts where each segment is empirically set to 5 s duration. The same methodology as described in Sections 5–8 was adopted to extract the sentiment from each segment. Fig. 1 shows a visualization of the avatar. A transcriber was used to obtain the text transcription of the audio.

Fig. 2 shows that our real-time multimodal sentiment analysis avatar analyzed a video and successfully detected its sentiment over time. The video related to a mobile and was collected from YouTube. Fig. 2 shows the sentiment of the first 11.5 s of the video detected by the avatar. In the initial 2 s, the reviewer expressed a positive sentiment about the product, followed by a negative sentiment from 2 to 4.4 s. This was followed by a positive review of the product expressed during the interval 4.4–8 s, and no sentiment expressed during the period 8–9.5 s. Finally, the reviewer expressed a positive sentiment about the product from 9.5 s till the end of the video.

10. Experimental results

In this section, we discuss the experimental results on the YouTube dataset and compare results using the approach proposed by [1]. Several supervised classifiers, namely Naïve Bayes, SVM, ELM, and Neural Networks, were employed on the fused feature vector to obtain the sentiment of each video segment. However, the best accuracy was obtained using the ELM classifier.

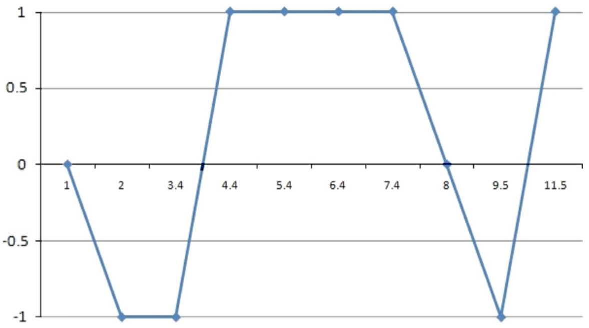

Fig. 2. Real-time multimodal sentiment analysis of a YouTube product review video.

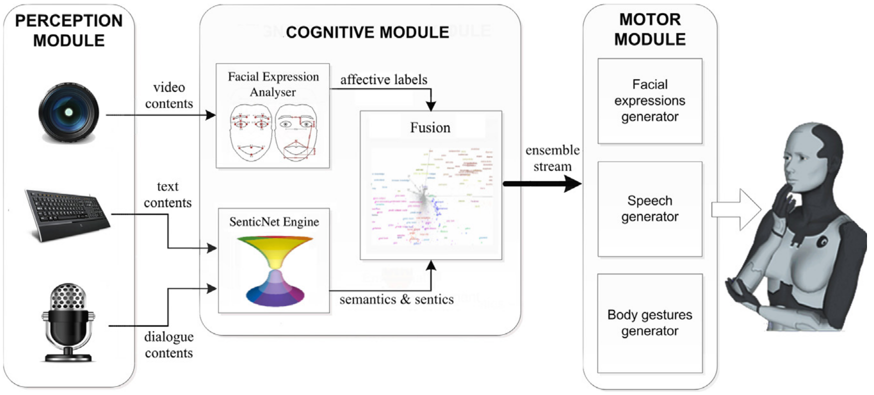

Fig. 1. Multimodal sentiment analysis avatar.

Table 5
Results of feature-level fusion.

Combination of modalities	Precision	Recall
Accuracy of the experiment carried out on Textual Modality	0.619	0.59
Accuracy of the experiment carried out on Audio Modality	0.652	0.671
Accuracy of the experiment carried out on Video Modality	0.681	0.676
Experiment using only visual and text-based features	0.7245	0.7185
Result obtained using visual and audio-based features	0.7321	0.7312
Result obtained using audio and text-based features	0.7115	0.7102
Accuracy of feature-level fusion of three modalities	0.782	0.771

Table 6
Results of decision-level fusion.

Combination of modalities	Precision	Recall
Experiment using only visual and text-based features	0.683	0.6815
Result obtained using visual and audio-based features	0.7121	0.701
Result obtained using audio and text-based features	0.664	0.659
Accuracy of decision-level fusion of three modalities	0.752	0.734

Results for feature-level fusion are shown in Table 5, from which it can be seen that our method outperforms [1] by 16.00% in terms of accuracy.

Table 6 shows the experimental results of decision-level fusion. Tables 5 and 6 show the experimental results obtained when only *audio and text*, *visual and text*, *audio and visual* modalities were used for the experiment. It is clear from these tables that the accuracy improves dramatically when audio, visual and textual modalities are used together. Finally, Table 5 also shows experimental results obtained when either the *visual* or *text* modality only, was used in the experiment.

10.1. Feature analysis

We also analyzed the importance of each feature used in the classification task. The best accuracy was obtained when all features were used together. However, GAVAM features were found to be superior in comparison to the features extracted by Luxand FSDK 1.7. Using only GAVAM features, we obtained an accuracy of 57.80% for the visual features-based sentiment analysis task. However, for the same task, 55.64% accuracy was obtained when we only used the features extracted by Luxand FSDK 1.7.

For the audio-based sentiment analysis task, MFCC and Spectral Centroid were found to produce a lower impact on the overall accuracy of the sentiment analysis system. However, exclusion of those features led to a degradation of accuracy for the audio-based sentiment analysis task. We also experimentally evaluated the role of certain audio features like *time domain zero crossing*, *root mean square*, *compactness*, but did not obtain a higher accuracy using any of them.

In the case of text-based sentiment analysis, we found that concept-gram features play a major role compared to SenticNet-based features. In particular, SenticNet-based features mainly helped detect associated sentiments in text in an unsupervised way. Note that our aim is to develop a multimodal sentiment analysis system where sentiment will be extracted from text in an unsupervised way using SenticNet as a knowledge base.

10.2. Performance comparison of different classifiers

On the same training and test sets, we ran the classification experiment using SVM, ANN and ELM. ELM outperformed ANN by 12% in terms of accuracy (see Table 7). However, we observed only a small difference in accuracy between the ELM and SVM classifiers.

Table 7
Comparison of classifiers.

Classifier	Recall (%)	Training time
SVM	77.03	2.7 min
ELM	77.10	25 s
ANN	57.81	2.9 min

In terms of training time, the ELM outperformed SVM and ANN by a huge margin (Table 7). As our eventual goal is to develop a real-time multimodal sentiment analysis engine, so we preferred the ELM as a classifier which provided the best performance in terms of both accuracy and training time.

11. Conclusion and future work

We have presented a multimodal sentiment analysis framework, which includes sets of relevant features for text and visual data, as well as a simple technique for fusing the features extracted from different modalities. In particular, our textual sentiment analysis module has been enriched by sentic-computing-based features, which have offered significant improvement in the performance of our textual sentiment analysis system. Visual features also play a key role to outperform the state-of-the-art.

As discussed in the literature, gaze- and smile-based facial expression features are usually found to be very useful for sentiment classification. Our future research will aim to incorporate gaze and smile features, for facial-expression-based sentiment classification, in addition to focusing on the use of audio modality for the multimodal sentiment analysis task. Furthermore, we will explore the possibility of developing a culture- and language-independent multimodal sentiment classification framework. Finally, we will strive to improve the decision-level fusion process using a cognitively-inspired fusion engine. Subsequently, we will work on reducing the time complexities of our developed methods, in order to get closer to the ambitious goal of developing a real-time system for multimodal sentiment analysis. Hence, another aspect of our future work will be to effectively analyze and appropriately address the system's time complexity requirements in order to create an efficient and reliable multimodal sentiment analysis engine.

References

[1] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: harvesting opinions from the web, In: Proceedings of the 13th International Conference on Multimodal Interfaces, ACM, Alicante, Spain, 2011, pp. 169–176.
[2] E. Cambria, N. Howard, J. Hsu, A. Hussain, Sentic blending: scalable multimodal fusion for continuous interpretation of semantics and sentics, In: IEEE SSCI, Singapore, 2013, pp. 108–117.
[3] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1) (2006) 489–501.
[4] G.-B. Huang, E. Cambria, K.-A. Toh, B. Widrow, Z. Xu, New trends of learning in computational intelligence, IEEE Comput. Intelli. Mag. 10 (2) (2015) 16–17.
[5] G. Huang, G.-B. Huang, S. Song, K. You, Trends in extreme learning machines: a review, Neural Netw. 61 (2015) 32–48.
[6] S. Decherchi, P. Gastaldo, R. Zunino, E. Cambria, J. Redi, Circular-ELM for the reduced-reference assessment of perceived image quality, Neurocomputing 102 (2013) 78–89.
[7] E. Cambria, P. Gastaldo, F. Bisio, R. Zunino, An ELM-based model for affective analogical reasoning, Neurocomputing 149 (2015) 443–455.
[8] E. Principi, S. Squartini, E. Cambria, F. Piazza, Acoustic template-matching for automatic emergency state detection: an ELM based algorithm, Neurocomputing 149 (2015) 426–434.
[9] S. Poria, E. Cambria, A. Hussain, G.-B. Huang, Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.

[10] H. Qi, X. Wang, S.S. Iyengar, K. Chakrabarty, Multisensor data fusion in distributed sensor networks using mobile agents, In: Proceedings of 5th International Conference on Information Fusion, 2001, pp. 11–16.

[11] Ekman, Paul, Friesen, Wallace V, O'Sullivan, Maureen, Chan, Anthony, Diacoyanni-Tarlatzis, Irene, Heider, Karl, Krause, Rainer, LeCompte, William Ayhan and Pitcairn, Tom and Ricci-Bitti, Pio E and others, Universals and cultural differences in the judgments of facial expressions of emotion. J. person. soc. psychol. 53 (4) (1987) 712–717.

[12] D. Matsumoto, More evidence for the universality of a contempt expression, Motiv. Emot. 16 (4) (1992) 363–368.

[13] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, 1978.

[14] W.V. Friesen, P. Ekman, Emfacs-7: Emotional Facial Action Coding System, Unpublished manuscript, University of California at San Francisco 2.

[15] A. Lanitis, C.J. Taylor, T.F. Cootes, A unified approach to coding and interpreting face images, In: Fifth International Conference on Computer Vision, 1995. Proceedings, IEEE, Cambridge, Massachusetts, USA, 1995, pp. 368–373.

[16] D. Datcu, L. Rothkrantz, Semantic audio–visual data fusion for automatic emotion recognition, In: Euromedia, Citeseer, 2008.

[17] M. Kenji, Recognition of facial expression from optical flow, IEICE Trans. Inf. Syst. 74 (10) (1991) 3474–3483.

[18] N. Ueki, S. Morishima, H. Yamada, H. Harashima, Expression analysis/synthesis system based on emotion space constructed by multilayered neural network, Syst. Comput. Jpn. 25 (13) (1994) 95–107.

[19] L.S.-H. Chen, Joint processing of audio–visual information for the recognition of emotional expressions in human–computer interaction (Ph.D. thesis), Citeseer, 2000.

[20] I.R. Murray, J.L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, J. Acoust. Soc. Am. 93 (2) (1993) 1097–1108.

[21] R. Cowie, E. Douglas-Cowie, Automatic statistical analysis of the signal and prosodic signs of emotion in speech, In: Fourth International Conference on Spoken Language, 1996. ICSLP 96, Proceedings, vol. 3, IEEE, Philadelphia, PA, USA, 1996, pp. 1989–1992.

[22] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, In: Fourth International Conference on Spoken Language, 1996, ICSLP 96, Proceedings, vol. 3, IEEE, Philadelphia, PA, USA, 1996, pp. 1970–1973.

[23] T. Johnstone, Emotional speech elicited using computer games, In: Fourth International Conference on Spoken Language, 1996, ICSLP 96, Proceedings, vol. 3, IEEE, Philadelphia, PA, USA, 1996, pp. 1985–1988.

[24] E. Navas, I. Hernáez, I. Luengo, 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. 14 (4) (2006) 1117–1127.

[25] H. Atassi, A. Esposito, A speaker independent approach to the classification of emotional vocal expressions, In: ICTAI, 2008, pp. 147–150.

[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, In: Interspeech, 2005, pp. 1517–1520.

[27] P. Pudil, F. Ferri, J. Novovicova, J. Kittler, Floating search methods for feature selection with nonmonotonic criterion functions, In: IAPR, 1994, pp. 279–283.

[28] K.R. Scherer, Adding the affective dimension: a new look in speech analysis and synthesis, In: ICSLP, 1996, pp. 1808–1811.

[29] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, K. Karpouzis, Multimodal emotion recognition from expressive faces, body gestures and speech, In: Artificial Intelligence and Innovations 2007: From Theory to Applications, 2007, pp. 375–388.

[30] J. Wiebe, Learning subjective adjectives from corpora, In: AAAI/IAAI, 2000, pp. 735–740.

[31] P.D. Turney, Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[32] E. Cambria, D. Olsher, D. Rajagopal, SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis, In: AAAI, Quebec City, 2014, pp. 1515–1521.

[33] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, In: HLT/EMNLP, Vancouver, BC, Canada, 2005, pp. 347–354.

[34] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, In: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2003, pp. 105–112.

[35] B. Pang, L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, In: ACL, Barcelona, 2004, pp. 271–278.

[36] C. Strapparava, A. Valitutti, Wordnet affect: an affective extension of wordnet., In: LREC, vol. 4, 2004, pp. 1083–1086.

[37] C.O. Alm, D. Roth, R. Sproat, Emotions from text: machine learning for text-based emotion prediction, In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 579–586.

[38] G. Mishne, Experiments with mood classification in blog posts, In: Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access, vol. 19, 2005.

[39] R. Xia, C. Zong, X. Hu, E. Cambria, Feature ensemble plus sample selection: domain adaptation for sentiment classification (extended abstract), In: IJCAI, Buenos Aires, 2015, pp. 4229-4233.

[40] C. Yang, K.H.-Y. Lin, H.-H. Chen, Building emotion lexicon from weblog corpora, In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, 2007, pp. 133–136.

[41] F.-R. Chaumartin, Upar7: a knowledge-based system for headline sentiment tagging, In: Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics, 2007, pp. 422–425.

[42] A. Esuli, F. Sebastiani, Sentiwordnet: a publicly available lexical resource for opinion mining, In: Proceedings of LREC, vol. 6, 2006, pp. 417–422.

[43] K.H.-Y. Lin, C. Yang, H.-H. Chen, What emotions do news articles trigger in their readers?, In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Glasgow, UK, 2007, pp. 733–734.

[44] A. Pak, P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, In: LREC, 2010, pp. 1320–1326.

[45] M. Hu, B. Liu, Mining and summarizing customer reviews, In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Seattle, Washington, USA, 2004, pp. 168–177.

[46] A. Balahur, J.M. Hermida, A. Montoyo, Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model, IEEE Trans. Affect. Comput. 3 (1) (2012) 88–101.

[47] C. Shan, S. Gong, P.W. McOwan, Beyond facial expressions: learning human emotion from body gestures, In: BMVC, 2007, pp. 1–10.

[48] M. Mansoorizadeh, N.M. Charkari, Multimodal information fusion application to human emotion recognition from face and speech, Multimed. Tools Appl. 49 (2) (2010) 277–297.

[49] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, S. Levinson, Audio–visual affect recognition, IEEE Trans. Multimed. 9 (2) (2007) 424–428.

[50] H. Gunes, M. Piccardi, Bi-modal emotion recognition from expressive face and body gestures, J. Netw. Comput. Appl. 30 (4) (2007) 1334–1345.

[51] T. Pun, T.I. Alecu, G. Chanel, J. Kronegg, S. Voloshynovskiy, Brain–computer interaction research at the computer vision and multimedia laboratory, University of Geneva, IEEE Trans. Neural Syst. Rehabil. Eng. 14 (2) (2006) 210–213.

[52] F. Wallhoff, Facial Expressions and Emotion Database, Technische Universität München.

[53] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, In: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, IEEE, Amsterdam, The Netherlands, 2005, pp. 5-9.

[54] M. Paleari, B. Huet, Toward emotion indexing of multimedia excerpts, In: International Workshop on Content-Based Multimedia Indexing, 2008, CBMI 2008, IEEE, London, UK, 2008, pp. 425–432.

[55] S. Poria, A. Gelbukh, E. Cambria, A. Hussain, G.-B. Huang, EmoSenticSpace: a novel framework for affective common-sense reasoning, Knowl. Based Syst. 69 (2014) 108–123.

[56] J.M. Saragih, S. Lucey, J.F. Cohn, Face alignment through subspace constrained mean-shifts, In: 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, Japan, 2009, pp. 1034–1041.

[57] F. Eyben, M. Wollmer, B. Schuller, Openear—introducing the munich open-source emotion and affect recognition toolkit, In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009, IEEE, Amsterdam, Netherlands, 2009, pp. 1–6.

[58] E. Cambria, A. Hussain, C. Havasi, C. Eckl. Sentic computing: exploitation of common sense for the development of emotion-sensitive systems. In: LNCS, vol. 5967, Springer, pp. 148–156, 2010.

[59] E. Cambria, A. Hussain, Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis, Springer, Cham, Switzerland, 2015.

Soujanya Poria received his BEng in Computer Science from Jadavpur University, India in 2013. He then joined Nanyang Technological University as a research engineer in the School of Electrical and Electronics Engineering and, later in 2015, he joined NTU Temasek Labs, where he is conducting research on sentiment analysis in multiple domains and different modalities. Since February 2014, Soujanya has also started his PhD studies at the University of Stirling (Computing Science and Mathematics). His research areas include natural language processing, opinion mining, cognitive science and multimodal sentiment analysis. In 2013, Soujanya received the best undergraduate thesis and researcher award from Jadavpur University. He was awarded Gold Plated Silver medal from the University and Tata Consultancy Service for his final year project during his undergraduate course. He is also a fellow of the Brain Sciences Foundation and a program committee member of SENTIRE, the IEEE ICDM workshop series on sentiment analysis.

Erik Cambria received his BEng and MEng with honors in Electronic Engineering from the University of Genoa in 2005 and 2008, respectively. In 2012, he was awarded his PhD in Computing Science and Mathematics following the completion of an EPSRC project in collaboration with the MIT Media Lab, which was selected as impact case study by the University of Stirling for the Research Excellence Framework (REF2014). After two long-term research visits at HP Labs India and Microsoft Research Asia, Dr Cambria worked as Lead Investigator of the Cognitive Science Programme at the National University of Singapore (Temasek Labs). Since 2014, he is an Assistant Professor at Nanyang Technological University (School of Computer Engineering), where he teaches and conducts research on natural language processing and information retrieval. Dr Cambria is Co-Editor in Chief of the Socio-Affective Computing Series and Associate Editor of Knowledge-Based Systems, Artificial Intelligence Review, and Cognitive Computation. He is recipient of several awards, e.g.,the Temasek Research Fellowship, and is involved in many international conferences as Workshop Organizer, e.g., ICDM SENTIRE, Track Chair, e.g., FLAIRS AI4BigData, PC Member, e.g., AAAI, and Keynote Speaker, e.g., CICLing.

Newton Howard's passion for science and technology began during his childhood. He pursued his interests in his studies and in 2000 while a graduate member of the Department of Mathematical Sciences at the University of Oxford, he proposed the Theory of Intention Awareness (IA). In 2002, he received a second doctoral degree in cognitive informatics and mathematics from the prestigious La Sorbonne in France. In 2007 he was awarded the habilitation a diriger des recherches (HDR) for his leading work on the Physics of Cognition (PoC) and its applications to complex medical, economical, and security equilibriums. Recently in 2014 he received his doctorate of philosophy from the University of Oxford specifically focusing on “The Brain Code” for work in neurodegenerative diseases. His work has made a significant impact on the design of command and control systems as well as information exchange systems used at tactical, operational and strategic levels. As the creator of IA, Dr. Howard was able to develop operational systems for military and law enforcement projects. These utilize an intent-centric approach to inform decision-making and ensure secure information sharing. His work has brought him into various academic and government projects of significant magnitude, which focus on science and the technological transfer to industry. While Dr. Howard's career formed in military scientific research, in 2002 he founded the Center for Advanced Defense Studies (CADS) a leading Washington, D.C, national security group. Currently, Dr. Howard serves as the Director of the Board. He also is a national security advisor to several U.S. Government organizations.

Guangbin Huang received the B.Sc degree in applied mathematics and M.Eng degree in computer engineering from Northeastern University, PR China, in 1991 and 1994, respectively, and Ph.D degree in electrical engineering from Nanyang Technological University, Singapore in 1999. During undergraduate period, he also concurrently studied in Wireless Communication department of Northeastern University, P. R. China. From June 1998 to May 2001, he worked as a Research Fellow in Singapore Institute of Manufacturing Technology (formerly known as Gintic Institute of Manufacturing Technology) where he has led/implemented several key industrial projects and also built up two R&D labs: Communication Information Technologies Lab and Mobile Communication Lab. He was the chief architect for several significant industrial projects including (Singapore Airlines) SATS Cargo Terminal 5 Information Tracking System. From May 2001, he has been working as an Assistant Professor and Associate Professor (tenured) in the School of Electrical and Electronic Engineering, Nanyang Technological University. He was a member of the Emergent Technologies Technical Committee of IEEE Computational Intelligence Society. He is a member of the Committee on Membership Development of IEEE Singapore Chapter. He serves as session chair, track chair and plenary talk chair for several international conferences. He is currently an associate editor of Neurocomputing, and IEEE Transactions on Systems, Man, and Cybernetics – Part B. He is a senior member of IEEE.

Amir Hussain obtained his BEng (with the highest 1st Class Honors) and PhD (in novel neural network architectures and algorithms) from the University of Strathclyde in Glasgow, Scotland, UK, in 1992 and 1997 respectively. He is currently a Professor of Computing Science, and founding Director of the Cognitive Signal Image and Control Processing Research (COSIPRA) Laboratory at the University of Stirling in Scotland, UK. His research interests are inter-disciplinary and industry focussed, and include multi-modal cognitive and sentic computing techniques and applications. He has published over 270 papers, including over a dozen books and 80 journal papers. He is the founding Editor-in-Chief of the journals: Cognitive Computation (Springer Neuroscience, USA), and Big Data Analytics (BioMed Central), and Chief-Editor of the Springer Book Series on Socio-Affective Computing, and SpringerBriefs on Cognitive Computation. He is an Associate Editor of the IEEE Transactions on Neural Networks and Learning Systems, a member of several Technical Committees of the IEEE Computational Intelligence Society (CIS), founding publications co-Chair of the IINNS Big Data Section and its annual INNS Conference on Big Data, and Chapter Chair of the IEEE UK and RI Industry Applications Society.