Journal of Computer Science

Review

Multimodal Sentiment Analysis: A Comparison Study

Intisar O. Hussien and Yahia Hasan Jazyah

ITC, Arab Open University, Kuwait

Article history
Received: 19-03-2018
Revised: 17-05-2018
Accepted: 09-06-2018

Corresponding Author:
Yahia Hasan Jazyah
ITC, Arab Open University,
Kuwait
yahia@aou.edu.kw

Abstract: Sentiments and emotions play a pivotal role in our daily lives. They assist decision making, learning, communication and situation awareness in human environments. Sentiment analysis is mainly focused on the automatic recognition of opinions' polarity, as positive or negative. Nowadays, sentiment analysis is replacing the old web based survey and traditional survey methods that conducted by deferent companies for finding public opinion about entities like products and services in order to improve their marketing strategy and product of advertisement, at the same time sentiment analysis improves customer service. Large number of videos is being uploaded online every day. Video files contain text, visual and audio features that complement each other. Multimodality is defined by analyzing more than one modality, Multimodal Sentiment Analysis refers to the combination of two or more input models in order to improve the performance of the analysis; a combination of text and audio-visual inputs is an example. The automatic analysis of multimodal opinion involves a deep understanding of natural languages, audio and video processing, whereas researchers are continuing to improve them. This paper focuses on multimodal sentiment analysis as text, audio and video, by giving a complete image of it and related dataset available and providing brief details for each type, in addition to that present the recent trend of researches in the multimodal sentiment analysis and its related fields will be explored.

Keywords: Sentiment Analysis, Multimodal, Affecting Analysis and Text, Audi and Visual Information

Introduction

Several scientific fields have a great interest of research because of their effective recognition and classification practically and theoretically, such as machine learning, signal processing, computer vision, computational linguistics, cognitive, social psychology and neuroscience. Such scientific fields are an emerging research area today (Picard, 2010) and (D'mello and Kory, 2015). People are now extensively using the social media environment, such as YouTube, Facebook, Blog and Microblog in order to express their opinions. People are increasingly making use of images, audio and videos on different social media platforms to disclose and express their opinions. Thus, it is highly crucial to mine opinions and point out sentiments from the various modalities (Cambria *et al*., 2014) and (KgaogeloLetsebe, 2017). At the same time, there is an increasing need to know not only what information a user conveys but also how it is being conveyed as described in Data Analysis. Many researches done by psychologists and neuroscientists have shown the effect of an emotion that plays a significant role in the rational actions of human beings as it is closely related to decision making (Damasio, 1994).

Many of the works, to date, on sentiment analysis have focused on textual data and number of resources have been created include the use of lexicons (Liu and Zhang, 2012) and (Pang *et al*., 2002), but very little of the literature examines the vocal correlates and other relevant aspects of emotion effects in human speech and video. A recent and modern development of multimodal sentiment analysis is the visual sentiment analysis. Users of social media frequently share their text messages along with images and videos, those kinds of visual multimedia are an additional source of information in expressing users' sentiment (Fig. 1).

DOI: 10.3844/jcssp.2018.804.818

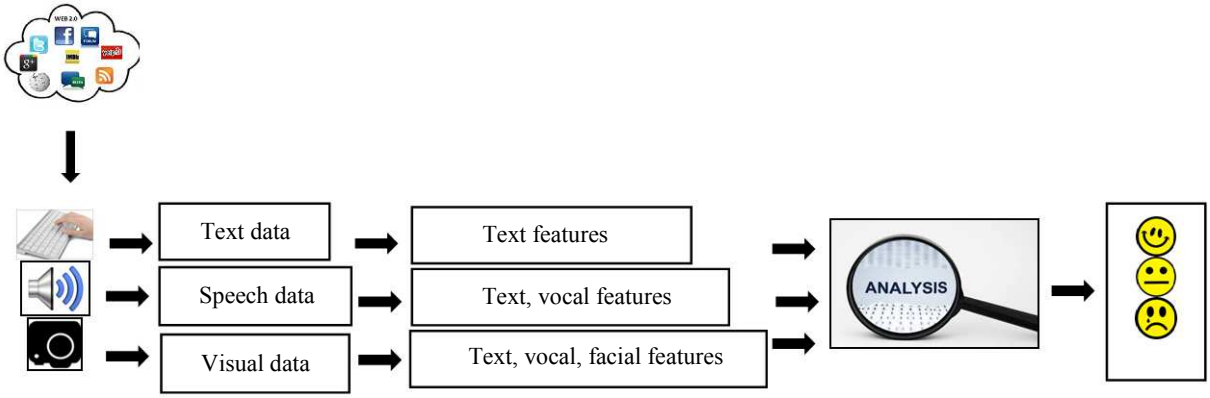

Fig. 1: General sentiment analysis

The aim of multimodal analysis is to increase the accuracy and achieve the best prediction. This research presents a detailed discussion of the literature, which describing human text opinion, vocal emotion, visual express and its principal findings. The parameters of voice and video which affected by emotion, will be described in details for a domain of specific emotions.

The remaining of paper is organized as follows: Part II presents the techniques and features of multimodal sentiment analysis, part III shows the multimodal sentiment analysis in more details, part IV presents the datasets of multimodal sentiment, part V presents the applications of sentimental analysis, part VI presents sentimental analysis approaches of challenge and gap and finally, part VII is the conclusion.

Multimodal Sentiment Analysis Techniques and Features

A. Multimodal Sentiment Analysis Technique

Multimodal fusion is the process of combining data collected from various modalities for analysis tasks. Multimodal fusion is a profusion information gain when using the fusion of multimodal with a better accuracy of the overall results, which helps to take a decision.

Three main levels (types) of fusion have been studied by researchers: Feature-level fusion (early fusion), decision-level fusion (late fusion) and hybrid fusion approach. Furthermore, there are many different multimodal fusion models such as Model-level fusion, Rule base model, Estimation based and Classification based methods.

Feature Level (Early Fusion) (Rosas et al., 2013)

Features of text, audio and video are extracted from various modalities separately. First, the previous features are treated as a general feature vector and then combined features are presented for analysis and classification. The advantage of this approach is that the relations of various multimodal features are completed at the beginning as an early stage, which may provide better achievement. The disadvantage of this approach is concurrence timing, as the obtained features belong to various modalities and can vary in many aspects; however, all features must be imported into the same format before the fusion analysis takes place.

Decision Level (Late Fusion) (Celli et al., 2014)

The features of each modality are extracted, analyzed and classified separately and all the results of analysis are merged in to a vector in order to obtain the final decision. The advantage of this approach is that it is easily compared to Feature Level fusion; since the fusion of decisions are gained from various modalities and have same form of data. One more advantage is that, no need for converting data to the same format; since every modality has used its best learning model and the most appropriate classifier for its features, however this could be considered as disadvantage for analysis because more than one classifier is used and mire-learning process for each model becomes uninteresting, boring and time consuming.

Hybrid Multimodal Fusion (Wöllmer et al., 2013)

It is a combination of both feature-level and decision-level fusion methods. It aims to obtain the advantages of both feature and decision level fusion approaches and overcomes the disadvantages of both.

B. Multimodal Sentiment Analysis Features

Multimodal seminal analysis features are a compilation of two or more different features, text, audio and image. In textual opinions, the only available source of information consists of the words in the opinion and the dependencies among them, which may sometime insufficient to convey the exact sentiment of the consumer. Instead, video opinions provide multimodal data in the form of vocal as well as visual responses. The vocal modulations in the recorded response help us

determine the tone of the speaker, whereas visual data can provide information regarding the emotional state of the speaker. So, a combination of text and video (with audio) data can help create a better sentiment analysis model.

Linguistic Features

Sentimental analysis gained from textual aims at the extraction of appraising meaning which starts by automatic detection of the state's subjective. Here, an overview is provided about the sentiment analysis approaches inNLP (Natural Language Processing), including supervised and unsupervised methods and future directions and limitations in the field (Table 1).

Feature types can be explicit or implicit; the explicit has four feature types: Syntactic, semantic, link-based and stylistic features, while the implicit focuses on semantic and linguistic rules (Sharef *et al*., 2016).

In natural language processing, Sentiment Analysis refers to find whether sentiment of a text which is written in natural language is positive, neural or negative (Obaidat *et al*., 2015), this can be achieved using supervised "Corpus-based" sentiment analysis approach, which relies on manually labeled samples, unsupervised "Knowledge or Lexicon-based" sentiment analysis approach, or hybrid (both lexicon-based and corpus-based) approach.

Supervised sentiment analysis (corpus based) aims at building predictive models for sentiment based by exploiting machine learning classifier that is trained on a labeled data, which in turn, test data based on it. This approach builds a feature vector of each text entry in which certain aspects or word frequencies are quantified and then, training the standard machine learning tools and validating them against reference annotated texts. Wiebe *et al*. (1999) use annotations that tag the evaluated content, but not its orientation, as a first approach in order to supervise the sentiment analysis, whereas the result of text classification is either subjective or objective. However, most of supervised approaches related to sentiment analysis are trained on specific domain and require a huge annotated corpus that manually labeled, this process is expensive and time consuming.

Unsupervised (lexicon based) approaches allow an estimation that is based on expert knowledge without the need to annotated data. The expert knowledge, which is used for the estimation, is often encoded in a lexicon, in which words or phrases are annotated with their sentimental meanings. These lexica can be manually annotated by means of raters who can interpret the meaning of words. Stone (1997) The General Inquirer (GI) is the one of the most widely used reference lexica for sentimental analysis, which includes a list of positive and negative terms. Wilson *et al*. (2005) define a method to increase the recall of unsupervised techniques that combines GI with other lexica (Wilson *et al*., 2005). (Pennebaker *et al*., 2015a) a new Linguistic Inquiry and Word Count (LIWC) are applied as a new method that counts the positive and negative affecting terms in text. On information retrieval, De Luca and Nürnberger (2006a) implemented methods, based on relations, to merge SynSets with hypernyms, hyponyms and context information.

Table 1: Recent papers on multimodal analysis

References	Language	Datasets	Modality: T/A/V	Feature Fusion/ Decision fusion	Features
Yamasaki *et al*. (2015)	English	1646 TED talk videos for 14 different tags	Text-Audio	Decision fusion	Linguistic, and Acoustic features
Rosas *et al*. (2013)	Spanish	Spanish videos + English videos	Text-Audio-Video	Feature level fusion	Linguistic, Acoustic. And Visual features
Lee and Narayanan (2005)	English	1187 calls of six 7200 ultrances real spoken dialog telephone machine agent	Audio	Feature level fusion	Acoustic, lexical, and discourse
Wöllmer *et al*. (2013)	English	Real YouTube dataset	Text-Audio-Video	Hybrid fusion	Polarized words, smile, gaze, pause, and voice
Rosas *et al*. (2013)	Spanish	MOUD	Text-Audio-Video	Feature level fusion	Acoustic, Linguistic, And Visual features
Morency *et al*. (2011)	Spanish	Spanish videos	Text-Audio-Video	Feature level fusion	Acoustic, Linguistic, And Visual features
DeVault *et al*. (2014)	English	351 video dataset (132F, 217M)	Text-Audio-Video	Multi Framework	Smile, 3D head orientation, intensity,speaking fraction, dynamic, speech dynamic, and gaze
Alam and Riccardi (2014)	English	404 YouTube vloggers/ YouTube personality dataset (210F, 194M)	Text-Audio-Video	Dicesion fusion	A-V features, POS psycholiguistic, lexical traits and emotional
Sarkar *et al*. (2014)	English	404 YouTube vloggers/YouTube personality dataset (210F, 194M)	Text-Audio-Video	-	A-V features, text, sentiment and demogaphic
Poria *et al*. (2017)	English	421SEAR, CK & eNTERFACE dataset	Text-Audio-Video	Dicesion fusion	66 FCP, Luxand and JAudio software, BOC, Sentic features and Negation
Poria *et al*. (2016)	English	47 YouTube dataset (20F, 27M)	Text-Audio-Video	Feature/ Decisoin fusion	Softwares Luxand FSDK 1.7,GAVAM, openEAR and Concept-gram and Sentic Net-based features
Poria *et al*. (2017)	English	47 YouTube dataset (20F, 27M)	Text-Audio-Video	Decisoin fusion	Softwares using CLM-Z and GAVAM, openEAR and using CNN
Siddiquie *et al*. (2015)	English	230 viddeos/ Rallying a Crowd (RAC) dataset	Text-Audio-Video	Feature/ Decisoin fusion	Softwares using CAFFEE and features, prosody, MFCC, or spectrogram and using SATSVM and DCM

Bradley and Lang (1999) and Dodds and Danforth (2010) Affecting Norms of English Words (ANEW) originally were not designed for sentimental analysis, but they are benefit in classified sentiment such as happiness motion on the text. However, there is a difficulty in lexicon construction. In order to make it deals with a variability of languages; it is very expensive if it is done manually and not reliable if it is done automatically. On the other hand, the semi supervised approach is a class of supervised approach that makes use of unlabeled data for training. It minimizes the cost associated with labeling process. However, it highly depends on performance of initial labeled set that is classified.

Sentiment analysis can be grouped into three different levels based on the target of study: Document level; the entire document is classified either positive or negative using machine learning approach or lexicon based approach. Sentences level; evaluation of opinion is done sentence by sentence in order to decide whether it is positive, neutral or negative. But the drawback of both levels, they provide high level of classification, the researches illustrate the previous described levels are (Abdul-Mageed *et al*., 2014; Abdul-Mageed and Diab, 2014), (Ibrahim *et al*., 2015), (Duwairi *et al*., 2014), (Al-Ayyoub *et al*., 2015), (Salameh *et al*., 2015), (Duwairi, 2015), (Al-Kabi *et al*., 2016), (ElSahar and El-Beltagy, 2015), (Wang *et al*., 2015) and (Ghareb *et al*., 2015). While Poria *et al*. (2014) enable the ability to distinguish the polarity of each aspect, it clarifies each aspect whether it is positive or negative.

Sentiment analysis systems are categorized into statistics-based and knowledge-based (Pang *et al*., 2002). Initially, the use of knowledge bases was popular for the identification of emotions and polarity in text, after that, the supervised statistical methods become in common by most researchers.

Pang *et al*. (2002), apply and compare the performance of a review dataset in different machine learning algorithms by using large textual features with accuracy 82.9%, where Socher *et al*. (2013), use a Recursive Neural Tensor Network (RNTN) and obtain an accuracy 85% using the same dataset.

Another approach by Yu and Hatzivassiloglou (2003) uses semantic orientation of words to identify polarity at sentence level. Melville *et al*. (2009) develop a framework that exploits word-class association information for domain dependent sentiment analysis. Other unsupervised or knowledge-based approaches to sentiment analysis include: Turney (2002) uses seed words to calculate polarity and semantic orientation of phrases; Hu *et al*. (2013) propose a mathematical model to extract emotional clues from blogs and used them for sentiment recognition; Gangemi *et al*. (2014) present an unsupervised frame-based approach to identify opinion holders and topics; and Sentic Computing, Cambria and Hussain (2015) use a hybrid approach for sentiment analysis that exploits an ensemble of deep learning, commonsense reasoning and linguistics to better grasp semantics and Sentics (i.e., denotative and connotative information) associated with natural language concepts (Fig. 2). Narr *et al*. (2011) apply language processing approach (NLP) to extract information from tweets and transform them into a semantic knowledge base.

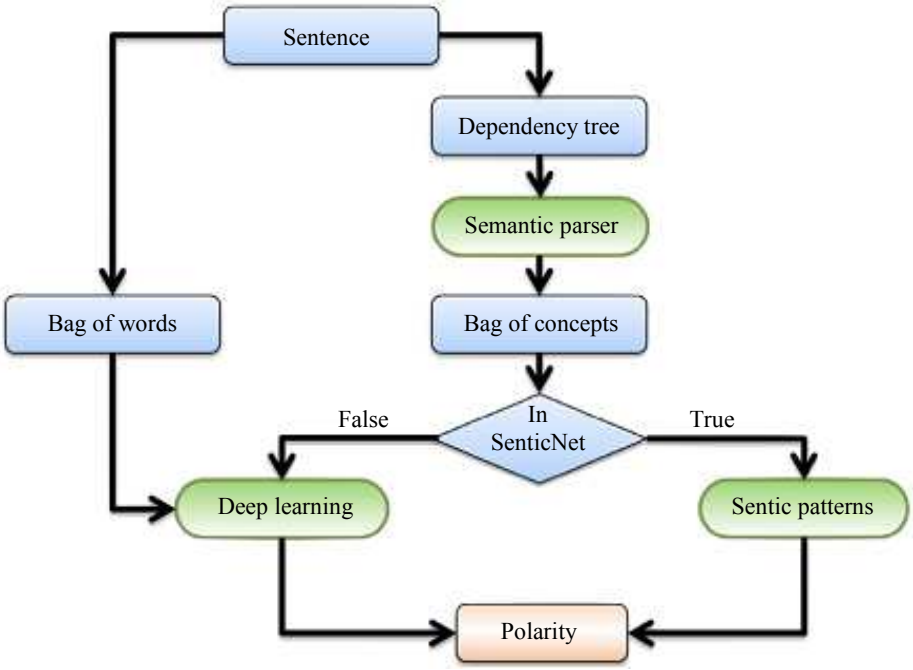

Fig. 2: Sentic computing framework (Poria *et al*., 2017)

Audio Features

Today, a rich body of literature has been established, including many surveys such as (Schuller *et al*., 2011), (Crouch and Khosla, 2012) and (Pérez-Rosas and Mihalcea, 2013). There is specific components feature for emotion and sentiment analysis through audio. Various prosodic and acoustic features have been used in the literature to learn how machines detect emotions (Navas *et al*., 2006), (Morrison *et al*., 2007), (Wu and Liang, 2011), (Murray and Arnott, 1993), (Luengo *et al*., 2005) and (Koolagudi *et al*., 2011). In psychological studies related to emotion, it is found that vocal parameters, especially pitch, intensity, speaking rate and voice quality play an important role in recognition of emotion and sentiment analysis (Murray and Arnott, 1993). There are different voice parameters which are affected by emotion such as Voice Quality, Utterance timing and Utterance pitch contour.

Many further works done and focus on sentiment analysis from the textual content as present in the speech, (Pereira *et al*., 2014) they used sentiment analysis in speech for information retrieval, their proposed approach takes a spoken query and retrieves documents. (Kaushik *et al*., 2103a) and its extension (Kaushik *et al*., 2103b) observe that sentiment analysis on natural speech data can be understand clearly even when faced with low word recognition rates and this is same as what proposed by (Metze *et al*., 2010). Pérez-Rosas and Mihalcea (2013) using speech recognition and focus on the linguistics reviews too.

Further studies show that not only the acoustic parameters are changing and depending on personality traits, but also through oral variations. Many researches are performed based on the types of features that are needed for better analysis (Muda *et al*., 2010). Researchers find that pitch and energy related features are playing a key role in affecting recognition. Other features that have been used by some researchers for feature extraction, which are affected by emotion, include Mel Frequency Cepstral Coefficients (MFCC), Log Frequency Power Coefficients (LFPC), Linear Prediction Cepstral Coefficients (LPCC), pause, teager energy operated based features and formants. Some of the effected audio features are mentioned briefly below:

- **Pitch**: It computes the standard deviation of the pitch level for the video. It represents the variation of voice intonation during the entire video
- **Intensity**: It measures the sound power of the spoken utterances in the video; the average voice intensity is computed over the whole video
- **Loudness**: It determines the perceived strength of the voice which is factored by the ear's sensitivity; the average loudness measure is computed over the entire video
- **Pause duration**: It calculates the number of audio samples that are identified as silence when the audio frames are extracted from the entire video. This feature can be interpreted as the percentage of the time when the speaker is silent
- **MFCC (Mel Frequency Cepstral Coefficients)**: One of the most commonly feature extraction method used in ASR, MFCCs are coefficients that collectively form a mel-frequency spectrum (MFC). The MFC computes the power of each frequency band of an audio clip
- **Spectral centroid**: One of the measures which is used in DSP in order to characterize spectrum, it indicates the center of mass of the spectrum, it provides an indication of the brightness of sound
- **Spectral flux**: It is a measure of how quickly the power spectrum of a signal is changing. This feature is usually calculated by comparing the power spectrum of one frame against the power spectrum of previous one, it is calculated by measuring the Euclidean distance between two normalized spectra.
- **Beat Histogram**: It shows the distribution of various beats by diagramming the strength of different rhythmic periodicities in a signal. It calculates the FFT from the output of the RMS of 256 windows

An acoustic study of emotions expressed in speech (Yildirim *et al*., 2004) investigates acoustic properties of speech associated with four different emotions (sadness, anger, happiness and neutral); they are intentionally expressed in speech by an actress. They aimed to obtain detailed acoustic knowledge and how speech is modulated when speaker's emotion changes from neutral to a certain emotional state. They experiment show happiness, anger, neutral and sadness share similar acoustic properties in a specific speaker. Speech associated with anger and happiness are characterized by longer utterance duration, shorter inter-word silence, higher pitch and energy values with wider ranges, showing the characteristics of exaggerated or hyper articulated speech, however this means that acoustic reparability is relatively poor.

Visual Features

Visual language is a type of non-verbal communication in which physical behavior, as opposed to words, are used to express or convey information. Such behavior includes facial expressions, body posture, gestures, eye movement, touch and the use of space.

Processing sentiment analysis using computer vision is a relatively recent area of research. The main research tasks in visual sentiment analysis is focusing on detecting, modeling and obtaining information of sentiment that expressed through facial, physical, gestures and any sentiment that can be observed in visual multimedia.

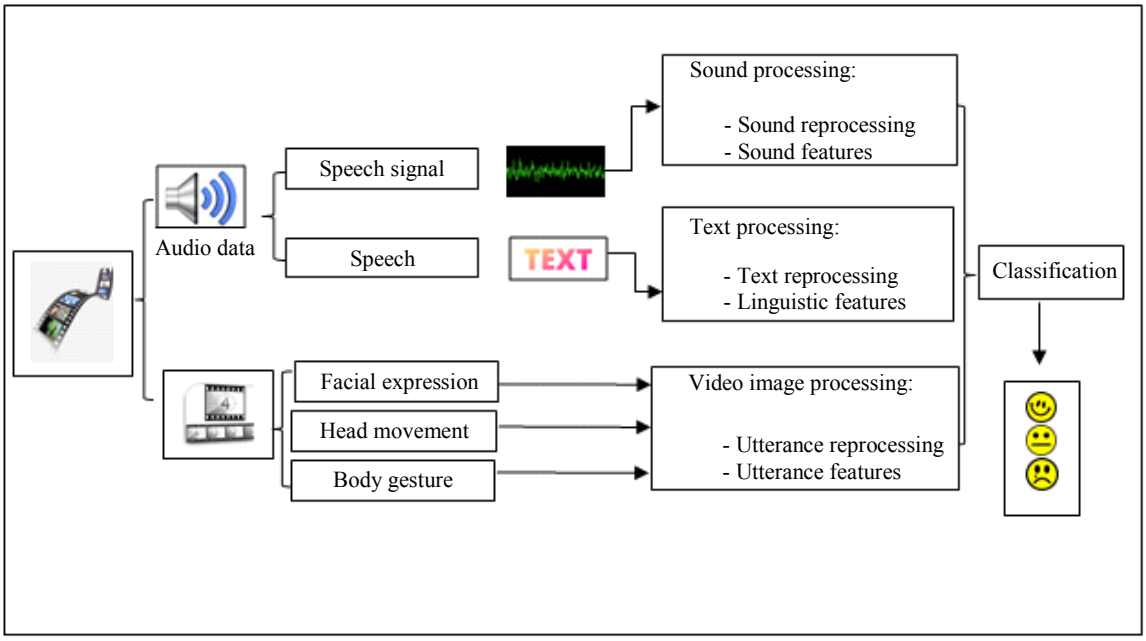

Fig. 3: Multimodal Sentiment Analysis Features/Video emotion process

Ekman and Keltner (1970) are pioneers in this field of research; they put through costly studies on facial emotions. They argued that it is possible to detect basic emotions as Anger, Joy, Sadness, Disgust and Surprise from cues of facial expressions. In this section, we present various studies on the use of visual features for multimodal affecting analysis (Fig. 3).

Facial Action Coding System

Many measurement systems for facial expressions were developed (Paul and Friesen, 1978), (Izard *et al*., 1983) and (Kring and Sloan, 1991). One of these systems, the Facial Action Coding System (FACS) developed by Paul and Friesen (1978) has been widely used. FACS depended on Action Units (AU) to reconstruct facial expressions. Human facial muscles are almost identical and AUs are based on movements of the human facial muscles, which consist of three basic parts: AU number, FACS name and muscular basis. FACS differentiates between various facial actions but cannot recognize emotions. But FACES were later complained with other resources to reconstruct emotions (Ekman *et al*., 2002), (Ekman *et al*., 1998) and (Ekman and Rosenberg, 1997; Matsumoto (1992) added new emotion (‘contempt’ or disrespect) to the set of six previously defined emotions.

Facial Feature

Most research works have concentrated on facial feature extraction for emotion and sentiment analysis, so many facial expression recognition techniques were introduce by researchers, some of these are Active Appearance Models (AAM) (Lanitis *et al*., 1995), Optical flow models (Yacoob and Davis, 1994), Active Shape Models (ASM) (Cootes *et al*., 1995), 3D Morphable Models (3DMM) (Blanz and Vetter, 1991), Muscle-based models (Ohta *et al*., 1998), 3D wireframe models (Cohen *et al*., 2003), Elastic net model (Kimura and Yachida, 1997), Geometry-based shape models (Verma *et al*., 2005), 3D Constrained Local Model (CLM) (Baltrušaitis *et al*., 2012), Generalized Adaptive View-based Appearance Model (GAVAM) (Morency *et al*., 2008), however most of those methods do not work well for videos as they do not model temporal information. In multimodal sentiment, we intend extracting temporal features from videos, a few methods which used temporal information (Bradley and Lang, 1999), (Yeasin *et al*., 2004), (Lien *et al*., 2000) and (Chang *et al*., 2004), Motion-Units facial motion (MU) (Cohen *et al*., 2003) and (Kring and Sloan, 2007) for affect recognition from videos.

An important side in video-based methods is preserve accurate tracking throughout the video sequence. A wide range of deformable models, such as muscle based models (Ohta *et al*., 1998), 3D wireframe models (Cohen *et al*., 2003), elastic net models (Kimura and Yachida, 1997) and geometry-based shape models (Verma *et al*., 2005) and (Davatzikos, 2001), have been used to track facial features in videos. After that, deformable models have demonstrated an improvement in both facial tracking and facial expression analysis accuracy, (Wen, 2003). Pantic and Rothkrantz (2000a),

Pantic and Rothkrantz (2000b) and Fasel and Luettin (2003) proposed automatic methods.

Body Gestures

Most research works have concentrated on facial feature extraction for emotion and sentiment analysis, however, there are some contributions based on features extracted from body gestures which provide valuable source of features for emotion and sentiment recognition. A relation between body gestures and emotion was explored in (De Meijer, 1989) include qualities and dimensions in different emotions. Another study was show that it is easy to distinguish the basic emotions from some simple statistical measures of motion's dynamics (Kapur *et al*., 2005).

A mathematical model to analyze body gestures for emotion expressiveness were developed by (Caridakis *et al*., 2007).

Piana *et al*. (2014), two feature were extracted facial and hand gesture features were gain and used in emotion analysis.

A successfully recent deep learning neural network was used to extract features automatically, in feature learning (LeCun *et al*., 2010), visual recognition (Kavukcuoglu *et al*., 2010), digit recognition (Hinton *et al*., 2006), image classification (Krizhevsky *et al*., 2012), musical signal processing (Hamel and Eck, 2010) and NLP (Chaturvedi *et al*., 2016).

A sentiment prediction framework was developed (Xu *et al*., 2014) to sentiment images using convolutional neural network. One of it is advantage, it is not require domain knowledge for visual sentiment. A deep 3D convolutional networks (C3D) have been proposed for spatio temporal feature learning (Tran *et al*., 2015), it show a successful feature' learning for spatio temporal in a comparison with 2D networks. A recurrent neural network (RNN) where developed by (Poria *et al*., 2017) to extract visual features (Fig. 4).

Multimodal Sentiment Analysis

Sentiments and emotions play a pivotal role in our daily lives. They assist decision making, learning, communication and situation awareness in human environments. Recently, most of researches in this field have focused on multimodal emotion recognition using visual and aural information. But at the same time, there is currently rare literature on multimodal sentiment analysis. Most of the work and available data resources are restricted to text opinion mining and the field of natural language processing. On the other side, most researches were based on English language, researches and sentiment analysis experiments were rarely based on other languages (especially the Arabic language) in comparison to English.

A. English Language

Lee and Narayanan (2005) explore domain specific emotion recognition from speech signals using data obtained from the application of real-world call center dialog. The experimental results of Language and discourse information, as well as acoustic features that most studies have focused on, show that significant improvements can be made by combining information sources in the same framework. However, their drawback is domain specific.

Eyben *et al*. (2010) contribute with three different point: First they address the task of tri-modal sentiment analysis by integrating three different modalities: Visual, audio and linguistic features, in order to determine the polarity of an input stream. Second, they present qualitative and statistical analyses that identify five multimodal features that are found helpful to differentiate between negative, neutral and positive sentiments: Polarized words, smile, gaze, pauses and voice pitch. And third, they introduce a new real dataset consisting of video opinions, which is collected from YouTube web site.

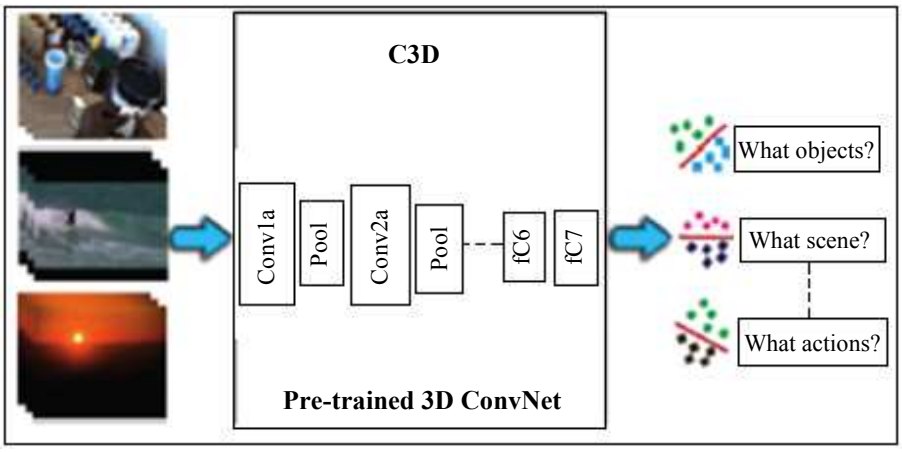

Fig. 4: 3D for extracting spatio-temporal generic video features (Poria *et al*., 2017)

Yamasaki *et al*. (2015) propose a method to accurately predict multiple impression-related user ratings for a given video talk. Their proposal considers multimodal features including linguistic as well as acoustic features, correlations between different user ratings (labels) and correlations between different feature types by using Single Markov Random Field (MRF) and the optimization of label assignment problem in order to obtain a consistent and multiple set of labels for a given video. Their experimental results on this dataset show that the proposed method obtains an accuracy of 93.3%.

D'mello and Kory (2015) design a survey and discuss both unimodal and multimodal accuracy comparison using statistical measures. The experiments compare the accuracy of different algorithms of different datasets using statistical methods.

Zeng *et al*. (2009) design a survey on multimodal emotion recognition; mainly, they focus on collecting and processing audio, visual and audio-visual material in order to identify the challenges that involve in multimodal data.

B. Other Languages

Rosas *et al*. (2013) and Morency *et al*. (2011) address the multimodal sentiment analysis by designing some experiments on a new dataset, which consisting of *Spanish* videos that are collected from the social media website, they combine the three features in comparative experiments, they show that the using of visual, audio and textual features jointly improves the use of one modality at a time (Rosas *et al*., 2013).

Pérez-Rosas *et al*. (2013) present a multimodal approach for utterance-level sentiment classification. The paper introduces a new multimodal dataset, which consists of sentiment annotated utterances that is extracted from video reviews, where each utterance is associated with a video, acoustic and linguistic DataStream. The experiments show that sentiment annotation of utterance-level visual data streams can be effectively performed and the use of multiple modalities can lead to a reduction in error rate of up to 10.5% as compared to the use of one modality at a time.

Dataset for Multimodal Sentiment Analysis

Many exhaustive surveys on sentiment analysis of text input are available, rarely surveys focus on the analysis of audio, video and multimodal input, one of the survey reviews the recent progress in the field of sentiment analysis with the focus on available datasets and sentiment analysis techniques are (D'mello and Kory, 2015) and (Zeng *et al*., 2009).

There are two main methodologies for dataset collection: Video recordings that depend on specific scripts and natural videos. Multimodal framework achieves better performance than unimodal systems, but improvement was much lower when it is trained on natural data versus acted data (D'mello and Kory, 2015). It is important to track and label the emotion of the opinion in a video and so, labeling is done at the utterance level, where every utterance is associated with sentiment label for both approaches.

There are few datasets available for multimodal sentimental analysis; the datasets in multimodal affect recognition that are recently covered are (Table 2):

- **YouTube Dataset**: The dataset was developed by Morency *et al*. (2011). The video set collected from YouTube has 47 videos, 20 of them are for females speakers and 27 for male ones that are randomly selected from youtube.com, all speakers express themselves in English. This dataset consists of 47 opinion videos with 280 utterances with manually annotated sentiment labels. The videos format is mp4 of size 360×480, each video in dataset are annotated with: Positive, negative or neutrally, 13, 12 and 22 respectively. "Towards multimodal sentiment analysis: Harvesting opinions from the web" (Morency *et al*., 2011)
- **MOUD Dataset**: The Multimodal Opinion Utterances Dataset are also collected from YouTube, it was developed by (Rosas *et al*., 2013). The final video set includes 21 male speakers and 84 female ones that are randomly selected from YouTube of different ages (15-60 years old) and different Spanish-speaking countries. This dataset consists of 498 Spanish opinion utterances from 55 unique individuals. "Multimodal Sentiment Analysis of Spanish Online Videos" (Rosas *et al*., 2013)
- **ICT-MMMO Dataset (the Institute for Creative Technologies Multi Modal Movie Opinion):** It was developed by (Wöllmer *et al*., 2013), the videos are collected from YouTube and ExpoTV and the videos review movies in English. The total number of videos is 386 (308 from YouTube and 78 from ExpoTV). Each YouTube video in dataset is annotated with: Positive, negative or neutrally, the modes are 228, 57 and 23 respectively. Each ExpoTV video in dataset is annotated with: Positive, negative or neutrally, the modes are 2, 62 and 14 respectively; however this data set had five sentiment labels
- **MOSI Dataset (Multimodal Opinion Sentiment Intensity dataset) (Zadeh *et al*., 2016):** This dataset was obtained from YouTube channels and consists of 93 videos where each one contains the opinions from one unique individual. 2199 is the total number of utterance, which are manually segmented. The dataset is annotated with labels for subjectivity, sentiment intensity for each frame; each opinion is annotated by visual features, each second hangs audio features. "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos"

Table 2: Multimodal sentiment analysis datasets

References	Dataset	Modality	Lang.	Speakers	Utterances	Sentiment classification	Features
Morency *et al*. (2011)	YouTube dataset	Text-Audio-Video	English	47 (27M, 20F)	280	Polarized words, smile, look away, pauses and pitch	20 visual +1000 linguistic +1941 acoustic
Rosas *et al*. (2013)	MOUD	Text-Audio-Video	Spanish	80 (65F, 15M)	498	Positive, Negative, and Neutral	40 visual+28 acoustic
Wöllmer *et al*. (2013)	ICT-MMMO	Text-Audio-Video	English	370 (228 positive 23 neutral, 119 negative)	-	Strongly Positive Strongly Negative, Weakly Positive Weakly Negative, Positive, Negative and Neutral	20 visual+1941 acoustic+1000 linguistic
Zadeh *et al*. (2016)	MOSI	Text-Audio-Video	English	93	2199	Strongly Positive Strongly Negative, Weakly Positive Weakly Negative, and Neutral	-

Sentimental Analysis Applications

Sentimental analysis can be used in several applications such as Marketing Strategies; for example, to understand and analyze customers' demands. It helps organization to increase innovation, retain customers and increase the operational efficiency. It can be used in Prediction to understand customer's needs and predict the future possibilities in every aspect which replace old surveys or create focus groups, which was much slower and much more expensive. Government policies; whereas politicians and governments often use sentiment analysis in order to understand how people feel about themselves and their policies. To contextualize the likes and dislikes of the user (Langlet and Clavel, 2016) and the ability to extract topic words from each user's speech.

Another domain for sentiment analysis is (Ellis *et al*. 2014) utilize multimodal sentiment analysis on broadcast video news to automatic analysis and summarization of TV programs. Multimodal sentiment analysis technologies can be also used to identify politically persuasive content (Siddiquie *et al*., 2015). Using this way and technologies will make it possible, easy and fast to obtain and mine opinions expressed through too many broadcast TV channels or any other online channels on the Internet (Langlet and Clavel, 2016). Many other applications include Recommender System, Summarization and Intelligence Comparison.

Multimodal Sentimental Analysis Challenges

Feature extraction in sentiment analysis is facing different problems such as redundancy, domain dependency, difficulty in implicit feature identification and limited work on Lexico-structural features. Followings are the general challenges in feature extraction that are identified by different researchers (Yildirim *et al*., 2004), (Pennebaker *et al*., 2015b) and (Redondo *et al*., 2007).

- **Domain Dependency:** Performance of classification and clustering, which based feature extraction techniques, is domain dependent that creates cross domain and generalization problems (Redondo *et al*., 2007). One solution for that clustering, clustering process is used to improve the categorization of the documents (De Luca and Nürnberger, 2006b). And (De Luca *et al*., 2004) introduce clustering process can enhance the semantic classification
- **High Dimensionality:** It means large feature sets that causes performance degradation due to computational problems and thus proper feature selection methods are essentially required (Wilson *et al*., 2005)
- **Different Writing Styles:** The same word can be considered positive in one situation and negative in another one. For example, the word 'long' is considered as a positive opinion in the sentence 'The laptop battery's life is long' but it is considered negative opinion in the sentence 'The laptop boot time is long'. And also people's opinions are change over time
- **Comparative Manner Expression**: A serious challenge in sentiment analysis comes from the fact that people usually express their opinions in a comparative manner; they express their positive and negative reviews in the same sentence
- **Context Quality:** Multimedia content on social media is a rich resource of data that provide us with scale, but the quality and the context of recorded material can vary and the data is limited to certain demographics that are more represented on the internet (Poria *et al*., 2017)

Conclusion

Sentiment analysis is mainly focused on the automatic recognition of opinions' polarity, as positive or negative. Multimodality is defined by analyzing more than one modality. Multimodal Sentiment Analysis refers to the combination of two or more input modes in order to improve the performance of the analysis; huge number of videos is being uploaded online continuously and so analysis of such media is important; the automatic analysis of multimodal opinion involves a deep understanding of natural languages, audio and video processing, whereas researchers are continuing to improve them. This paper provides a comprehensive overview about the multimodal sentimental concept and goal and at the end it discusses some challenges related to the field. We found that most of researches

are based on English language and rarely depend on other languages. Also we present most available datasets. This review encourages for further research in this field; and in future we will focus on the methods targeted to Arabic language.

Acknowledgment

This research was supported and funded by the research sector - Arab Open University - Kuwait Branch under decission number 18012. We thank everybody who assisted us to improve the study.

Authors Contributions

Intisar O. Hussien: Participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript, designing the research plan and organizing the study.

Yahia Hasan Jazyah: Participated in all experiments, coordinated the data-analysis and contributed to the writing of the manuscript.

Ethics

We testify that this research paper submitted to the Journal of Computer Science, title: "Multimodal Sentiment Analysis: A Comparison Study" has not been published in whole or in part elsewhere.

This research project was conducted with full compliance of research ethics norms of Arab Open University - Kuwait.

References

Abdul-Mageed, M. and M.T. Diab, 2014. SANA: A large scale multi-genre, multi-dialect lexicon for Arabic subjectivity and sentiment analysis. Proceedings of the 9th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), pp: 1162-1169.

Abdul-Mageed, M., M. Diab and S. Kübler, 2014. SAMAR: Subjectivity and sentiment analysis for Arabic social media. Comput. Speech Lang., 28: 20-37. DOI: 10.1016/j.csl.2013.03.001

Alam, F. and G. Riccardi, 2014. Predicting personality traits using multimodal information. Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, Nov. 07-07, ACM, Orlando, Florida, USA, pp: 15-18. DOI: 10.1145/2659522.2659531

Al-Ayyoub, M., S.B. Essa and I. Alsmadi, 2015. Lexicon-based sentiment analysis of Arabic tweets. Int. J. Soc. Netw. Min., 2: 101-114. DOI: 10.1504/IJSNM.2015.072280

Al-Kabi, M., M. Al-Ayyoub, I. Alsmadi and H. Wahsheh, 2016. A prototype for a standard Arabic sentiment analysis corpus. Int. Arab J. Inf. Technol., 13: 163-170.

Baltrušaitis, T., P. Robinson and L.P. Morency, 2012. 3D constrained local model for rigid and non-rigid facial tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Jun. 16-21, IEEE Xplore Press, Providence, RI, USA, pp: 2610-2617. DOI: 10.1109/CVPR.2012.6247980

Blanz, V. and T. Vetter, 1999. A morphable model for the synthesis of 3D faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Aug. 08-13, ACM, Los Angeles, CA, USA, pp: 187-194. DOI: 10.1145/311535.311556

Bradley, M.M. and P.J. Lang, 1999. Affective Norms for English Words (ANEW): Instruction manual and affective ratings. Technical report C-1, The Center for Research in Psychophysiology, University of Florida.

Cambria, E. and A. Hussain, 2015. Sentic computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. 1st Edn., Springer, Cham, ISBN-10: 3319236547, pp: 176.

Cambria, E., H. Wang and B. White, 2014. Guest editorial: Big social data analysis. Knowledge-Based Syst., 69: 1-2. DOI: 10.1016/j.knosys.2014.07.002

Caridakis, G., G. Castellano, L. Kessous, A. Raouzaiou and L. Malatesta *et al*., 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. Proceedings of the 4th IFIP International Conference on Artificial Intelligence Applications and Innovations, (IAI' 07), Springer, Boston, MA, pp: 375-388. DOI: 10.1007/978-0-387-74161-1_41

Celli, F., B. Lepri, J.I. Biel, D. Gatica-Perez and G. Riccardi *et al*., 2014. The workshop on computational personality recognition 2014. Proceedings of the 22nd ACM International Conference on Multimedia, Nov. 03-07, ACM, Orlando, Florida, USA, pp: 1245-1246. DOI: 10.1145/2647868.2647870

Chang, Y., C. Hu and M. Turk, 2004. Probabilistic expression analysis on manifolds. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 27-Jul. 2, IEEE Xplore Press, Washington, DC, USA, pp: II-520-II-527. DOI: 10.1109/CVPR.2004.1315208

Chaturvedi, I., Y.S. Ong, I.W. Tsang, R.E. Welsch and E. Cambria, 2016. Learning word dependencies in text by means of a deep recurrent belief network. Knowledge-Based Syst., 108: 144-154. DOI: 10.1016/j.knosys.2016.07.019

Cohen, I., N. Sebe, F.G. Gozman, M.C. Cirelo and T.S. Huang, 2003. Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 18-20, IEEE Xplore Press, Madison, WI, USA, pp: I-595-I-601. DOI: 10.1109/CVPR.2003.1211408

Cootes, T.F., C.J. Taylor, D.H. Cooper and J. Graham, 1995. Active shape models-their training and application. Comput. Vis. Image Understand., 61: 38-59. DOI: 10.1006/cviu.1995.1004

Crouch, S. and R. Khosla, 2012. Sentiment analysis of speech prosody for dialogue adaptation in a diet suggestion program. ACM SIGHIT Record, 2: 8-8. DOI: 10.1145/2180796.2180800

Damasio, A.R., 1994. Descartes' Error: Emotion, Reason and the Human Brain. 18th Edn., G.P. Putnam, New York, ISBN-10: 0399138943, pp: 312.

Davatzikos, C., 2001. Measuring biological shape using geometry-based shape transformations. Image Vis. Comput., 19: 63-74. DOI: 10.1016/S0262-8856(00)00056-1

De Luca, E.W. and A. Nürnberger, 2006a. Rebuilding lexical resources for information retrieval using sense folder detection and merging methods. Proceedings of the 5th International Conference on Language Resources and Evaluation, (REC' 06).

De Luca, E.W. and A. Nürnberger, 2006b. Using clustering methods to improve ontology-based query term disambiguation. Int. J. Intell. Syst., 21: 693-709. DOI: 10.1002/int.20155

De Luca, E.W. and A. Nürnberger, 2004. Improving ontology-based sense folder classification of document collections with clustering methods. Proceedings of the 2nd International Workshop on Adaptive Multimedia Retrieval, (AMR' 04), pp: 72-86.

De Meijer, M., 1989. The contribution of general features of body movement to the attribution of emotions. J. Nonverbal Behav., 13: 247-268. DOI: 10.1007/BF00990296

DeVault, D., R. Artstein, G. Benn, T. Dey and E. Fast *et al*., 2014. SimSensei kiosk: A virtual human interviewer for healthcare decision support. Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems, May 05-09, pp: 1061-1068.

Dodds, P.S. and C.M. Danforth, 2010. Measuring the happiness of large-scale written expression: Songs, blogs and presidents. J. Happiness Stud., 11: 441-456. DOI: 10.1007/s10902-009-9150-9

D'mello, S.K. and J. Kory, 2015. A review and meta-analysis of multimodal affect detection systems. ACM Comput. Surveys, 47: 43-43. DOI: 10.1145/2682899

Morrison, D., R. Wang and L.C. De Silva, 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Commun., 49: 98-112. DOI: 10.1016/j.specom.2006.11.004

Duwairi, R.M., 2015, April. Sentiment analysis for dialectical Arabic. Proceedings of the 6th International Conference on Information and Communication Systems, Apr. 7-9, IEEEE Xplore Press, Amman, Jordan, pp: 166-170. DOI: 10.1109/IACS.2015.7103221

Duwairi, R.M., R. Marji, N. Sha'ban and S., Rushaidat, 2014. Sentiment analysis in Arabic tweets. Proceedings of the 5th International Conference on Information and Communication Systems, Apr. 1-3, IEEE Xplore Press, Irbid, Jordan, pp: 1-6. DOI: 10.1109/IACS.2014.6841964

Ekman, P. and D. Keltner, 1970. Universal facial expressions of emotion. California Mental Health Res. Digest, 8: 151-158.

Ekman, P., E. Rosenberg and J. Hager, 1998. Facial Action Coding System Affect Interpretation Dictionary (FACSAID).

Ekman, P., W.V. Friesen and J.C. Hager, 2002. FACS investigator's guide. A Human Face.

Ekman, P. and E.L. Rosenberg, 1997. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA.

Ellis, J.G., B. Jou and S.F. Chang, 2014. Why we watch the news: A dataset for exploring sentiment in broadcast video news. Proceedings of the 16th International Conference on Multimodal Interaction, Nov. 12-16, ACM, Istanbul, Turkey, pp: 104-111. DOI: 10.1145/2663204.2663237

ElSahar, H. and S.R. El-Beltagy, 2015. Building Large Arabic Multi-Domain Resources for Sentiment Analysis. In: Computational Linguistics and Intelligent Text Processing, Gelbukh, A. (Ed.), Springer, Cham, pp: 23-34.

Eyben, F., M. Wöllmer, A. Graves, B. Schuller and E. Douglas-Cowie *et al*., 2010. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. J. Multimodal User Interfaces, 3: 7-19. DOI: 10.1007/s12193-009-0032-6

Fasel, B. and J. Luettin, 2003. Automatic facial expression analysis: A survey. Patt. Recognit., 36: 259-275. DOI: 10.1016/S0031-3203(02)00052-3

Gangemi, A., V. Presutti and D.R. Recupero, 2014. Frame-based detection of opinion holders and topics: A model and a tool. IEEE Comput. Intell. Magazine, 9: 20-30. DOI: 10.1109/MCI.2013.2291688

Ghareb, A.S., A.R. Hamdan, A.A. Bakar and M.R. Yaakub, 2015. Hybrid statistical rule-based classifier for Arabic text mining. J. Theoretical Applied Inform. Technol., 71: 194-204.

Hamel, P. and D. Eck, 2010. Learning features from music audio with deep belief networks. Proceedings of the 11th International Society for Music Information Retrieval Conference, Aug. 9-13, Utrecht, Netherlands, pp: 339-344.

Hinton, G.E., S. Osindero and Y.W. Teh, 2006. A fast learning algorithm for deep belief nets. Neural Comput., 18: 1527-1554. DOI: 10.1162/neco.2006.18.7.1527

Hu, X., J. Tang, H. Gao and H. Liu, 2013. Unsupervised sentiment analysis with emotional signals. Proceedings of the 22nd International Conference on World Wide Web, May 13-17, ACM, Rio de Janeiro, Brazil, pp: 607-618. DOI: 10.1145/2488388.2488442

Ibrahim, H.S., S.M. Abdou and M. Gheith, 2015. Sentiment analysis for modern standard Arabic and colloquial.

Izard, C.E., L.M. Dougherty and E.A. Hembree, 1983. A system for identifying Affect Expressions by holistic judgments (AFFEX). Instructional Resources Center, University of Delaware.

Kapur, A., A. Kapur, N. Virji-Babul, G. Tzanetakis and P.F. Driessen, 2005. Gesture-based affective computing on motion capture data. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Oct. 22-24, Springer, Beijing, China, pp: 1-7. DOI: 10.1007/11573548_1

Kaushik, L., A. Sangwan and J.H. Hansen, 2013a. Sentiment extraction from natural audio streams. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, May 26-31, IEEE Xplore Press, Vancouver, BC, Canada, pp: 8485-8489. DOI: 10.1109/ICASSP.2013.6639321

Kaushik, L., A. Sangwan and J.H. Hansen, 2013b. Automatic sentiment extraction from YouTube videos. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Dec. 8-12, IEEE Xplore Press, Olomouc, Czech Republic, pp: 239-244. DOI: 10.1109/ASRU.2013.6707736

Kavukcuoglu, K., P. Sermanet, Y.L. Boureau, K. Gregor and M. Mathieu *et al*., 2010. Learning convolutional feature hierarchies for visual recognition. Proceedings of the 23rd International Conference on Neural Information Processing Systems, Dec. 06-09, Curran Associates Inc., Vancouver, British Columbia, Canada, pp: 1090-1098.

KgaogeloLetsebe, 2017. The benefits of sentiment data analysis. Portals journalistJohannesburg,

Kimura, S. and M. Yachida, 1997. Facial expression recognition and its degree estimation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 17-19, IEEE Xplore Press, San Juan, Puerto Rico, USA, pp: 295-300. DOI: 10.1109/CVPR.1997.609338

Koolagudi, S.G., N. Kumar and K.S. Rao, 2011. Speech emotion recognition using segmental level prosodic analysis. Proceedings of the International Conference on Devices and Communications, Feb. 24-25, IEEE Xplore Press, Mesra, India, pp: 1-5. DOI: 10.1109/ICDECOM.2011.5738536

Kring, A.M. and D. Sloan, 1991. The Facial Expression Coding System (FACES): A users guide. Unpublished Manuscript.

Kring, A.M. and D.M. Sloan, 2007. The Facial Expression Coding System (FACES): Development, validation and utility. Psychol. Assess., 19: 210-210. DOI: 10.1037/1040-3590.19.2.210

Krizhevsky, A., I. Sutskever and G.E. Hinton, 2012. Imagenet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems, Dec. 03-06, Curran Associates Inc., Lake Tahoe, Nevada, pp: 1097-1105.

Langlet, C. and C. Clavel, 2016. Grounding the detection of the user's likes and dislikes on the topic structure of human-agent interactions. Knowledge-Based Syst., 106: 116-124. DOI: 10.1016/j.knosys.2016.05.038

Lanitis, A., C.J. Taylor and T.F. Cootes, 1995. Automatic face identification system using flexible appearance models. Image Vis. Comput., 13: 393-401. DOI: 10.1016/0262-8856(95)99726-H

LeCun, Y., K. Kavukcuoglu and C. Farabet, 2010. Convolutional networks and applications in vision. Proceedings of the IEEE International Symposium on Circuits and Systems, May 30-Jun. 2, IEEE Xplore Press, Paris, France, pp: 253-256. DOI: 10.1109/ISCAS.2010.5537907

Lee, C.M. and S.S. Narayanan, 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process., 13: 293-303. DOI: 10.1109/TSA.2004.838534

Lien, J.J.J., T. Kanade, J.F. Cohn and C.C. Li, 2000. Detection, tracking and classification of action units in facial expression. Robot. Autonomous Syst., 31: 131-146. DOI: 10.1016/S0921-8890(99)00103-7

Liu, B. and L. Zhang, 2012. A Survey of Opinion Mining and Sentiment Analysis. In: Mining Text Data, Aggarwal, C. and C. Zhai (Eds.), Springer US, pp: 415-463.

Luengo, I., E. Navas, I. Hernáez and J. Sánchez, 2005. Automatic emotion recognition using prosodic parameters. Proceedings of the 9th European Conference on Speech Communication and Technology, Sept. 4-8, Lisbon, Portugal, pp: 493-496.

Matsumoto, D., 1992. More evidence for the universality of a contempt expression. Motivat. Emot., 16: 363-368. DOI: 10.1007/BF00992972

Melville, P., W. Gryc and R.D. Lawrence, 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Jun. 28-Jul. 01, ACM, Paris, France, pp: 1275-1284. DOI: 10.1145/1557019.1557156

Metze, F., A. Batliner, F. Eyben, T. Polzehl and B. Schuller *et al*., 2010. Emotion recognition using imperfect speech recognition. ISCA.

Morency, L.P., R. Mihalcea and P. Doshi, 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. Proceedings of the 13th International Conference on Multimodal Interfaces, Nov. 14-18, ACM, Alicante, Spain, pp: 169-176. DOI: 10.1145/2070481.2070509

Morency, L.P., J. Whitehill and J. Movellan, 2008. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition, Sept. 17-19, IEEE Xplore Press, Amsterdam, Netherlands, pp: 1-8. DOI: 10.1109/AFGR.2008.4813429

Muda, L., M. Begam and I. Elamvazuthi, 2010. Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. arXiv preprint arXiv:1003.4083.

Murray, I.R. and J.L. Arnott, 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoustical Society Am., 93: 1097-1108. DOI: 10.1121/1.405558

Narr, S., E.W. De Luca and S. Albayrak, 2011. Extracting semantic annotations from twitter. Proceedings of the 4th Workshop on Exploiting Semantic Annotations in Information Retrieval, Oct. 28-28, ACM, Glasgow, Scotland, UK, pp: 15-16. DOI: 10.1145/2064713.2064723

Navas, E., I. Hernaez and I. Luengo, 2006. 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: 1117-1127. DOI: 10.1109/TASL.2006.876121

Obaidat, I., R. Mohawesh, M. Al-Ayyoub, A.S. Mohammad and Y. Jararweh, 2015. Enhancing the determination of aspect categories and their polarities in Arabic reviews using lexicon-based approaches. Proceedings of the IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, Nov. 3-5, IEEE Xplore Press, Amman, Jordan, pp: 1-6. DOI: 10.1109/AEECT.2015.7360595

Ohta, H., H. Saji and H. Nakatani, 1998. Recognition of facial expressions using muscle-based feature models. Proceedings of the 14th International Conference on Pattern Recognition, Aug. 20-20, IEEE Xplore Press, Brisbane, Queensland, Australia, pp: 1379-1381. DOI: 10.1109/ICPR.1998.711959

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (NLP' 02), Association for Computational Linguistics, pp: 79-86. DOI: 10.3115/1118693.1118704

Pantic, M. and L.J. Rothkrantz, 2000. Expert system for automatic analysis of facial expressions. Image Vis. Comput., 18: 881-905. DOI: 10.1016/S0262-8856(00)00034-2

Pantic, M. and L.J.M. Rothkrantz, 2000. Automatic analysis of facial expressions: The state of the art. IEEE Trans. Patt. Anal. Mach. Intell., 22: 1424-1445. DOI: 10.1109/34.895976

Paul, E. and W. Friesen, 1978. Facial action coding system investigator's guide.

Pennebaker, J.W., R.L. Boyd, K. Jordan and K. Blackburn, 2015a. The development and psychometric properties of LIWC2015. LIWC.net, Austin.

Pennebaker, J.W., R.L. Boyd, K. Jordan and K. Blackburn, 2015b. The development and psychometric properties of LIWC2015. The University of Texas at Austin.

Pereira, J.C., J. Luque and X. Anguera, 2014. Sentiment retrieval on web reviews using spontaneous natural speech. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-9, IEEE Xplore Press, Florence, Italy, pp: 4583-4587. DOI: 10.1109/ICASSP.2014.6854470

Pérez-Rosas, V. and R. Mihalcea, 2013. Sentiment analysis of online spoken reviews. INTERSPEECH.

Pérez-Rosas, V., R. Mihalcea and L.P. Morency, 2013. Utterance-level multimodal sentiment analysis. ACL.

Piana, S., A. Stagliano, F. Odone, A. Verri and A. Camurri, 2014. Real-time automatic emotion recognition from body gestures. arXiv preprint arXiv:1402.5047.

Picard, R.W., 2010. Affective computing: from laughter to IEEE. IEEE Trans. Affective Comput., 1: 11-17. DOI: 10.1109/T-AFFC.2010.10

Poria, S., B. Agarwal, A. Gelbukh, A. Hussain and N. Howard, 2014. Dependency-based semantic parsing for concept-level text analysis. Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Apr. 06-12, Springer, Kathmandu, Nepal, pp: 113-127. DOI: 10.1007/978-3-642-54906-9_10

Poria, S., E. Cambria, N. Howard, G.B. Huang and A. Hussain, 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing, 174: 50-59. DOI: 10.1016/j.neucom.2015.01.095

Poria, S., E. Cambria, R. Bajpai and A. Hussain, 2017. A review of affective computing: From unimodal analysis to multimodal fusion. Inform. Fus., 37: 98-125. DOI: 10.1016/j.inffus.2017.02.003

Redondo, J., I. Fraga, I. Padrón and M. Comesaña, 2007. The Spanish adaptation of ANEW (affective norms for English words). Behav. Res. Meth., 39: 600-605. DOI: 10.3758/BF03193031

Rosas, V.P., R. Mihalcea and L.P. Morency, 2013. Multimodal sentiment analysis of Spanish online videos. IEEE Intell. Syst., 28: 38-45. DOI: 10.1109/MIS.2013.9

Salameh, M., S. Mohammad and S. Kiritchenko, 2015. Sentiment after translation: A case-study on Arabic social media posts. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May 31-Jun. 5, Association for Computational Linguistics, Denver, Colorado, pp: 767-777.

Sarkar, C., S. Bhatia, A. Agarwal and J. Li, 2014. Feature analysis for computational personality recognition using YouTube personality data set. Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition, Nov. 07-07, ACM, Orlando, Florida, USA, pp: 11-14. DOI: 10.1145/2659522.2659528

Schuller, B., A. Batliner, S. Steidl and D. Seppi, 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun., 53: 1062-1087. DOI: 10.1016/j.specom.2011.01.011

Sharef, N.M., H.M. Zin and S. Nadali, 2016. Overview and future opportunities of sentiment analysis approaches for big data. J. Comput. Sci., 12: 153-168. DOI: 10.3844/jcssp.2016.153.168

Siddiquie, B., D. Chisholm and A. Divakaran, 2015. Exploiting multimodal affect and semantics to identify politically persuasive web videos. Proceedings of the ACM on International Conference on Multimodal Interaction, Nov. 09-13, ACM, Seattle, Washington, USA, pp: 203-210. DOI: 10.1145/2818346.2820732

Socher, R., D. Chen, C.D. Manning and A. Ng, 2013. Reasoning with neural tensor networks for knowledge base completion. Proceedings of the 26th International Conference on Neural Information Processing Systems, Dec. 05-10, Curran Associates Inc., Lake Tahoe, Nevada, pp: 926-934.

Stone, P.J., 1997. Thematic Text Analysis: New Agendas for Analyzing Text Content. In: Text Analysis for the Social Sciences, Roberts, C.W. (Ed.), Lawrence Erlbaum, Mahwah, NJ, pp: 33-54.

Tran, D., L. Bourdev, R. Fergus, L. Torresani and M. Paluri, 2015. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, Dec. 7-13, IEEE Xplore Press, Santiago, Chile, pp: 4489-4497. DOI: 10.1109/ICCV.2015.510

Wang, H., A. Hanafy, M. Bahgat, S. Noeman and O.S. Emam *et al*., 2015. A system for extracting sentiment from large-scale Arabic social data. Proceedings of the 1st International Conference on Arabic Computational Linguistics, Apr. 17-20, IEEE Xplore Press, Cairo, Egypt, pp: 71-77. DOI: 10.1109/ACLing.2015.17

Turney, P.D., 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Jul. 07-12, Association for Computational Linguistics, Philadelphia, Pennsylvania, pp: 417-424. DOI: 10.3115/1073083.1073153

Verma, R., C. Davatzikos, J. Loughead, T. Indersmitten and R. Hu *et al*., 2005. Quantification of facial expressions using high-dimensional shape transformations. J. Neurosci. Meth., 141: 61-73. DOI: 10.1016/j.jneumeth.2004.05.016

Wen, Z., 2003. Capturing subtle facial motions in 3d face tracking. Proceedings of the 9th IEEE International Conference on Computer Vision, Oct. 13-16, IEEE Xplore Press, Nice, France, pp: 1343-1350. DOI: 10.1109/ICCV.2003.1238646

Wiebe, J.M., R.F. Bruce and T.P. O'Hara, 1999. Development and use of a gold-standard data set for subjectivity classifications. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, Jun. 20-26, College Park, Maryland, pp: 246-253. DOI: 10.3115/1034678.1034721

Wilson, T., J. Wiebe and P. Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Oct. 06-08, Vancouver, British Columbia, Canada, pp: 347-354. DOI: 10.3115/1220575.1220619

Wöllmer, M., F. Weninger, T. Knaup, B. Schuller and C. Sun. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intell. Syst., 28: 46-53. DOI: 10.1109/MIS.2013.34

Wu, C.H. and W.B. Liang, 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. Affective Comput., 2: 10-21. DOI: 10.1109/T-AFFC.2010.16

Xu, C., S. Cetintas, K.C. Lee and L.J. Li, 2014. Visual sentiment prediction with deep convolutional neural networks. arXiv preprint arXiv:1411.5731.

Yacoob, Y. and L. Davis, 1994. Computing spatio-temporal representations of human faces. Doctoral dissertation, Department of Computer Science, University of Maryland at College Park.

Yamasaki, T., Y. Fukushima, R. Furuta, L. Sun and K. Aizawa *et al*., 2015. Prediction of user ratings of oral presentations using label relations. Proceedings of the 1st International Workshop on Affect and Sentiment in Multimedia, Oct. 30-30, ACM, Brisbane, Australia, pp: 33-38. DOI: 10.1145/2813524.2813533

Yeasin, M., B. Bullot and R. Sharma, 2004. From facial expression to level of interest: A spatio-temporal approach. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Jun. 27-Jul. 2, IEEE Xplore Press, Washington, DC, USA, pp: II-922-II-927. DOI: 10.1109/CVPR.2004.1315264

Yildirim, S., M. Bulut, C.M. Lee, A. Kazemzadeh and Z. Deng *et al*., 2004. An acoustic study of emotions expressed in speech. Proceedings of the 8th International Conference on Spoken Language Processing, Oct. 4-8, International Speech Communication Association, Jeju Island, Korea, pp: 2193-2196.

Yu, H. and V. Hatzivassiloglou, 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (NLP' 03), Association for Computational Linguistics, pp: 129-136. DOI: 10.3115/1119355.1119372

Zadeh, A., R. Zellers, E. Pincus and L.P. Morency, 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259.

Zeng, Z., M. Pantic, G.I. Roisman and T.S. Huang, 2009. A survey of affect recognition methods: Audio, visual and spontaneous expressions. IEEE Trans. Patt. Anal. Mach. Intell., 31: 39-58. DOI: 10.1109/TPAMI.2008.52