

Accepted Manuscript

A Survey of Multimodal Sentiment Analysis

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller,
Shih-Fu Chang, Maja Pantic

PII: S0262-8856(17)30119-1
DOI: doi: [10.1016/j.imavis.2017.08.003](https://doi.org/10.1016/j.imavis.2017.08.003)
Reference: IMAVIS 3638

To appear in: *Image and Vision Computing*

Received date: 31 July 2016
Revised date: 25 May 2017
Accepted date: 22 August 2017



Please cite this article as: Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, Maja Pantic, A Survey of Multimodal Sentiment Analysis, *Image and Vision Computing* (2017), doi: [10.1016/j.imavis.2017.08.003](https://doi.org/10.1016/j.imavis.2017.08.003)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

A Survey of Multimodal Sentiment Analysis

Mohammad Soleymani^{a,*}, David Garcia^b, Brendan Jou^{c,**}, Björn Schuller^{d,e,a}, Shih-Fu Chang^c, Maja Pantic^{e,f}

^aSwiss Center for Affective Sciences, University of Geneva, Geneva, Switzerland

^bChair of Systems Design, ETH Zürich, Zurich, Switzerland

^cElectrical Engineering Department and Computer Science Department, Columbia University, New York, NY, USA

^dChair of Complex and Intelligent Systems, University of Passau, Passau, Germany

^eDepartment of Computing, Imperial College London, London, United Kingdom

^fHuman Media Interaction, EEMCS, University of Twente, Enschede, The Netherlands

Abstract

Sentiment analysis aims to automatically uncover the underlying attitude that we hold towards an entity. The aggregation of these sentiment over a population represents opinion polling and has numerous applications. Current text-based sentiment analysis rely on the construction of dictionaries and machine learning models that learn sentiment from large text corpora. Sentiment analysis from text is currently widely used for customer satisfaction assessment and brand perception analysis, among others. With the proliferation of social media, multimodal sentiment analysis is set to bring new opportunities with the arrival of complementary data streams for improving and going beyond text-based sentiment analysis. Since sentiment can be detected through affective traces it leaves, such as facial and vocal displays, multimodal sentiment analysis offers promising avenues for analyzing facial and vocal expressions in addition to the transcript or textual content. These approaches leverage emotion recognition and context inference to determine the underlying polarity and scope of an individual's sentiment. In this survey, we define sentiment and the problem of multimodal sentiment analysis and review recent developments in multimodal sentiment analysis in different domains, including spoken reviews, images, video blogs, human-machine and human-human interaction. Challenges and opportunities of this emerging field are also discussed leading to our thesis that multimodal sentiment analysis holds a significant untapped potential.

Keywords: sentiment, affect, sentiment analysis, human behavior analysis, computer vision, affective computing

1. Introduction

Sentiment is a long-term disposition evoked when a person encounters a specific topic, person, or entity [1]. Understanding people's position, attitude or opinion towards a certain entity has many applications. For example, companies are interested in understanding how their products or their brand is perceived among their customers [2]. Political parties are interested in opinion polling to gauge voting intentions [3]. Automatic sentiment analysis is the computational understanding of one's position, attitude or opinion towards an entity, person or topic [4]. With the advent of the World Wide Web and shortly after, the social web, individuals are enabled to broadly express their opinions through these media. This has provided a very rich resource for opinion mining and sentiment analysis and

promoted the development of automatic sentiment analysis [5, 4, 6]. Text-based sentiment analysis has long been the standard bearer in this area and only recently has sentiment analysis from other modalities, such as speech and vision, begun to be considered.

Liu and Zhang [4] defined sentiment analysis as a problem of automatic identification of four components of a sentiment, including, entity, aspect, opinion holder, aspect's sentiment. For example, in the sentence "Sally likes the screen resolution in Nexus 6P," "Nexus 6P" is the entity, "screen resolution" is the aspect, Sally is the opinion holder and the associated sentiment is positive. A successful automatic sentiment analysis system should be able to extract all these four components correctly. The available user-generated data on the Internet, containing people's opinion or sentiment, is unstructured and noisy [6]. Challenges such as negation, irony and ambiguous phrases with implicit hints add to this challenge. Therefore, correct extraction of all these components is very challenging.

Sentiment analysis from text is a well-researched topic that now enjoys a number of industry solutions. Text-based sentiment analysis has been applied to a broad set of applications including movie box-office performance prediction [7], stock market performance prediction [8] and

*Corresponding author

**Now at Google Inc

Email addresses: mohammad.soleymani@unige.ch (Mohammad Soleymani), dgarcia@ethz.ch (David Garcia), bjou@caa.columbia.edu (Brendan Jou), bjoern.schuller@imperial.ac.uk (Björn Schuller), shih.fu.chang@columbia.edu (Shih-Fu Chang), m.pantic@imperial.ac.uk (Maja Pantic)

election outcome prediction [9]. Today though, we are seeing a shift toward an increasingly multimodal social web. For example, vloggers post their opinions on YouTube¹, and photos commonly accompany user posts on Instagram² and Twitter³. In the research community, there are currently three major lines of investment in multimodal sentiment analysis:

- multimodal sentiment analysis in spoken reviews and vlogs [10, 11, 12],
- multimodal sentiment analysis in human-machine and human-human interaction [13],
- and visual sentiment analysis analyzing images and their associated tags posted on social media [14].

And yet, multimodal sentiment analysis is still in its infancy and more research and industry investment is needed to demonstrate its full potential.

A number of recent studies have attempted to recognize sentiment expressed in social multimedia from multimodal signals, including visual, audio and textual information. Video blogs (vlogs) or spoken reviews that are posted on social multimedia platforms, such as YouTube, contain expressions of sentiment, e.g., a video depicting a user talking about a product or a movie. Typically, speech transcripts along with facial and vocal expressions are analyzed separately and the results of unimodal, text-based sentiment analysis are fused in post to form a “multimodal sentiment analysis” system [10, 11]. Similarly, sentiment analysis can be also done using speech analysis as demonstrated in numerous studies [15, 16, 17].

Another emerging application for multimodal sentiment analysis is sentiment analysis in human-avatar or human-human interaction. Clavel and Callejas [18] posited that sentiment expressed in the interaction between a person and an Embodied Conversational Agent (ECA) can be used to improve the quality of the interaction. An ECA is a computer-generated character which can imitate vocal, facial and body expressions to enrich an interaction between human and machines. One effort in multimodal sentiment analysis is the European Horizon 2020 Program Project on Automatic Sentiment Analysis in the Wild (SEWA)⁴, focused on building multimodal human behavior analysis tools to extract sentiment in response to videos such as product advertisement. Non-verbal user sentiments is analyzed as well as verbal feedback in a dyadic human-human interaction on the watched video. Sentiment analysis is also component in development in the ARIA VALUSPA European Horizon 2020 Project⁵ for improving human-ECA interactions.

A recent development in multimodal sentiment analysis is visual sentiment analysis. Social media users often share text messages with accompanying images/video, and these visual multimedia are an additional channel of information in expressing user sentiment. Mid-level visual sentiment representations are one useful construct for extracting sentiment and entities in text-based sentiment analysis [29]. The pioneering effort of Borth *et al.* [14] in visual automatic sentiment analysis proposed to detecting adjective-noun pairs as a mid-level representation toward sentiment analysis in images/video. Each adjective-noun pair was manually assigned a sentiment score and the resulting ontology called VSO was released along with an image dataset. Machine learning models trained with visual content features were also used to automatically detect these adjective-noun pairs in images. This line of research was further expanded by the introduction of a Multilingual Visual Sentiment Ontology (MVSO) [23], containing multi-lingual adjective-noun pairs from multiple languages along with an even larger image dataset and deep learning-based classifiers.

There are several existing surveys covering automatic sentiment analysis in text [4, 5] or in a specific domain, such as human-agent interaction [18]. We focus on multimodal sentiment analysis irrespective of its domain and aim to provide an overview of the sentiment analysis for researchers in computer vision, affective computing and multimodal interaction communities who are not necessarily familiar with the concepts related to sentiment analysis in text. An overview of the reviewed modalities and example media are given in Table 1. We also discuss the challenges and opportunities of multimodal sentiment analysis as an emerging field. In the remainder of the survey, we define sentiment in Section 2. Section 3 reviews existing computational methods in text analysis, visual sentiment analysis and multimodal sentiment analysis. Applications of multimodal sentiment analysis are given in Section 4. Existing challenges and perspectives on multimodal sentiment analysis are discussed in Section 5.

2. Definition and Terminology

2.1. Problem Definition

Affect, feeling, emotion, sentiment and opinion are often used interchangeably in the literature [30]. There are many definitions of emotion and its related concepts, often dependent on the emotion theory they subscribe to [31]. Scherer [31] tried to define emotions and differentiate them from other affective phenomena such as mood. Scherer’s perspective is mainly based on the component process model and cognitive appraisal theory [32, 33]. According to Scherer’s account, emotions are short-term phenomena which require triggers, involves cognitive appraisals, bodily reactions, action tendencies (e.g., fight or flight), expressions (e.g., facial and vocal) and subjective feelings. Mood is a long-term diffuse affect state with no

¹<https://www.youtube.com>

²<https://www.instagram.com>

³<https://twitter.com>

⁴<http://www.sewaproject.eu>

⁵<http://aria-agent.eu>

Table 1: Overview of the reviewed methods with an illustrated example of the medium and typical sentiment analysis approaches.

Modality	Example data	Typical features & methods
Text	 <p>product review</p>	Lexicon-based dictionaries; bag-of-words; word embeddings [19] in combination with classifiers such as SVM or deep neural networks [4, 20]
Speech	 <p>speech</p>	Paralinguistic features, e.g., pitch, in combination with classifiers such as SVM or deep recurrent neural networks [21, 15]
Visual	 <p>image</p>	Mid-level visual concepts corresponding to Adjective-Noun Pairs that carry strong sentiments [14] through convolutional neural networks [22, 23]; additional features include facial expression [24, 25]; facial action units; visual aesthetics [26]
Multimodal	 <p>vlogs</p>	Multimodal fusion of text, facial expression and paralinguistic features [10, 12, 27, 28]

apparent trigger that can last hours or days [31]. Feelings are a subjective experience of emotion.

Deonna and Teroni [1] discussed the definition of affective phenomena, including emotion, sentiment, emotional dispositions and character traits from a philosophical perspective. Deonna and Teroni [1] define sentiment as a disposition or a love-or-hate deep-seated opinion that comes in different forms, such as, “the affection you may have for your hamster, your devotion to your country, your dislike for the banking establishment, and your great fondness for the most recent electronic gadget.” Sentiment only manifests itself when the holder of sentiment is facing a situation in which the entity or object is involved or evoked. Deonna and Teroni argue that sentiment can be identified or traced through the affective interactions with the object or entity. For example, if person A holds a positive sentiment towards person B, A evaluates positively a situation where B is in a pleasant situation. As a result, the sentiment-holder is emotionally sensitive to the fortune of the object or entity.

Munezero *et al.* [30] provide an in-depth discussion on the difference between sentiment, opinion, emotion and feeling in the context of emotion recognition and sentiment analysis in text. Munezero and colleagues differentiate between emotion and sentiment based on their duration; emotion is short-term whereas sentiment is long-term. They define opinions as judgments that are open to

dispute and uncertain but do not need to be emotionally charged. Opinion is an expression of personal interpretations of information.

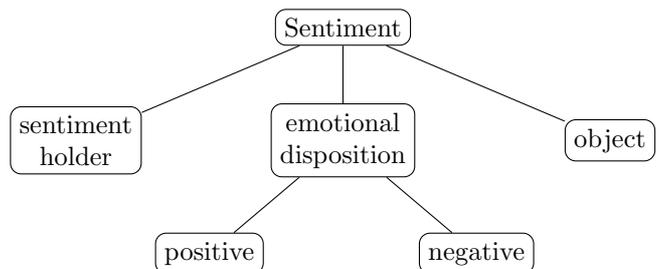


Figure 1: Schematic structure of sentiments; based on a figure from [30].

In sum, unlike emotions that typically have external manifestations, sentiment and opinion do not necessarily manifest themselves in behavior or expressions. Emotions involve a person, emotional experience including subjective feeling and bodily changes and a target. Sentiments involve a sentiment holder, an emotional disposition, i.e., polarity (positive or negative), and an object. A schematic representation of opinion and sentiment are given in Figures 1 and 2. Additional distinctions between affect, feeling, and opinion exist, for further discussion on this, we refer the reader to [30].

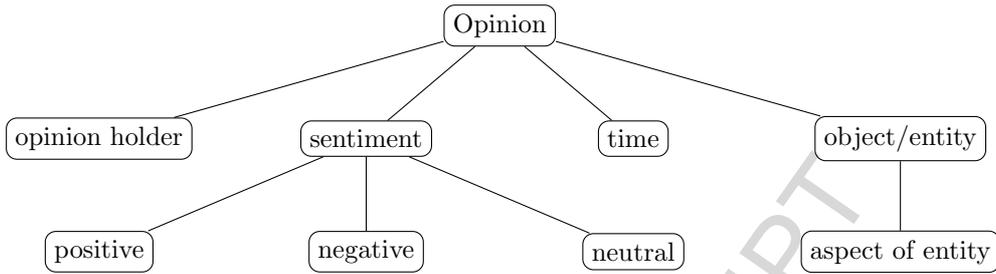


Figure 2: Schematic structure of opinions; partially reconstructed from [30].

2.2. Opinion Mining, Sentiment Analysis, and Emotion Recognition

Opinion mining and sentiment analysis have been used inter-changeably in the literature [5]. Opinion mining is the automatic method to extract and analyze the subjective judgments on different aspects of an item or entity. Kim and Hovy [34] defined an opinion with a quadruple including topic, holder, claim and sentiment. According to Kim and Hovy, sentiment analysis is identifying the sentiment of a claim. Similarly, Liu and Zhang [4] formalize the problem of opinion mining as a problem of identifying a tuple containing an entity, aspect of entity, time, opinion holder and opinion polarity. Sentiment analysis should be able to capture sentiment holder, entity and aspect of entity in addition to the polarity.

Emotion recognition is the automatic identification of an episodic emotional reaction, often of a single person; unlike opinions, emotions are short-term [30]. Emotion recognition can be used for identifying the polarity of sentiment when the stimulus is the entity of interest. Cambria *et al.* [6] suggested that opinion mining and sentiment analysis focus on two different problems of polarity detection and emotion recognition. However, if we hold to the definition of sentiment as a disposition with polarity, then sentiment analysis cannot be boiled down to an emotion recognition problem.

Given these definitions and the interconnections between emotions, sentiment and opinion, a careful consideration is needed when addressing each of these problems. We first need to define our problem and clearly delineate whether it is sentiment analysis or emotion recognition. As a result, unlike Clavel and Callejas [18], we do not believe that emotion recognition in human-avatar interactions is sentiment analysis. And yet, emotion recognition from human behavior can still be a component for sentiment analysis in human-machine or human-human interactions.

2.3. Representations of Sentiment

Sentiment analysis is mainly focused on the automatic recognition of opinions' polarity, i.e., positive or negative. One less-studied aspect of the existing sentiment representations is the intensity of sentiment. Zadeh [27] proposed to perform multilevel sentiment analysis to consider the intensity of sentiment.

Fontaine *et al.* [35] suggested that in addition to arousal, valence and dominance, predictability is also required to represent emotions. Similarly, some suggested augmenting polarity only representation in sentiment analysis with additional dimensions, discrete emotional representation or appraisals [18, 30]. It is, however, unclear how a sentiment as an affective disposition can be represented by discrete emotions such as disgust. We can argue that the emotion of a person is a result of the sentiment he holds; for example, one might feel sadness due to the sickness of a loved one. In the previous example, the positive sentiment towards the loved one and the appraisal of her unfortunate situation by the sentiment holder result in an emotion, i.e., sadness. However, discrete emotions cannot represent the sentiment itself. According to cognitive appraisal theory [33], a concurrent set of evaluations or appraisals of an object or event results in emotions. For example, the appraisal of a traffic jam for a driver stuck in the traffic is goal incongruence - since the traffic jam is obstructing the goal of the driver to reach her destination. Although appraisal of an entity is an important factor in forming an opinion or sentiment, it is not appropriate for sentiment representation. First, sentiment is often formed after multiple exposures and not an episodic appraisal (e.g., intrinsic pleasantness or goal relevance); second, it is not straightforward to use appraisals in sentiment analysis. To sum up, if we take the definition of sentiment as an emotionally charged opinion, sentiment analysis will be better represented in a uni-dimensional space describing the polarity and its intensity.

3. Computational Approaches for Sentiment Analysis

3.1. Sentiment in Text

The field of sentiment analysis from textual data started as an alternative to topic detection, aiming at the extraction of evaluative meaning. Back in 1992, Hearst proposed the automatic detection of *directionality* as a measurement of subjective states, based on models from cognitive linguistics [36], followed by Sack's proposition of the identification of point of view as an approximation to subjective content [37]. In the early 2000s, numerous works tackled

the question of sentiment analysis from text. The influential 2008 review of Pang and Lee captured this rise in interest [5], and subsequent new methods and applications have been integrated in comprehensive benchmarks in recent research [38]. Here, we provide an overview of approaches to sentiment analysis in Natural Language Processing (NLP), including supervised and unsupervised methods, and future directions and limitations in the field.

Supervised sentiment analysis aims at building predictive models for sentiment based on annotated datasets from which learning is automated. This approach builds a feature vector of each text entry in which certain aspects or word frequencies are quantified, to then train standard machine learning tools and validate them against reference annotated texts. A first approach to supervised sentiment analysis is the classification of texts as subjective or objective [39], using annotations that tag evaluative content but not its orientation. One of the earliest motivations for supervised approaches was the analysis of sentiment with respect to stock trading [40], leading to the development of a voting classifier that combined various methods to extract sentiment from finance message boards. Further developments focus on user-generated reviews, taking the star-rating of the review as a sentiment label. Naïve Bayes classifiers and Support Vector Machines (SVM) [41] proved to be effective machine learning models adopted in early approaches to the problem. SVM in particular has been a useful model for supervised sentiment analysis in general product reviews [42]. These approaches calculate features over bag-of-words models of the reviews, for which a more detailed review can be found in Pang and Lee [5]. A notable improvement over this approach is the use of tree banks [43], a model in which compositionality of meanings is taken into account in a hierarchical fashion,

Most supervised approaches to sentiment analysis are trained on certain domain or communication context, such as social media or news. Arguably the most prolific domain for sentiment analysis is the classification of polarity of Twitter short public messages, called tweets. The SemEval competition [44] has included a tweet sentiment polarity task since 2013, in which developers of sentiment analysis methods submit their tools and are independently evaluated on a test dataset. A SVM fed by a wide range of features achieved the best results in 2013 [45], and an improved model with fewer features but more lexicon resources was introduced in 2014 [46]. The best-performing approach in the 2015 edition of the competition combines SVM with other methods including maximum-entropy and stochastic gradient descent optimizations [47]. In the latest edition at the time of writing [48], a tool called Swiss-Cheese [49] achieved the best results to date by training convolutional neural networks with large datasets of Tweets with emoticons. An alternative benchmark compared various supervised methods for Twitter sentiment [50], finding that a previous approach by some of the authors of [51], based on a bootstrap parametric ensemble framework, preforms best on average.

Domain adaptation aims at developing supervised sentiment classifiers that can be applied across contexts [52]. This approach is of special relevance, since the main caveat of supervised methods in sentiment analysis is the necessity of labeled text datasets to train the classifier. The adaptation of classifiers across domains is a far from trivial task, since some contexts are more formal or produce longer texts than others. For example, the formality and length of movie reviews allow the creation of tree banks [43], but the application of this approach to social media is still an open question.

When labeled data is scarce or unavailable for certain applications, *unsupervised* approaches allow an estimation based on expert knowledge without annotated data. The expert knowledge used for the estimation is often encoded in a lexicon, in which words or phrases are annotated with their sentiment meanings. These lexica can be manually annotated, asking raters to interpret the meaning of words. One of the most widely used reference lexica for sentiment analysis is the General Inquirer (GI) [53], which takes a list of positive and negative terms. The GI is often combined with other lexica [54], as a way to increase the recall of unsupervised techniques. Another widely applied method is Linguistic Inquiry and Word Count (LIWC) [55], a software that counts positive and negative affect terms in text. Another notable reference lexicon is Affective Norms for English Words (ANEW) [56], which was not originally designed for sentiment analysis but has proven useful in measuring sentiment as happiness expressed through text [57]. This application motivated extensions to much larger word-bases [58], and versions in multiple languages [59, 60, 61]. Most of these lexica are produced by crowdsourcing annotations and can be extended to capture a wide range of emotional states, like in the NRC lexicon [45].

Emotion lexica can also be automatically estimated through word-level approaches. Large scale corpora allow the log-linear estimation of adjective semantic orientation from conjunctions [62], as well as through Pointwise Mutual Information in cocurrences [63] and search results [64] with reference terms. Alternative approaches use the annotation of sentiment concepts from a common sense database to map text into Ekman's affect classes [65], or apply fuzzy rules from a lexicon including a wide variety of affect classes [66].

One of the main caveats of using simple word frequency counts is the oversimplification of sentiment as an average of word valences at the word level. As explained by Das and Smith [40], unsupervised sentiment analysis can calculate more precise sentence-level classifications through conceptual dependencies, inspired in the concept of language as a semantic processor that composes the meaning of a sentence [67]. This is implemented through contextual valence shifters: rules that detect changes in valence by identifying negations, amplifications, etc. [68]. This way, lexica of annotations at the word level, such as the MPQA lexicon [69], are applied in sentence-level classifi-

cation through valence shifters [54]. Improving sentence-level unsupervised classification is the aim of a wide variety of unsupervised sentiment analysis methods (see [70] for a more in depth review on this subtopic), and it has led to three of the best performing methods in recent independent benchmarks [38]: SentiStrength [71], VADER [72], and Umigon [73].

One of the main open questions in sentiment analysis is the inclusion of ambiguity, taking into account the variance of sentiment perception. Das and Chen [40] already pointed to inter-coder variability, as two raters of stock-related posts disagreed 28% of the time. Subasic and Huettner [66] proposed to apply fuzzy logic to cope with ambiguity. Further research included additional evaluation metrics inspired in inter-coder reliability [71], and at the important role of high-quality annotations for training data [74]. The task of sentiment analysis is not one of just maximizing accuracy, but also validity, i.e., not having only a precise approximation but also to capture human variability in the perception and expression of sentiment. Recent research has shown that standard unsupervised sentiment analysis methods are consistent with the ratings provided by the authors of classified texts [75]. Yet sentiment analysis has some limitations; a recent study showed that the aggregation of LIWC and VADER scores on Twitter and Facebook posts over the span of months are weakly correlated with trait emotionality and general emotion measures in questionnaires [76]. This illustrates the importance of validating statistical measures based on sentiment analysis, testing if aggregated statistics are sufficiently correlated with standard questionnaire measurements.

Perhaps the most promising development in sentiment analysis is the application of deep learning. Representation learning can leverage large scale datasets to compute word embeddings that are relevant for sentiment analysis, producing automatically extended lexica [77]. While the inference of word categories based on deep learning methods is achieving results very close to those of human annotators [78], recent work found that extrapolating word sentiment continuous variables based on word embeddings still requires significant work [79]. Deep Recurrent Neural Networks have been applied to the task of subjectivity detection [20], and word vector representations can combine supervised and unsupervised learning when applied to sentiment analysis [80]. In terms of Twitter sentiment analysis competitions, so far the most notable achievements of deep learning are some of the top positions in various SemEval competitions [81, 49], which have applications to languages other than English [82]. Currently, one of the most widely used packages for deep learning is based on tree bank annotations and trained on movie review datasets [43], and is included in the Stanford CoreNLP suite⁶.

The problem of sentiment analysis in text is far from solved, but the developments of the last years make the application of sentiment analysis a reality. We must nevertheless acknowledge that there is no universal solution yet, and the performance of tools varies widely depending on the context, formality, and type of text that is being analyzed [38]. Domain knowledge and validation are still necessary ingredients in the application of sentiment analysis, and only a careful selection of tools can guarantee a correct and valid approach to quantifying sentiment from text.

3.2. Sentiment Analysis in Speech

Analysis of speech in search of emotional and affective cues has a comparably long tradition [83]. In the meanwhile, a rich body of literature has been established, including a range of recent surveys such as [84].

However, targetting explicitly *sentiment* exclusively from spoken utterances is a comparably young field. Focusing on the acoustic side of spoken language, the border between sentiment and emotion analysis is often very weak, as, e. g., in [85]. In [21], Mairesse, Polifroni, and Di Fabrizio focus on pitch-related features and observed that also without textual cues, pitch contains information on sentiment.

A number of further works focus on sentiment analysis exclusively from the textual content as present in the speech. For example, Costa Pereira *et al.* [15] proposed using sentiment analysis in speech for information retrieval. Their proposed approach takes a spoken query and retrieves documents whose opinions resemble the query. Similarly, Pérez-Rosas, and Mihalcea [86] focus on the linguistics of spoken reviews after using speech recognition. Kaushik *et al.* [16] and its extension [17] observe that sentiment analysis on natural spontaneous speech data can be realised even when faced with low word recognition rates – a trend that has been noticed also in the recognition of valence from spontaneous speech by Metze and colleagues [87].

3.3. Visual Sentiment Analysis

While there have been related lines of research in vision-based emotion recognition for some time, e.g., [25, 24], conducting sentiment analysis by computer vision is a relatively recent area of research. The principal research tasks in “visual sentiment analysis” revolve around modeling, detecting and leveraging sentiment expressed by means of facial or bodily gestures or sentiment associated with visual multimedia. In the former, the aim is to model and detect sentiment from visually-observable expressions of the sentiment displayed by an individual. And in the latter, the aim is to detect the sentiment that visual multimedia like an image expresses as intended by its author or evokes in human observers.

While automatic extraction of facial expression and bodily gestures from image/video recordings of individuals (and groups) is a well established research area (e.g.,

⁶<http://stanfordnlp.github.io/CoreNLP>

see [88, 89, 90, 91] for recent work on the topic), the same cannot yet be claimed for vision-based sentiment analysis of non-verbal expressions of sentiment. A small number of works have investigated multimodal sentiment analysis on vlogs and reviews [10], on video recordings [92, 93, 94], and from visual behavioral displays [95, 96]. *Visual sentiment* manifests itself frequently in our daily interactions with the visual world, from moments when we harbor that disappointment at seeing our favorite sports team lose a game on livestream video, to moments of amazement when struck by the beauty of an intricate painting or captivating photograph. Such sentiments are birthed out of our wealth of experiences, dispositions and opinions and, in the case of visual sentiment, manifest as a result of the interactions we have with visual elements. And so, given that sentiment is always towards an object or entity, visual sentiment likewise is defined toward an object, scene or event present in the visual content. For example, an image showing delicious food, a beautiful garden or an exotic wedding likely expresses the positive sentiment held by the publisher of the image; and by experiencing these images, a positive sentiment may be triggered in viewers. When these sentimental experiences are distilled into a set of say, semantic labels, we can then construct computer vision problems to learn functional mappings from low-level visual multimedia, i.e., the raw pixels, motion, etc., to the high-level sentiment labels in classification, localization and summarization tasks.

Among the earliest work in visual sentiment analysis, Wang *et al.* [97] investigated adjective associations organized into 12 adjective-adjective word pairs over 100 images annotated by 42 subjects. They used a variety of color features including lightness, saturation and sharpness features in conjunction with support vector regression to predict the presence of these pairs like *warm-cool*, *brilliant-gloomy*, and *vibrant-desolate*. In [98], Yanulevskaya *et al.* proposed “holistic” image features, composed of codebooks over local color histogram and Gabor features, in conjunction with support vector machines (SVMs) for sentiment prediction on an image dataset popularly used in psychology studies called the International Affective Picture System [99]. Later, Siersdorfer *et al.* [100] proposed the use of global and local color histograms with SIFT [101] features with SVMs on a much larger dataset of 586k social images for sentiment analysis using weak sentiment labels assigned using SentiWordNet [102] on image tags, and Jia *et al.* [103] proposed using similar color features but using a specialized graphical model instead of SVMs for experiments over 23k digital images of paintings with adjectives like *pretty*, *causal*, *romantic* and *jaunty*.

Around the same time, attribute learning and mid-level feature representations began to rise in popularity in computer vision, e.g., ObjectBank [104]. And as a result, several mid-level sentiment representations were proposed for visual sentiment analysis. The challenge in visual sentiment analysis up until this point was that the visual variance was far too wide and often resulted in training

instances clustering poorly in feature spaces; for example, consider the possible visual spectrum for the appearance of a *cup* or *laptop* compared to sentiment semantics like *positive* or *warm*. The hope with a mid-level representation approach was that a proxy representation that could be detected in vision systems with higher fidelity could also be used to aid visual sentiment analysis. In [105], Yuan *et al.* proposed the use of 102 mid-level features they called SentiBank for sentiment which was simply derived from scene attributes in the SUN Attribute dataset [106], i.e. semantic attributes like *still water*, *ice* and *hiking* were included. Whereas these were mostly noun concepts, in [107, 14], Borth *et al.* proposed a set of visual classifiers forming a mid-level representation called SentiBank. The representation consisted of a set of 1200 linear SVM outputs ($F1 = 0.6$) where the SVMs were trained using a taxonomy of a semantic construct called adjective-noun pairs (ANPs), i.e., versus the adjective-adjective construct of [97]. The proposed ANPs combined a “noun” for visual detectability and an “adjective” for sentiment modulation of the object described by noun semantics, resulting in pairs like *cute dog*, *beautiful sunset*, *disgusting food* and *terrible accident*.

The adjective-noun pair mid-level representations of [14] lowered the entry bar for a slew of computer vision methods and went on to drive a number of applications in animated image sequences [108, 109], aesthetics analysis [26], emotion analysis [108, 110], automatic content curation [111, 112, 113] and others. These ANPs proposed in [14] were formed by first using seed keyword from *Plutchik’s Wheel of Emotion* [114] to query Flickr⁷ and YouTube’s API for images and video. By mining tags associated with these visual data and then performing part-of-speech tagging to identify adjectives and nouns, a pool of adjective-noun pair candidates could be formed by pairwise combinations. The ANP candidates are then filtered by sentiment strength, named entities, and popularity before subsampling to avoid too many adjectives being paired to the same noun. This final set of ANPs was called the Visual Sentiment Ontology (VSO) [14]⁸, although actual ontological structures by manual labeling were not actually introduced until later in [115]. These ANPs were also used to mine Flickr images to train the SentiBank detector bank [107] described earlier and also applied toward a sentiment analysis study from Twitter images using crowd-sourced sentiment labels [14]. In [23], Jou *et al.* proposed an ontology with extended breadth and volume in a Multilingual Visual Sentiment Ontology (MVSO)⁹, including 15630 ANPs from 12 major languages and 7.37M images from over 235 countries [116], with an expanded English ANP corpus. MVSO used a similar ANP mining process to VSO, focusing around Flickr performing many more stages of candidate filtering, including diversity of users for

⁷<https://www.flickr.com>

⁸<https://visual-sentiment-ontology.appspot.com>

⁹<http://mvso.cs.columbia.edu>

a given ANP, semantic correctness and language-specific syntax. In addition, they also explored generating ontology structures automatically by either grouping ANP by exact translations with a pivot language or approximate ANP groupings using word embeddings [23, 117]. These multilingual ANPs and their associated detector banks have found applications in cross-lingual visual sentiment analysis [23], portrait analysis [117] and image query expansion diversification [118].

More recently, with the success of convolutional neural networks (CNNs) [119], spurred on by the performance of AlexNet [120] on image classification in the ImageNet Large Scale Visual Recognition Competition [121], there has also been a rush to apply CNNs, and more broadly neural networks, in visual sentiment analysis. Xu *et al.* [122] used an AlexNet model pre-trained on ImageNet simply as a feature extractor [123] with SVMs and logistic regression classifiers, while You *et al.* [124] fine-tuned the same AlexNet model, both for visual sentiment detection on a set of Twitter and Tumblr images. Campos *et al.* [125, 126] explored fine-tuning in combination with SVMs and logistic regression classifiers, rather than disjointly, for visual sentiment analysis on the same set of Twitter images as in [124] with 6.1% absolute accuracy improvement. For adjective-noun pair detection, Chen *et al.* [127] proposed an AlexNet model fine-tuned from an ImageNet model for an expanded set of 2089 ANPs in VSO [14] with a top-5 accuracy of 19.1%. Jou *et al.* [23] fine-tuned from DeepSentiBank to get six AlexNet-styled networks for ANP detection in English, Spanish, Italian, French, German and Chinese, achieving a top-5 accuracy of 21.7% on 4342 ANPs for the largest corpus, English. Recently, Narihira *et al.* [128] and Jou *et al.* [22] also later proposed custom multitask network structures for VSO [14] that incorporate both adjective-only and noun-only detection. And Mathews *et al.* [129] proposed coupling a CNN with Long Short-Term Memory (LSTM) in a network called SentiCap for sentimentally biasing visual captioning of images, allowing images to now be summarized by short sentences that include both noun-based object groundings and affective adjective qualifiers.

All these work in furthering and applying visual sentiment analysis point to the potential in the higher accuracy techniques, like with CNNs [23, 125, 126, 124, 127, 128, 22, 129], as well as increased coverage, like with multilingual [23, 125, 117, 22] and multiple content source methods [14, 115, 124]. And with the increasing number of publicly available computer vision models/libraries and visual sentiment datasets, visual sentiment analysis is poised to see growth in both of these directions. Even so, the multifaceted nature of sentiment indicates that visual sentiment analysis alone will not be able to fully measure and/or describe our experiential disposition and opinions in multimedia data. For example, visual content might not be able to understand the context or extract the entity.

3.4. Multimodal Sentiment Analysis

Multimodal sentiment analysis has been only addressed very recently, with only a handful of notable work mostly focusing on analyzing sentiment in vlogs. To the best of our knowledge, Morency *et al.* [10] were the first to consider multimodal sentiment analysis. They proposed analyzing audiovisual content in addition to text for sentiment analysis. They collected 47 videos depicting a monologue. They manually selected 30 second excerpts that only cover one topic and transcribed the videos manually. Every video received three sentiment labels, namely, positive, neutral and negative. In total, the dataset contains 498 excerpts, each containing one sentence. A qualitative analysis on the short dataset demonstrated that the following features are associated with the expressed sentiment: word polarity (language); smiling and looking away (visual); pause and pitch (audio). Pitch and speech pause were extracted using openEAR [130]. The authors measures durations for smiling and looking away using a commercial facial expression analysis software. Transcribed speech was analyzed to spot the lexicons with positive or negative polarity according to [69]. Hidden Markov Models were utilized for sentiment classification, which took utterance-level trimodal features as input. Overall, trimodal sentiment detection outperformed the unimodal ones, and an average F1 score of 0.55 was achieved in a leave-one-out cross-validation. Even though the dataset was small and the text analysis method was simple, this work demonstrated the potential strength of multimodal sentiment analysis. In an extension of their work, Poria *et al.* [28] explored performing sentiment analysis on videos using a combination of facial expression, audio data such as vocal pitch, and textual features from uttered sentences. To extract text features, they used text2vec [19], which in addition to part-of-speech features were used to train a deep CNN. They, however, did not use the deep CNN for sentiment analysis but rather for feature extraction by removing the softmax layer and replacing it with a SVM. Facial fiducial landmarks were extracted from faces. Audio features were extracted using openSMILE [131]. Due to the low number of samples in the neutral class, they discarded the neutral samples. The best modality was again text, and multimodal fusion reached an accuracy of 88.6% for two-class classification (negative vs. positive). More recently, Poria *et al.* [132] further extended this work using a more extensive set of features including a larger number of audio features and, for text, features derived from the sentic computing paradigm. A larger number of low-level audio features were extracted. For text, they used sentic computing paradigm [133], which evaluates higher-level linguistic concepts beyond the statistical descriptors offered by, for example, bag-of-words representations. Using these features, the authors were able to improve the sentiment detection accuracy on three classes up to $F1 = 0.78$.

Wölmer *et al.* [11] attempted multimodal sentiment analysis in movie reviews on online user-generated videos. A dataset of 370 videos depicting a monologue reviewing

a movie was collected from ExpoTV¹⁰ and YouTube. Out of these 370 videos, 228 were positive, 23 were neutral and 119 were negative according the annotators' labels and original ratings on ExpoTV. For analyzing the text, the reviews were both transcribed manually and using automatic speech recognition (ASR). The authors further performed cross-domain analysis by training on a large corpus of written reviews. The large-scale corpus, in total, contained 102622 written reviews for 4901 movies and was collected from Metacritic¹¹. They also explored the application of online knowledge sources (OKS), i.e., WordNet [134], ConceptNet [135], and General Inquirer [136] for inferring the speaker's sentiment. Unigram and trigram bag-of-words feature representation were extracted after Porter stemming and removing stop words. Facial expressions were analyzed, and smile intensity, gaze direction and head pose were extracted as features. Acoustic low-level descriptors (LLD) were extracted by openSMILE [131] from the audio channel. Features were pooled at utterance-level. Linguistic features were used to train a linear SVM, and audiovisual features were used to train a Bidirectional Long-Short-Term-Memory (BLSTM) recurrent neural network [137]. The cross-corpus n-gram analysis trained on the Metacritic database led to the best classification accuracy ($F1 = 0.73$). The domain-specific text analysis achieved a similar performance. Interestingly, audiovisual analysis without any text analysis did not fare much worse, with accuracy dropping to $F1 = 0.66$. Given that the content was a spoken review, speech prosody appeared to be the most informative channel for improving the performance of multimedial fusion when added to the text-based method.

Pérez Rosas *et al.* [12] analyzed 105 videos from YouTube in Spanish-language depicting a person expressing opinions on different topics. Videos were manually segmented into 30 seconds segments, each covering a single topic. Videos were labeled manually in three classes of negative, neutral and positive; however, only 4 videos were labeled as neutral class. The videos were manually transcribed to analyze the text. A bag-of-words representation was reconstructed, and words with frequency below 10 were discarded. The unigram feature set was used as the feature vector representing the text. Similar to [10], smile duration and looking away were detected using a commercial facial expression analysis software. Pause duration, pitch, intensity and loudness were extracted from the audio track. Text modality performed the best, reaching an accuracy of nearly 65%. Visual modality and, particularly, smile duration were found to be important for sentiment detection. The multimodal early fusion increased the accuracy up to 75%. This work showed the potential of using multimodal sentiment analysis and particularly language-independent modalities such as visual modality on a dataset collected in a language other than English. However, the size of

the dataset is limited, and there is still room to perform a more in-depth analysis for each modality and modality fusion strategies.

Zadeh [27] proposed to detect sentiment intensity from text, audio and visual modalities. He constructed a dataset of 93 vlog posts from YouTube in English. The videos were manually transcribed. Subjectivity was determined at the sentence level. Subjectivity was defined as an expression of a private state using three rules, namely, explicit mention of a private state, e.g., "*I also love the casting of Mark Strong as Sinestro.*"; mentioning a private state, e.g., "*Shia LaBeouf said that the second movie lacked um heart.*"; and an implicit reference to an opinion, e.g., "*I would never recommend watching this movie.*" Only video segments containing subjectivity were annotated on a seven-point scale (from -3 to +3), and each video received five labels. In total, 2199 excerpts contained subjectivity. A set of features were extracted from audio, visual and text modalities. From text, ngrams (up to three) were extracted. Simple audio features, e.g., MFCC and peak slope were extracted from audio. Facial action units, facial landmarks and head pose were extracted from the visual modality. He first trained a model to recognize subjective from objective sentences. His experiment showed that multimodal approach outperformed unimodal ones for recognizing subjectivity using a SVM. His preliminary analysis yielded significant results from visual and text modalities for sentiment intensity detection ($\rho = 0.49$). Even though the work was presented in a doctoral consortium and only preliminary results were included, it demonstrated two novelties. First, it showed the effectiveness of multimodal analysis in pre-processing related to sentiment analysis, i.e., subjectivity recognition. Second, it dealt with sentiment at multiple levels instead of simple positive vs. negative classification.

Ellis *et al.* [138] performed multimodal sentiment analysis on broadcast video news. They collected a dataset of 929 sentence-length excerpts which were annotated on Amazon Mechanical Turk¹². Text and multimedia content were separately annotated on three levels of negative, neutral and positive. They first showed a sentence that was spoken on the news and asked the Turkers to label it. They subsequently showed the video in which the same sentence is spoken, to also be labeled. They found that in about 21.5% of cases, the sentiment labels between the transcription and multimedia content differed. They extracted low-level descriptors using openSMILE [131] from the audio tracks. They then identified whether the speaker's face was depicted in the frames before extracting features that are related to facial expressions, e.g., dense Local Binary Patterns (LBP) histograms. Analysis of the sentiment in text was performed using an off-the-shelf solution [139] that reached an accuracy of $\sim 46\%$ for three-class classification. They found that television anchors

¹⁰The original site is no longer active at the time of writing.

¹¹<http://www.metacritic.com>

¹²<http://www.mturk.com>

have a unique pattern for expressing sentiment, and by training person-specific models, they increased an accuracy by $\sim 12\%$ with the visual modality and $\sim 6\%$ with the audio modality. Within modalities, anchor-specific audio sentiment detection performed the best with the accuracy of 62.56% followed by anchor-specific visual sentiment detection at 56.85%. Their findings demonstrated the significance of multimodal analysis for multimedia content in both understanding its polarity and its automatic sentiment analysis.

There are a number of notable work that do not directly address or mention sentiment but are closely related. McDuff *et al.* [140] presents an example of using facial expression analysis to assess the candidate preferences of American voters. They collected 611 responses to five video clips from a US presidential election debate. They demonstrated that voter preference (or opinion) can be determined with an accuracy of 73% using only the facial expressions in response to those videos. Madzlan *et al.* [141] analyzed vlogs for automatic identification of attitudes. Attitudes are defined by the authors as social affective states that vloggers intend to convey. They used speech prosody and facial expressions to automatically recognize amusement, impatience, friendliness, enthusiasm and frustration. They found pitch to be the most important audio feature, followed by intensity and voice quality. Siddiquie *et al.* [142] used multimodal analysis to detect propaganda videos. They used audiovisual affective analysis in addition to sentiment analysis in the comments to identify politically persuasive videos online.

In summary, the burgeoning field of multimodal sentiment analysis shows great promise in accurately capturing the real essence of expressed sentiments. Even in the complete absence of text, most of the methods are able to identify sentiment fairly accurately due to the affective nature of sentiment. However, audiovisual methods are only effective in understanding sentiment's polarity; entity aspect extraction and subjectivity recognition still remains in the domain of text analysis.

4. Applications of Sentiment Analysis

Sentiment analysis in language is being commercially used to summarize reviews and customer opinions. We are not only able to aggregate the opinions at scale, but also get that feedback immediately at low cost. Before sentiment analysis, companies had to either perform surveys or create focus groups, which was much slower and much more expensive. With the emergence of opinions posted in multimedia on social media, e.g., spoken reviews on YouTube, sentiment analysis has the ability to become an increasingly crowd-sourced and low-cost endeavor.

Sentiment we hold towards an object or a person have an impact on our interpersonal relationship and interactions. Therefore, sentiment analysis can be used for enhancing human-machine and human-human interactions. The work of Langlet and Clavel [143, 13] is an example

of utilizing automatic sentiment analysis in human-agent interaction. They argue that if a user and the ECA share a sentiment towards an entity, then the human is more likely to find the ECA likable. ECAs are finding their way in many different applications, from online education to customer service. They propose to use agent's utterances to extract the expressions of like or dislike from the user. Langlet and Clavel annotated statement pairs (i.e., an ECA statement followed by a human response) to contextualize the likes and dislikes of the user. In their more recent work [143], they also added the ability to extract topic words from each user's speech turn to identify the relevant entities. Their proposed approach is purely based on language analysis which requires an ASR to transcribe the speech from the user highly accurately. Thus, the scope and methodology of their work is similar to text-based sentiment analysis

The current work on visual sentiment analysis including SentiBank [14] found its way into many multimedia content analysis work [144, 145] that are not directly linked to sentiment analysis. However, the ANPs were rather used as higher-level concepts and attributes for automatic description of the visual content.

A new domain where sentiment analysis is finding its way is multimedia analytics. For example, the work of Ellis *et al.* [138] utilizes multimodal sentiment analysis on broadcast video news which can be used for automatic analysis and summarization of TV programs. Multimodal sentiment analysis technologies can be also used to identify politically persuasive content [142]. These technologies will make it possible to mine opinions expressed through countless broadcast television channels or online channels on the Internet.

5. Challenges and Perspectives

5.1. Challenges

5.1.1. Methodological challenges

Since most of the current sentiment analysis is data-driven, the capacity of machine learning models is limited to a specific domain where the training data come from. Domain adaptation is an open issue that needs to be addressed, for example, adapting a model trained on sentiment analysis in product reviews for analyzing microblog posts. Other important challenges of sentiment analysis include how to handle ambiguous situations and irony. For example, a sarcastic comment praising an object intends to convey a negative sentiment; however, conventional sentiment analysis methods often incorrectly interpret such expressions. A number of methods have been proposed to identify sarcasm in language [4]. The problem is far from solved, however, because humor is culture-specific, and it is very challenging for a machine to learn unique (and often quite specific) cultural references. We argue that multimodal sentiment analysis can be made more successful in identifying the sarcastic comments by taking advantage

of vocal and facial expressions. Moreover, people express sentiment for social reasons that are not related to their internal dispositions. For example, a person might express like or dislike sentiments to conform with a certain cultural norm or to express and differentiate his/her identity. Finally, machine-based sentiment analysis is limited to the external manifestations of sentiment, and we do not have the ability conclusively determine an individual's unexpressed viewpoints.

Sentiment analysis can be carried out on two different types of data, each with their own issues. Sentiment analysis on human-machine and human-human interactions requires datasets that are very similar to the ones used in emotion recognition. Therefore, it faces the same problems of limited size and uncertain ground-truth. Recently, McDuff *et al.* [146] demonstrated how a large number of emotional responses including sentiment can be captured using webcams over the Internet. Although this limits the quality of audiovisual capture, these techniques provide the scale that is not accessible in a laboratory. There is also the issue of labeling private data recorded in the laboratory, which limits the tedious task of labeling to people who are authorized to access the data. As a result, we are not only bounded by the amount of data we can record in the laboratory but also by a limited ability to label large amount of data.

The second source of data containing multimodal sentiment is multimedia content on social media. Social media is a rich resource of data that provides us with scale. The problem is that the quality and the context of the recorded material can vary, and the data is limited to certain demographics that are more represented on the Internet. However, since the data is public, it can be easily labeled through crowdsourcing.

Looking at the existing work on multimodal sentiment analysis, it seems people are more likely to express positive or negative opinions, and as a result, there is a lack of neutral opinions expressed online in all the reviewed multimodal sentiment analysis studies. This can be possibly addressed by automatically detecting the subjective expressions that are not emotionally charged.

Even though multimodal sentiment analysis is showing promising results, the core part of sentiment analysis remains a text analysis problem. Entity extraction, aspect of entity extraction and identification of the holder of opinion can be only tackled by natural language processing. We envision that multimodal sentiment analysis can add a new dimension and improve the existing sentiment analysis techniques but cannot completely replace them.

Given the significance of the problem, construction of a large-scale publicly available benchmark with permissive license such as the ones developed in MediaEval is desirable [147].

5.1.2. Ethics

Sentiment is a private state and mining a private state of our own person raises legitimate ethical concerns. Ma-

chines that are able to go beyond individual human intelligence in understanding our own opinions and attitudes has incredible potential in areas like mental healthcare, but also is cause for questions around privacy.

Sentiment analysis on social media, as a data-driven technique, may introduce a bias in decisions or higher level analytics. For example, sentiment analysis that only considers users of a single social media platform, e.g., Twitter, might magnify the importance the demographics of such a platform. Suppose companies or even political parties relied on social media analytics to assess the importance of certain policy decisions and implementations. For example, if more white males expressed emotionally charged or strong opinions on Twitter, companies or organizations using sentiment analysis tool may be more likely to listen to them given the agnosticism with which machine learning tools often treat data [148]. Moreover, data-driven methods can learn the language of the dominant demographics, further undermining the opinion expressed by some individuals and people groups.

Likewise, automatic sentiment analysis can be also a tool for limiting the freedom of speech. The social web provides an open platform for people to express and share their opinions. However, sentiment analysis can become a tool for oppressive regimes to identify dissents or apply censorship at scale [149]. Similarly, hate speech, racist comments and/or malicious propaganda can be identified with automatic sentiment analysis and either promoted or suppressed in response. And throughout all of this, since at its core, machine-based data-driven techniques form the core of modern sentiment analysis, it is still prone to errors which may ultimately result in ill-informed decisions and poor consequences.

5.2. Perspectives

The Internet has moved from a principally text-based communication medium to one of widespread multimedia. The hope and charge for multimodal sentiment analysis is to both integrate across many modalities for sentiment understanding as well as complement tasks that have traditionally been isolated to single modes, e.g., text-based subjectivity analysis [27]. Given the 'wildness' of Web data, we expect a key area that multimodal sentiment analysis will distinguish itself is in the presence of missing or incomplete data, especially given the volatility and varied veracity found in Internet sources. Emotion recognition, being at the heart of sentiment analysis, also stands to gain much in robustness and reliability from moving toward multimodal emotion recognition [150].

One promising new avenue for sentiment analysis is its usage in human-human and human-ECA interaction. Avatars and virtual agents are appearing in many domains, and improving the quality of their interaction will be of great interest. The growing interest in this domain will bring new challenges and problems which can be addressed by researchers from NLP as well as affective computing communities. Currently, aspect, entity and opin-

ion holder extraction remains solely in the domain of text analysis. However, the recent advances in computer vision and concept-detection will bring new opportunities in the automatic identification of the opinion holder, e.g., face recognition and aspect and entity recognition, including visual object recognition. Sentiment-specific emotion recognition technologies shall be also further developed to automatically identify behavioral patterns associated with sentiment traces rather than the current universal emotion recognition tools.

6. Conclusions

In this paper, we presented an overview of the concept and goals of multimodal sentiment analysis, reviewed the state of the art, and discussed challenges and perspectives related to the field. A growing body of work published in the last half decade has demonstrated the great strides and promise of multimodal sentiment analysis. Our review of the existing literature demonstrates that multimodal sentiment analysis is a promising approach to leverage complementary channels of information for sentiment analysis and often outperforms the unimodal methods. It also holds the potential to enhance other tools that currently benefit from unimodal sentiment analysis, such as entity recognition and subjectivity analysis. We hope the review encourages further cross-disciplinary efforts in this emerging field.

7. Acknowledgments

The work of Soleymani is supported by his Ambizione grant from the Swiss National Science Foundation. The work of Pantic and Schuller is partially supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA). We would like to thank Julien Deonna and Cristina Soriano for invaluable discussions on the definition of sentiment.

References

- [1] J. Deonna, F. Teroni, *The emotions: A philosophical introduction*, Routledge, 2012.
- [2] B. J. Jansen, M. Zhang, K. Sobel, A. Chowdury, Twitter power: Tweets as electronic word of mouth, *Journal of the American Society for Information Science and Technology* 60 (11) (2009) 2169–2188. doi:10.1002/asi.21149.
- [3] P. Melville, W. Gryc, R. D. Lawrence, Sentiment analysis of blogs by combining lexical knowledge with text classification, in: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009, pp. 1275–1284.
- [4] B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis, in: *Mining Text Data*, Springer U.S., 2012, pp. 415–463. doi:10.1007/978-1-4614-3223-4_13.
- [5] B. Pang, L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval* 2 (12) (2008) 1–135. doi:10.1561/1500000011.
- [6] E. Cambria, B. Schuller, Y. Xia, C. Havasi, New avenues in opinion mining and sentiment analysis, *IEEE Intelligent Systems* 28 (2) (2013) 15–21. doi:10.1109/MIS.2013.30.
- [7] S. Asur, B. A. Huberman, Predicting the future with social media, in: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2010, pp. 492–499. doi:10.1109/WI-IAT.2010.63.
- [8] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, *Journal of Computational Science* 2 (1) (2011) 1–8. doi:10.1016/j.jocs.2010.12.007.
- [9] A. Tumasjan, T. Sprenger, P. Sandner, I. Welp, Predicting elections with Twitter: What 140 characters reveal about political sentiment, in: *International AAAI Conference on Web and Social Media*, 2010, pp. 178–185.
- [10] L.-P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis, in: *ACM International Conference on Multimodal Interfaces (ICMI)*, New York, New York, USA, 2011, p. 169. doi:10.1145/2070481.2070509.
- [11] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, L.-P. Morency, YouTube movie reviews: Sentiment analysis in an audio-visual context, *IEEE Intelligent Systems* 28 (3) (2013) 46–53. doi:10.1109/MIS.2013.34.
- [12] V. Pérez Rosas, R. Mihalcea, L. P. Morency, Multimodal sentiment analysis of Spanish online videos, *IEEE Intelligent Systems* 28 (3) (2013) 38–45. doi:10.1109/MIS.2013.9.
- [13] C. Langlet, C. Clavel, Adapting sentiment analysis to face-to-face human-agent interactions: From the detection to the evaluation issues, in: *IEEE Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 14–20. doi:10.1109/ACII.2015.7344545.
- [14] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, Large-scale visual sentiment ontology and detectors using adjective noun pairs, in: *ACM International Conference on Multimedia*, 2013, pp. 223–232. doi:10.1145/2502081.2502282.
- [15] J. C. Pereira, J. Luque, X. Anguera, Sentiment retrieval on web reviews using spontaneous natural speech, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4583–4587. doi:10.1109/ICASSP.2014.6854470.
- [16] L. Kaushik, A. Sangwan, J. H. Hansen, Sentiment extraction from natural audio streams, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8485–8489.
- [17] L. Kaushik, A. Sangwan, J. H. L. Hansen, Automatic sentiment extraction from YouTube videos, in: *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 239–244. doi:10.1109/ASRU.2013.6707736.
- [18] C. Clavel, Z. Callejas, Sentiment analysis: From opinion mining to human-agent interaction, *IEEE Transactions on Affective Computing* (2015) 74–93. doi:10.1109/TAFFC.2015.2444846.
- [19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* 26, Curran Associates, 2013, pp. 3111–3119.
- [20] O. Irsoy, C. Cardie, Opinion mining with deep recurrent neural networks, in: *EMNLP*, 2014, pp. 720–728.
- [21] F. Mairesse, J. Polifroni, G. Di Fabbrizio, Can prosody inform sentiment analysis? Experiments on short spoken reviews, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5093–5096.
- [22] B. Jou, S.-F. Chang, Deep cross residual learning for multi-task visual recognition, in: *ACM International Conference on Multimedia (MM)*, 2016.
- [23] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, S.-F. Chang, Visual affect around the world: A large-scale multilingual visual sentiment ontology, in: *ACM International Conference on Multimedia (MM)*, 2015, pp. 159–168. doi:10.1145/2733373.2806246.
- [24] Z. Zeng, M. Pantic, G. I. Roisman, T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Transactions on Pattern Analysis and Ma-*

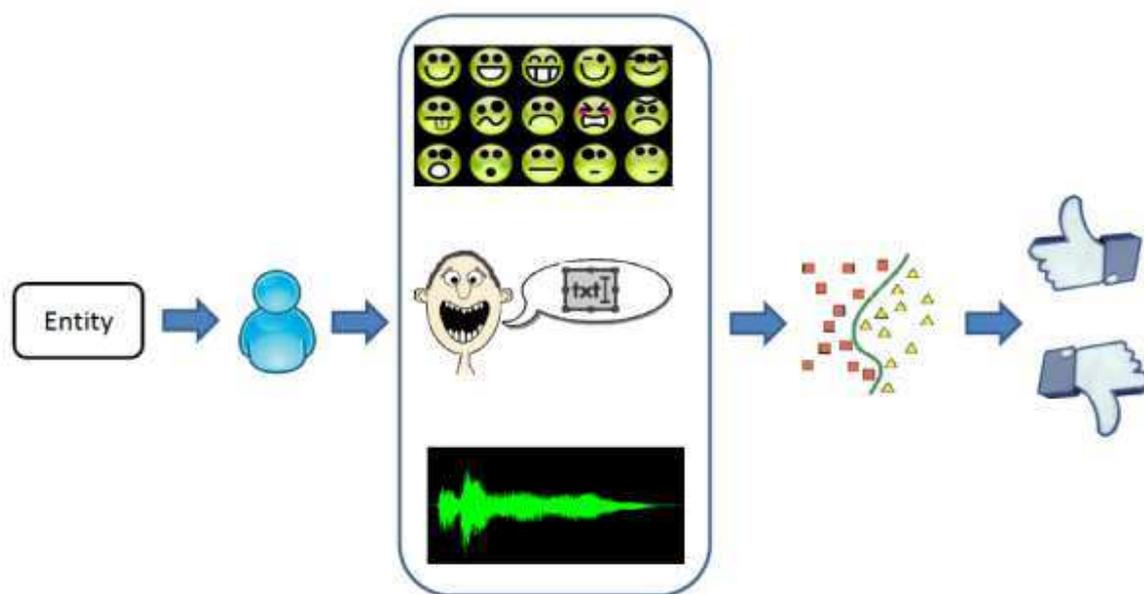
- chine Intelligence 31 (1) (2009) 39–58. doi:10.1109/TPAMI.2008.52.
- [25] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: A survey of registration, representation, and recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (6) (2015) 1113–1133. doi:10.1109/TPAMI.2014.2366127.
- [26] S. Bhattacharya, B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, M. Shah, Towards a comprehensive computational model for aesthetic assessment of videos, in: *ACM International Conference on Multimedia (MM)*, 2013. doi:10.1145/2502081.2508119.
- [27] A. Zadeh, Micro-opinion sentiment intensity analysis and summarization in online videos, in: *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 587–591. doi:10.1145/2818346.2823317.
- [28] S. Poria, E. Cambria, A. Gelbukh, Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, in: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 2539–2544.
- [29] K. Yatani, M. Novati, A. Trusty, K. N. Truong, Analysis of adjective-noun word pair extraction methods for online review summarization, in: *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Vol. 22, 2011, p. 2771.
- [30] M. D. Munezero, C. S. Montero, E. Sutinen, J. Pajunen, Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text, *IEEE Transactions on Affective Computing* 5 (2) (2014) 101–111. doi:10.1109/TAFFC.2014.2317187.
- [31] K. R. Scherer, What are emotions? And how can they be measured?, *Social Science Information* 44 (4) (2005) 695–729. doi:10.1177/0539018405058216.
- [32] D. Sander, D. Grandjean, K. R. Scherer, A systems approach to appraisal mechanisms in emotion, *Neural Netw.* 18 (4) (2005) 317–352. doi:10.1016/j.neunet.2005.03.001.
- [33] K. R. Scherer, A. Schorr, T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*, Oxford University Press, 2001.
- [34] S.-M. Kim, E. Hovy, Determining the sentiment of opinions, in: *Proceedings of the 20th International Conference on Computational Linguistics, COLING '04*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004. doi:10.3115/1220355.1220555. URL <https://doi.org/10.3115/1220355.1220555>
- [35] J. R. Fontaine, K. R. Scherer, E. B. Roesch, P. C. Ellsworth, The world of emotions is not two-dimensional, *Psychological Science* 18 (12) (2007) 1050–1057. doi:10.1111/j.1467-9280.2007.02024.x.
- [36] M. A. Hearst, Direction-based text interpretation as an information access refinement, *Text-based intelligent systems: Current research and practice in information extraction and retrieval* (1992) 257–274.
- [37] W. Sack, On the computation of point of view, in: *AAAI*, 1994, p. 1488.
- [38] F. N. Ribeiro, M. Araújo, P. Gonçalves, F. Benevenuto, M. A. Gonçalves, A benchmark comparison of state-of-the-practice sentiment analysis methods, *arXiv preprint arXiv:1512.01818*.
- [39] J. M. Wiebe, R. F. Bruce, T. P. O'Hara, Development and use of a gold-standard data set for subjectivity classifications, in: *Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 246–253. doi:10.3115/1034678.1034721.
- [40] S. R. Das, M. Y. Chen, Yahoo! for Amazon: Sentiment parsing from small talk on the web, *For Amazon: Sentiment Parsing from Small Talk on the Web*.
- [41] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: Sentiment classification using machine learning techniques, in: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 10, 2002, pp. 79–86.
- [42] K. Dave, S. Lawrence, D. M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, in: *International Conference on World Wide Web (WWW)*, 2003, pp. 519–528. doi:10.1145/775152.775226.
- [43] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 1631, 2013, p. 1642.
- [44] P. Nakov, T. Zesch, D. Cer, D. Jurgens (Eds.), *International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, 2015. URL <http://www.aclweb.org/anthology/S15-2>
- [45] S. M. Mohammad, S. Kiritchenko, X. Zhu, Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets, *Atlanta, USA*, 2013.
- [46] Y. Miura, S. Sakaki, K. Hattori, T. Ohkuma, TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data, in: *International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 628–632.
- [47] M. Hagen, M. Potthast, M. Büchner, B. Stein, Webis: An ensemble for Twitter sentiment detection, in: *International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, 2015, pp. 582–589. URL <http://www.aclweb.org/anthology/S15-2097>
- [48] P. Nakov, A. Ritter, S. Rosenthal, V. Stoyanov, F. Sebastiani, SemEval-2016 task 4: Sentiment analysis in Twitter, in: *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16*, Association for Computational Linguistics, San Diego, California, 2016.
- [49] J. Deriu, M. Gonzenbach, F. Uzdilli, A. Lucchi, V. De Luca, M. Jaggi, Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision, *Proceedings of SemEval (2016)* 1124–1128.
- [50] A. Abbasi, A. Hassan, M. Dhar, Benchmarking Twitter sentiment analysis tools, in: *LREC*, 2014, pp. 823–829.
- [51] A. Hassan, A. Abbasi, D. Zeng, Twitter sentiment analysis: A bootstrap ensemble framework, in: *IEEE International Conference on Social Computing (SocialCom)*, 2013, pp. 357–364.
- [52] J. Blitzer, M. Dredze, F. Pereira, et al., Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification, in: *ACL*, Vol. 7, 2007, pp. 440–447.
- [53] P. J. Stone, *Thematic text analysis: New agendas for analyzing text content*, *Text Analysis for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum (1997) 33–54.
- [54] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: *Conference on Human Language Technology (HLT)*, 2005, pp. 347–354. doi:10.3115/1220575.1220619.
- [55] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of LIWC2015, *UT Faculty/Researcher Works*.
- [56] M. M. Bradley, P. J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, *Tech. rep.* (1999). doi:10.1109/MIC.2008.114.
- [57] P. S. Dodds, C. M. Danforth, Measuring the happiness of large-scale written expression: Songs, blogs, and presidents, *Journal of Happiness Studies* 11 (4) (2010) 441–456.
- [58] A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas, *Behavior research methods* 45 (4) (2013) 1191–1207. doi:10.3758/s13428-012-0314-x.
- [59] J. Redondo, I. Fraga, I. Padrón, M. Comesaña, The Spanish adaptation of ANEW (Affective Norms for English Words), *Behavior research methods* 39 (3) (2007) 600–605.
- [60] M. L. Võ, M. Conrad, L. Kuchinke, K. Urton, M. J. Hofmann, A. M. Jacobs, The berlin affective word list reloaded (bawl-r), *Behavior research methods* 41 (2) (2009) 534–538.
- [61] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, et al., Human language reveals a universal positivity

- bias, *Proceedings of the National Academy of Sciences* 112 (8) (2015) 2389–2394.
- [62] V. Hatzivassiloglou, K. R. McKeown, Predicting the semantic orientation of adjectives, in: Annual meeting of the association for computational linguistics and conference of the European chapter of the association for computational linguistics, 1997, pp. 174–181.
- [63] P. Turney, M. L. Littman, Unsupervised learning of semantic orientation from a hundred-billion-word corpus, Tech. rep., National Research Council Canada (2002). doi:10.4224/8914027.
- [64] P. D. Turney, Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews, in: Annual Meeting on Association for Computational Linguistics, 2002, pp. 417–424. doi:10.3115/1073083.1073153.
- [65] H. Liu, H. Lieberman, T. Selker, A model of textual affect sensing using real-world knowledge, in: ACM International Conference on Intelligent User Interfaces (IUI), New York, New York, USA, 2003. doi:10.1145/604045.604067.
- [66] P. Subasic, A. Huettner, Affect analysis of text using fuzzy semantic typing, *IEEE Transactions on Fuzzy Systems* 9 (4) (2001) 483–496. doi:10.1109/91.940962.
- [67] R. C. Schank, L. Tesler, A conceptual dependency parser for natural language, in: Conference on Computational linguistics, 1969, pp. 1–3.
- [68] L. Polanyi, A. Zaenen, Contextual Valence Shifters, 2006, pp. 1–10. doi:10.1007/1-4020-4102-0_1.
- [69] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, *International Conference on Language Resources and Evaluation (LREC)* 39 (2) (2005) 165–210. doi:10.1007/s10579-005-7880-9.
- [70] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, *Computational Linguistics* 37 (2) (2011) 267–307.
- [71] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, *Journal of the American Society for Information Science and Technology* 61 (12) (2010) 2544–2558.
- [72] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.
- [73] C. Levallois, Umigon: Sentiment analysis for tweets based on lexicons and heuristics, in: International Workshop on Semantic Evaluation (SemEval), Vol. 13, 2013.
- [74] I. Mozetič, M. Grčar, J. Smilović, Multilingual Twitter sentiment classification: The role of human annotators, *PloS one* 11 (5) (2016) e0155036.
- [75] D. Garcia, A. Kappas, D. Küster, F. Schweitzer, The dynamics of emotions in online interaction, arXiv preprint arXiv:1605.03757.
- [76] A. Beasley, W. Mason, Emotional states vs. emotional words in social media, in: ACM Web Science Conference, 2015, p. 31.
- [77] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, Learning sentiment-specific word embedding for Twitter sentiment classification, in: Association for Computational Linguistics, 2014, pp. 1555–1565.
- [78] E. Fast, B. Chen, M. S. Bernstein, Empath: Understanding topic signals in large-scale text, in: CHI Conference on Human Factors in Computing Systems, 2016, pp. 4647–4657. doi:10.1145/2858036.2858535.
- [79] P. Mandra, E. Keuleers, M. Brysbaert, How useful are corpus-based methods for extrapolating psycholinguistic variables?, *The Quarterly Journal of Experimental Psychology* 68 (8) (2015) 1623–1642.
- [80] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 142–150.
- [81] D. Tang, F. Wei, B. Qin, T. Liu, M. Zhou, Cooolll: A deep learning system for Twitter sentiment classification, in: International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 208–212.
- [82] J. Deriu, A. Lucchi, V. De Luca, A. Severyn, S. Müller, M. Cieliebak, T. Hofmann, M. Jaggi, Leveraging large amounts of weakly supervised data for multi-language sentiment classification, in: Proceedings of the 26th International Conference on World Wide Web, WWW '17, International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 2017, pp. 1045–1052. doi:10.1145/3038912.3052611. URL <https://doi.org/10.1145/3038912.3052611>
- [83] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotion in speech, in: IEEE International Conference on Spoken Language Processing (ICSLP), Vol. 3, 1996, pp. 1970–1973.
- [84] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, *Speech Communication* 53 (9) (2011) 1062–1087.
- [85] S. Crouch, R. Khosla, Sentiment analysis of speech prosody for dialogue adaptation in a diet suggestion program, *ACM SIGHIT Record* 2 (1) (2012) 8–8.
- [86] V. Pérez-Rosas, R. Mihalcea, Sentiment analysis of online spoken reviews, in: INTERSPEECH, 2013, pp. 862–866.
- [87] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, S. Steidl, Emotion recognition using imperfect speech recognition, in: INTERSPEECH 2010, ISCA, 2010, pp. 478–481.
- [88] O. Rudovic, V. Pavlovic, M. Pantic, Context-sensitive dynamic ordinal regression for intensity estimation of facial action units, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (5) (2015) 944–958. doi:10.1109/TPAMI.2014.2356192.
- [89] R. Walecki, O. Rudovic, V. Pavlovic, M. Pantic, Copula ordinal regression for joint estimation of facial action unit intensity, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [90] S. Kaltwang, S. Todorovic, M. Pantic, Doubly sparse relevance vector machine for continuous facial behavior estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* in press. doi:10.1109/TPAMI.2015.2501824.
- [91] V. Belagiannis, A. Zisserman, Recurrent human pose estimation, CoRR abs/1605.02914. URL <http://arxiv.org/abs/1605.02914>
- [92] K. Bousmalis, M. Mehu, M. Pantic, Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools, *Image and Vision Computing* 31 (2) (2013) 203 – 221, affect Analysis In Continuous Input. doi:<http://dx.doi.org/10.1016/j.imavis.2012.07.003>. URL <http://www.sciencedirect.com/science/article/pii/S0262885612001059>
- [93] K. Bousmalis, S. Zafeiriou, L. P. Morency, M. Pantic, Infinite hidden conditional random fields for human behavior analysis, *IEEE Transactions on Neural Networks and Learning Systems* 24 (1) (2013) 170–177. doi:10.1109/TNNLS.2012.2224882.
- [94] K. Bousmalis, S. Zafeiriou, L. P. Morency, M. Pantic, Z. Ghahramani, Variational infinite hidden conditional random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (9) (2015) 1917–1929. doi:10.1109/TPAMI.2014.2388228.
- [95] M. A. Nicolaou, Y. Panagakis, S. Zafeiriou, M. Pantic, Robust canonical correlation analysis: Audio-visual fusion for learning continuous interest, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1522–1526. doi:10.1109/ICASSP.2014.6853852.
- [96] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, M. Pantic, Robust correlated and individual component analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38 (8) (2016) 1665–1678. doi:10.1109/TPAMI.2015.2497700.
- [97] W.-N. Wang, Y.-L. Yu, S.-M. Jiang, Image retrieval by emotional semantics: A study of emotional space and feature extraction, in: IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2006. doi:10.1109/ICSMC.

- 2006.384667.
- [98] V. Yanulevskaia, J. van Gemert, K. Roth, A.-K. Herbold, N. Sebe, J.-M. Geusebroek, Emotional valence categorization using holistic image features, in: IEEE International Conference on Image Processing (ICIP), 2008. doi:10.1109/ICIP.2008.4711701.
- [99] P. J. Lang, M. M. Bradley, B. N. Cuthbert, International Affective Picture System (IAPS): Technical manual and affective ratings, Tech. rep., National Institute of Mental Health (NIMH) Center for the Study of Emotion and Attention (CSEA) (1997).
- [100] S. Siersdorfer, E. Minack, F. Deng, J. Hare, Analyzing and predicting sentiment of images on the social web, in: ACM International Conference on Multimedia (MM), 2010. doi:10.1145/1873951.1874060.
- [101] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision (IJCV)* 60 (2) (2004) 91–110. doi:10.1023/B:VISI.0000029664.99615.94.
- [102] S. Baccianella, A. Esuli, F. Sebastiani, SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, in: International Conference on Language Resources and Evaluation (LREC), 2010, pp. 2200–2204.
- [103] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, J. Tang, Can we understand van Gogh’s mood?: Learning to infer affects from images in social networks, in: ACM International Conference on Multimedia (MM), 2012. doi:10.1145/2393347.2396330.
- [104] L.-J. Li, H. Su, E. P. Xing, L. Fei-Fei, Object Bank: A high-level image representation for scene classification & semantic feature sparsification, in: Annual Conference on Neural Information Processing Systems (NIPS), 2010. doi:10.1007/s11263-013-0660-x.
- [105] J. Yuan, S. McDonough, Q. You, J. Luo, SentiWordNet: Image sentiment analysis from a mid-level perspective, in: ACM International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM), 2013. doi:10.1145/2502069.2502079.
- [106] G. Patterson, J. Hays, SUN attribute database: Discovering, annotating, and recognizing scene attributes, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2012. doi:10.1109/CVPR.2012.6247998.
- [107] D. Borth, T. Chen, R. Ji, S.-F. Chang, SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content, in: ACM International Conference on Multimedia, 2013, pp. 459–460. doi:10.1145/2502081.2502268.
- [108] B. Jou, S. Bhattacharya, S.-F. Chang, Predicting viewer perceived emotions in animated GIFs, in: ACM International Conference on Multimedia (MM), 2014, pp. 213–216. doi:10.1145/2647868.2656408.
- [109] Z. Cai, D. Cao, R. Ji, Video (GIF) sentiment analysis using large-scale mid-level ontology, arXiv preprint arXiv:1506.00765 (2015) 1–5.
- [110] Y.-G. Jiang, B. Xu, X. Xue, Predicting emotions in user-generated videos, in: Conference of the Association for the Advancement of Artificial Intelligence (AAAI), 2014.
- [111] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, S.-F. Chang, Predicting viewer affective comments based on image content in social media, in: ACM International Conference on Multimedia Retrieval (ICMR), 2014. doi:10.1145/2578726.2578756.
- [112] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, S.-F. Chang, Object-based visual sentiment concept analysis and application, in: ACM International Conference on Multimedia (MM), 2014. doi:10.1145/2647868.2654935.
- [113] Y.-Y. Chen, T. Chen, T. Liu, H.-Y. M. Liao, S.-F. Chang, Assistive image comment robot - A novel mid-level concept-based representation, *IEEE Transactions on Affective Computing* 6 (3) (2015) 298–311. doi:10.1109/TAFFC.2014.2388370.
- [114] R. Plutchik, *Emotion: A Psychoevolutionary Synthesis*, Harper & Row, 1980.
- [115] D. Borth, Visual learning of socio-video semantics, Ph.D. thesis, Technische Universität Kaiserslautern (2014).
- [116] B. Jou, M. Yuying Qian, S.-F. Chang, SentiCart: Cartography and geo-contextualization for multilingual visual sentiment, in: ACM International Conference on Multimedia Retrieval (ICMR), 2016. doi:10.1145/2911996.2912022.
- [117] N. Pappas, M. Topkara, M. Redi, B. Jou, T. Chen, H. Liu, S.-F. Chang, Multilingual visual sentiment concept matching, in: ACM International Conference on Multimedia Retrieval (ICMR), 2016. doi:10.1145/2911996.2912016.
- [118] H. Liu, B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, S.-F. Chang, Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology, in: ACM International Conference on Multimedia Retrieval (ICMR), 2016. doi:10.1145/2911996.2912030.
- [119] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, 1998. doi:10.1109/5.726791.
- [120] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: Annual Conference on Neural Information Processing Systems (NIPS), 2012.
- [121] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)* 115 (3) (2015) 211–252. doi:10.1007/s11263-015-0816-y.
- [122] C. Xu, S. Cetintas, K.-C. Lee, L.-J. Li, Visual sentiment prediction with deep convolutional neural networks, arXiv preprint arXiv:1411.5731 (2014) 1–7.
- [123] A. S. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2014. doi:10.1109/CVPRW.2014.131.
- [124] Q. You, J. Luo, H. Jin, J. Yang, Robust image sentiment analysis using progressively trained and domain transferred deep networks, in: Conference of the Association for the Advancement of Artificial Intelligence (AAAI), 2014.
- [125] V. Campos, B. Jou, X. Giró-i-Nieto, From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction, arXiv preprint arXiv:1604.03489 (2016) 1–12.
- [126] V. Campos, A. Salvador, X. Giró-i-Nieto, B. Jou, Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction, in: ACM International Workshop on Affect & Sentiment in Multimedia, 2015, pp. 57–62. doi:10.1145/2813524.2813530.
- [127] T. Chen, D. Borth, T. Darrell, S.-F. Chang, DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks, arXiv preprint arXiv:1410.8586 (2014) 1–7.
- [128] T. Narihira, D. Borth, S. X. Yu, K. Ni, T. Darrell, Mapping images to sentiment adjective noun pairs with factorized neural nets, arXiv preprint arXiv:1511.06838 (2015) 1–8.
- [129] A. Mathews, L. Xie, X. He, SentiCap: Generating image descriptions with sentiments, in: Conference of the Association for the Advancement of Artificial Intelligence (AAAI), 2015.
- [130] F. Eyben, M. Wöllmer, B. Schuller, OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit, in: International Conference on Affective Computing and Intelligent Interaction 2009 and Workshops, 2009, pp. 1–6. doi:10.1109/ACII.2009.5349350.
- [131] F. Eyben, M. Wöllmer, B. Schuller, openSMILE: The Munich versatile and fast open-source audio feature extractor, in: ACM International Conference on Multimedia (MM), 2010, pp. 1459–1462. doi:10.1145/1873951.1874246.
- [132] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, *Neurocomputing* 174 (2016) 50–59. doi:10.1016/j.neucom.2015.01.095.
- [133] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems, in: Development of Multimodal Interfaces: Active Listening and Synchrony, Springer, 2010, pp. 148–156.

- [134] G. A. Miller, WordNet: A lexical database for English, *Communications of the ACM* 38 (11) (1995) 39–41. doi:10.1145/219717.219748.
- [135] R. Speer, C. Havasi, Representing general relational knowledge in conceptnet, 2012, pp. 3679–3686.
- [136] P. J. Stone, D. C. Dunphy, M. S. Smith, *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, 1966.
- [137] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework, *Image and Vision Computing* 31 (2) (2013) 153–163.
- [138] J. G. Ellis, B. Jou, S.-F. Chang, Why we watch the news: A dataset for exploring sentiment in broadcast video news, in: *ACM International Conference on Multimodal Interaction*, 2014, pp. 104–111. doi:10.1145/2663204.2663237.
- [139] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Vol. 1631, 2013, p. 1642.
- [140] D. McDuff, R. E. Kaliouby, E. Kodra, R. Picard, Measuring voter’s candidate preference based on affective responses to election debates, in: *Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 369–374. doi:10.1109/ACII.2013.67.
- [141] N. A. Madzlan, J. G. Han, F. Bonin, N. Campbell, Automatic recognition of attitudes in video blogs-prosodic and visual feature analysis., in: *INTERSPEECH*, 2014, pp. 1826–1830.
- [142] B. Siddiquie, D. Chisholm, A. Divakaran, Exploiting multimodal affect and semantics to identify politically persuasive web videos, in: *ACM International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 203–210. doi:10.1145/2818346.2820732.
- [143] C. Langlet, C. Clavel, Grounding the detection of the user’s likes and dislikes on the topic structure of human-agent interactions, *Knowledge-Based Systems*[In press]. doi:10.1016/j.knsys.2016.05.038.
- [144] A. Khosla, A. Das Sarma, R. Hamid, What makes an image popular?, in: *International Conference on World Wide Web (WWW)*, 2014, pp. 867–876. doi:10.1145/2566486.2567996.
- [145] C. Schulze, D. Henter, D. Borth, A. Dengel, Automatic detection of CSA media by multi-modal feature fusion for law enforcement support, in: *ACM International Conference on Multimedia Retrieval (ICMR)*, 2014, pp. 353:353–353:360. doi:10.1145/2578726.2578772.
- [146] D. McDuff, R. E. Kaliouby, J. F. Cohn, R. W. Picard, Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads, *IEEE Transactions on Affective Computing* 6 (3) (2015) 223–235. doi:10.1109/TAFFC.2014.2384198.
- [147] M. Larson, M. Soleymani, M. Eskevich, P. Serdyukov, R. Ordelman, G. Jones, The community and the crowd: Multimedia benchmark dataset development, *IEEE Multimedia* 19 (3) (2012) 15–23. doi:10.1109/MMUL.2012.27.
- [148] M. Williams, *Data Science Popul Austin: Privilege and Supervised Machine Learning* (2014). URL <https://vimeo.com/163292139>
- [149] D. Morrison, Toward automatic censorship detection in microblogs, in: W.-C. Peng, H. Wang, J. Bailey, S. V. Tseng, B. T. Ho, Z.-H. Zhou, L. A. Chen (Eds.), *Trends and Applications in Knowledge Discovery and Data Mining*, Springer International Publishing, 2014, pp. 572–583. doi:10.1007/978-3-319-13186-3_51.
- [150] S. K. D’Mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, *ACM Computing Surveys (CSUR)* 47 (3) (2015) 43:1–43:36. doi:10.1145/2682899.

Graphical Abstract



RIPT

AC

Highlights

- Sentiment and sentiment analysis are defined.
- Current work on multimodal sentiment analysis is reviewed and summarized.
- Challenges and opportunities in multimodal sentiment analysis are discussed.

ACCEPTED MANUSCRIPT