Neurocomputing 174 (2016) 50–59

Contents lists available at ScienceDirect

Neurocomputing

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

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

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

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

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2015.01.095
0925-2312/

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

2. Motivations

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

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

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

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

3. Related work

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

3.1. Video: emotion and sentiment analysis from facial expressions

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

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

3.2. Audio: emotion and sentiment recognition from speech

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

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

Table 1
Sample of SenticNet data.

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