Cognitive Computation (2021) 13:229–230
https://doi.org/10.1007/s12559-020-09738-0

Affect Recognition for Multimodal Natural Language Processing

Soujanya Poria[1] · Ong Yew Soon[2] · Bing Liu[3] · Lidong Bing[4]

Published online: 29 June 2020

Language is inherently multimodal. It has many forms of appearance, like speech, gestures, facial expressions, and head-nods. In an ideal human-machine conversational system, machines should understand this multimodal language. Understanding human language also largely depends on the machines' ability to interpret emotions. Emotional sensitivity can prevent desultory answers provided by these machines, thus making conversations more natural and engaging. For us humans, emotions aid our learning, communication, and decision-making. Hence, over the past two decades, there has been a significant effort to incorporate cognitive capabilities into machines so that they can interpret, comprehend and express emotions. Computational analysis of human multimodal language is an emerging research area in natural language processing (NLP). It expands the horizons of NLP to study language used in face to face communication and in online multimedia. This form of language contains modalities of language (in terms of spoken text), visual (in terms of gestures and facial expressions) and acoustic (in terms of changes in the voice tone). At its core, this research area is focused on modeling the three modalities and their complex interactions. This special issue on Affect Recognition in Human Multimodal Language aims to facilitate the growth of this new research direction in the community. The challenges of modeling human multimodal language can be split into two major categories: (1) studying each modality individually and modeling each in a manner that can be linked to other modalities (also known as intramodal dynamics, (2) linking the modalities by modeling the interactions between them (also known as intermodal dynamics). Common forms of these interactions include complementary or correlated information across modes. Intrinsic to each modality, modeling human multimodal language is complex due to factors such as idiosyncrasy in communicative styles, non-trivial alignment between modalities and unreliable or contradictory information across modalities. Therefore, computational analysis of multimodal language becomes a challenging research area. This special issue aimed at bringing together contributions from both academics and practitioners in the context of affect recognition in multimodal language with a primary focus on:

1. Multimodal affect recognition in monologues,
2. Effective multimodal fusion.
3. Detecting affect in dyadic and multiparty multimodal conversations. Detecting affect in conversations is more challenging than monologues. This is primarily due to the presence of complex inter-dependency between speaker states in the conversation,

This special issue consists of 4 papers. Each of these has undergone several rounds of revisions and has been reviewed by three or more reviewers.

In "Exploring Perception Uncertainty for Emotion Recognition in Dyadic Conversation and Music Listening," Jing Han, Zixing Zhang, Zhao Ren1, and Bjorn Schuller attempt to exploit the perception uncertainty measure to adjust and enhance emotion prediction performance. The ground perception uncertainty is quantified with an inter-annotator agreement. In particular, the authors have adopted a multi-task learning framework with two tasks: initial emotion recognition and perception uncertainty prediction. The final emotion is adjusted with the predicted uncertainty and trained with a multi-task objective. The evaluations on

✉ Soujanya Poria
sporia@sutd.edu.sg

Ong Yew Soon
ASYSOng@ntu.edu.sg

Bing Liu
liub@uic.edu

Lidong Bing
l.bing@alibaba-inc.com

[1] Singapore University of Technology and Design, Singapore, Singapore

[2] Nanyang Technological University, Singapore, Singapore

[3] University of Illinois at Chicago, Chicago, IL, USA

[4] DAMO Academy, Alibaba Group, Singapore, Singapore

two datasets—RECOLA and emoMusic—show the efficacy of this approach.

The work “Analyzing Connections Between User Attributes, Images, and Text” by Laura Burdick, Rada Mihalcea, Ryan L. Boyd, and James W. Pennebaker blends psychology with artificial intelligence. In this work, the authors study co-relations among humans by relying on a dataset of images and captions provided by the 1350 individuals. Authors also explore the problem of gender and personality prediction by relying on unimodal and multimodal features. The evaluation studies demonstrate the superiority of the multimodal features over unimodal.

In the paper titled “Emoji helps! A Multi-modal Siamese Architecture for Tweet-user Verification” by Chanchal Suman, Sriparna Saha, Pushpak Bhattacharyya, Rohit Shyamkant Chaudhari, the problem of authorship verification from Tweets is addressed. Authorship verification is an important problem of natural language processing which deals with the process of examining the characteristics of a questioned text in order to draw conclusions on its authorship. It has a long list of applications including analysis of long fraud documents, terrorist conspiracy texts, short letters, blog posts, emails, SMS, Twitter streams, or Facebook status updates to check the authenticity and identify fraudulence. In the current work, given a set of known tweets of a particular author, the task is to predict whether an unknown tweet is written by the same author or not. As an additional modality, emojis present in the tweet is also utilized apart from tweet-text. Usage of emojis in Twitter platforms is common and it is a representation of emotions of the authors. These emojis represent some writing styles of the authors as there are many emojis that represent a single emotion and everybody has his/her own choice. In the current paper, authors have manually created and annotated a data set for authorship verification involving tweets and the emojis. Then, a Siamese-based multimodal tweet verification framework is proposed for the verification of tweet-user. Here inputs are two tweets and the corresponding emojis and the task is to predict whether both of them are written by the same author or not. Experimental results on the newly developed data set in comparison with the single-modal (text-based) approach establish the efficacy of using emoji as an additional modality for tweet-user verification.

Finally, in “Emotion aided Dialogue Act Classification for Task-Independent Conversations in a Multi-Modal Framework” by Tulika Saha, Dhawal Gupta, Sriparna Saha, Pushpak Bhattacharyya, authors have worked on the problem of dialogue act classification. Dialogue Act Classification (DAC) is an important step towards building dialogue systems. It helps in understanding the communicative intention of the user and is represented as a function of the speaker’s utterance. This is often posed as a sequence labeling problem. Existing systems mostly consider the text data (narrations of dialogues) available for determining the DAs. A combination of multiple modalities (like facial expression, acoustic) often helps in gathering crucial cues to better identify the communicative intention and emotional state of the speaker. Moreover, as emotion detection (ED) and DAC are related tasks, simultaneously solving both these tasks in a multi-task framework can help in analyzing the affect of emotion in the automatic identification of the DAs. The authors have worked along these directions. The contributions of the paper are twofold: (1) A DL-based multi-tasking network has been developed in a multimodal framework (acoustic and text) for solving two tasks (DAC and emotion recognition (ER)) jointly; (2) An open-access, standard ER-based multimodal corpus IEMOCAP is extended by annotating the corresponding DA labels for solving both the tasks (DAC and ER). Detailed experimental results illustrated the utilities of a multimodal multi-tasking framework compared with single-task frameworks and single-modal frameworks.

Acknowledgments The guest editors are grateful to the Editor-in-Chief, Amir Hussain, and to the many reviewers who kindly agreed to serve for this special issue and submitted their insightful reviews in a timely manner.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.