MixedEmotions: An Open-Source Toolbox for Multimodal Emotion Analysis

Paul Buitelaar, Ian D. Wood, Sapna Negi, Mihael Arcan, John P. McCrae, Andrejs Abele, Cécile Robin, Vladimir Andryushechkin, Housam Ziad, Hesam Sagha, Maximilian Schmitt, Björn W. Schuller, J. Fernando Sánchez-Rada, Carlos A. Iglesias, Carlos Navarro, Andreas Giefer, Nicolaus Heise, Vincenzo Masucci, Francesco A. Danza, Ciro Caterino, Pavel Smrž, Michal Hradiš, Filip Povolný, Marek Klimeš, Pavel Matějka, and Giovanni Tummarello

***Abstract*—Recently, there is an increasing tendency to embed functionalities for recognizing emotions from user-generated media content in automated systems such as call-centre operations, recommendations, and assistive technologies, providing richer and more informative user and content profiles. However, to date, adding these functionalities was a tedious, costly, and time-consuming effort, requiring identification and integration of diverse tools with diverse interfaces as required by the use case at hand. The MixedEmotions Toolbox leverages the need for such functionalities by providing tools for text, audio, video, and linked data processing within an easily integrable plug-and-play platform. These functionalities include: 1) for text processing: emotion and sentiment recognition; 2) for audio processing: emotion, age, and gender recognition; 3) for video processing: face detection and tracking, emotion recognition, facial landmark localization, head pose estimation, face alignment, and body pose estimation; and 4) for linked data: knowledge graph integration. Moreover, the MixedEmotions Toolbox is open-source and free. In this paper, we present this toolbox in the context of the existing landscape, and provide a range of detailed benchmarks on standard testbeds showing its state-of-the-art performance. Furthermore, three real-world use cases show its effectiveness, namely, emotion-driven smart TV, call center monitoring, and brand reputation analysis.**

***Index Terms*—Emotion analysis, open source toolbox, affective computing, linked data, audio processing, text processing, video processing.**

This work was supported by the European Unions Horizon 2020 Programme research and innovation programme under Grant 644632 (MixedEmotions). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yi-Hsuan Yang. *(Corresponding author: Hesam Sagha.)*

P. Buitelaar, I. D. Wood, S. Negi, M. Arcan, J. P. McCrae, A. Abele, C. Robin, V. Andryushechkin, and H. Ziad are with the National University of Ireland Galway, Galway, Ireland (e-mail: paul.buitelaar@insight-centre.org; ian.wood@insight-centre.org; sapna.negi@insight-centre.org; mihael.arcan@insight-centre.org; John.McCrae@insight-centre.org; andrejs.abele@insight-centre.org; cecile.robin@insight-centre.org; vladimir.andryushechkin@insight-centre.org; housam.ziad@insight-centre.org).

H. Sagha was with the Chair of Complex & Intelligent Systems, University of Passau, Passau 94032, Germany. He is now with audEERING GmbH, Gilching 82205, Germany (e-mail: hesamsga81@gmail.com).

M. Schmitt is with the Chair of Complex & Intelligent Systems, University of Passau, Passau 94032, Germany (e-mail: maximilian.schmitt@uni-passau.de).

B. W. Schuller is with the Chair of Complex & Intelligent Systems, University of Passau, Passau 94032, Germany, and also with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: bjoern.schuller@imperial.ac.uk).

J. F. Sánchez-Rada and C. A. Iglesias are with the GSI Universidad Politécnica de Madrid, Madrid 28040, Spain (e-mail: jf.sanchez@upm.es; cif@dit.upm.es).

C. Navarro is with the Paradigma Digital, Madrid 28224, Spain (e-mail: cnavarro@paradigmadigital.com).

A. Giefer and N. Heise are with the Deutsche Welle, Bonn 53113, Germany (e-mail: andreas.giefer@dw.com; nicolaus.heise@dw.com).

V. Masucci, F. A. Danza, and C. Caterino are with the Expert Systems, Modena 41123, Italy (e-mail: vmasucci@expertsystem.com; fadanza@gmail.com; ccaterino@expertsystem.com).

P. Smrž and M. Hradiš are with the Brno University of Technology, Brno-střed 60190, Czech Republic (e-mail: smrz@fit.vutbr.cz; ihradis@fit.vutbr.cz).

F. Povolný, M. Klimeš, and P. Matějka are with the Phonexia, Brno-Krlovo Pole 612 00, Czech Republic (e-mail: filip.povolny@phonexia.com; klimes@phonexia.com; matejka@phonexia.com).

G. Tummarello is with the Siren Solutions, Dublin, Ireland (e-mail: giovanni@siren.solutions).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes "ex-entities.png," "ex-linked-data.pdf," "ex-social-tv.png," "ex-call-center.png," and "ex-brand-reputation.png." Contact paul.buitelaar@insight-centre.org for further questions about this work.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

I. Motivation & Introduction

ANY Media content (e.g., social media, TV/Radio program) contains a vast amount of information which can be harvested for various analysis from a content perspective (e.g., reputation analysis [1], content emotion analysis [2]) and a content-authors perspective (e.g., user profiling and recommendation [3], [4], user community analysis [5], [6]). Nevertheless, as part of this information, the emotional aspects of the media content has not received its well-deserved attention and its utility of those aspects have not yet been well-exploited in real-world or commercial scenarios. Emotions are important part of human life as they enhance communication and understanding between people. Similarly, incorporating emotion-related information into multimedia content and multimedia analysis could enhance usability and user-adaptability. Although some research advances have been made in this direction (such as: emotion analysis of users' audio or video for enriching users' profiles for media recommendation [3], [4], [7], affect prediction from movies [8], or speech [9]), they have not gone further than research, and reproduction of such algorithms is time consuming and fault-prone.

The MixedEmotions Toolbox[1] introduced herein fills this gap by providing a plug-and-play and ready-to-use set of emotion recognition modules that can be used in isolation or in combination through predefined or configurable workflows. It provides a unified solution for large-scale emotion analysis on heterogeneous, multilingual, text, speech, video, and social media data streams, leveraging open access and proprietary data sources including modules for collection of social media data, and exploiting social context by leveraging social network graphs. It also includes entity linking and knowledge graph technologies for semantic-level emotion information aggregation and integration. Available free tools have been adapted and included in the platform alongside tools developed by the authors of this paper.

This paper describes the current version of the MixedEmotions Toolbox, including its underlying architecture, the modules it comprises and their capabilities, and applications of the platform in three representative multimedia-related use cases: Social TV, Brand Reputation Management, and Call Center Operations.

Before describing the toolbox in detail, we describe what *emotion* actually is and how it is represented and provide a quick review of existing emotion analysis platforms as well as an overview of requirements for emotion analysis on big-data.

A. What is Emotion?

One of the most complete and accepted definitions of emotion is proposed by Scherer [10] through a component process model, in which an emotion is a synchronization of different cognitive and physiological components in response to a stimulus event. The expression of emotions through facial and vocal changes is originated from the 'somatic nervous system' component. Moreover, emotions and preferences (as stable emotions with low behavioural impacts) can be conveyed through verbal or written *content*, such as product reviews, opinions, and suggestions. Therefore, analysing the facial and vocal changes as well as verbal and written content provides clues for automatic emotion recognition.

B. Quick Overview of Emotion Representations Used

Various representation schemes for emotions have been proposed, each based on particular criteria. Ekman's *six basic emotions* (Anger, Fear, Surprise, Happiness, Disgust, Sadness) are based on the universality of those emotions [11]; Plutchik's *wheel of emotion* is further based on contrast and closeness of emotions [12]; Russel's *Circumplex model* is constructed to capture the core affect in a two dimensional (Arousal and Valence) model [13], [14]; Osgood identified three primary dimensions of emotion expression (Pleasure, Arousal, Dominance) [15]; and more recently, Fontaine et al. identified a fourth dimension (unpredictability) [16]. Arousal reflects the level of energy in the emotion (e.g., pleased vs. ecstatic); valence reflects the hedonic tone (e.g., pleasant vs. unpleasant); dominance represents the sense of control or dominant nature of the emotion (e.g., fear vs. anger); and unpredictability refers to the appraisal of expectedness or familiarity.

In the MixedEmotions Toolbox, the preferred emotion representation model is the four dimensional model, combined with emotion intensity as a fifth dimension and a level of confidence in the measurement. However, due to limitations in available gold standard data and error-prone human ability to map perceived emotions into these dimensions, some modules in the MixedEmotions Toolbox represent emotions as a subset of these dimensions. For emotion representation in audio and video processing, we chose a two (arousal and valence) or three (+ dominance) dimensional emotion model. The choice of the dimensional model is due (among others) to: (i) it can be mapped not only to the six basic emotions but to a myriad of emotion categories, (ii) emotions which resemble each other are located in the vicinity of each other, (iii) it is easier to define continuous values as the output of machine learning systems (such as neural networks), and (iv) it is easier to handle the decision fusion of different subsystems in the continuous domain. In the analysis of text, there were previously no substantial resources annotated with a dimensional emotion model; however, resources and tools that utilize Ekmans' six basic emotions were available. In addition, there are many resources available for "sentiment analysis", which is essentially just the Valence dimension. For this reason, several toolbox modules for text analysis utilize these representation schemes, and functionality is provided for translating to and from a dimensional representation. New data annotated with a four dimensional model is provided alongside models for detecting emotion with this scheme utilising the new data.

C. Existing Emotion Analysis Platforms

Some web services for emotion analysis from textual contents, facial expressions, and speech already exist. Table I summarizes some known services along with their characteristics. As can be seen in the table, all the services are for the analysis of only one modality such as facial, textual, or speech. Moreover, most of the services are not free and not open-source. The MixedEmotions Toolbox overcomes these limitations by providing multi-modal, open-source, free, and user-friendly emotion analyzers.

D. Emotion Analysis in Big Data and Pre-requisites

To deploy a multifaceted emotion analyzer for big-data, the seven "Vs" of big data (Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value) should be addressed. Among them, Variety encompasses multimodality (audio, video, text) and multilinguality/multiculturalism, and Veracity emerges from subjectivity of assessments (annotations). These aspects have been addressed for: (i) the textual modality by: automatic translation [17], defining multilingual WordNet Grid [18], and (ii) for the audio modality by: analyzing within or between language family emotion recognition [19], feature transfer

[1]MixedEmotions Toolbox is the outcome of the European Project MixedEmotions (https://mixedemotions-project.eu/). Note that it is not about the 'co-occurrence of different emotions' (as the psychological term 'mixed emotions'), but about the 'emotions from mixed modalities'.

TABLE I
A SHORT LIST OF AVAILABLE EMOTION ANALYZER SERVICES FOR T(EXTUAL), F(ACIAL), AND S(PEECH) CONTENTS

Service	Modality	Open Source	Free
IBM Watson AlchemyLanguage (www.ibm.com/watson),Bitext (www.bitext.com)	T	No	No
MoodPatrol (market.mashape.com/soulhackerslabs/)	T	No	No
Synesketch (krcadinac.com/synesketch)	T	Yes	Yes
Microsoft Cognitive Services (www.microsoft.com/cognitive-services)	F	No	No
IMOTIONS (wwww.imotions.com)	F	No	No
Affectiva Emotion API (www.affectiva.com)	F	No	Free/Enterprise Editions
EmoVu (www.emovu.com), CrowdEmotions (www.crowdemotion.co.uk)	F	No	?
Nviso (www.nviso.ch/technology.html), SkyBiometry (www.skybiometry.com)	F	No	Limited/Non-Free Editions
audEERING SensAI (www.audeering.com/technology/sensai/)	S	Yes	Free Research Edition (openSMILE)
Good Vibrations (www.good-vibrations.nl)	S	No	No
Vokaturi (www.vokaturi.com/)	S	No	Limited/Enterprise Editions

learning between languages [20], model transfer learning [21], language identification [22], audio denoising [23], and decision aggregation through cooperative speaker models [24]. Regarding the Volume and Velocity, there is a need for fast computation. This has been investigated using End-to-End approaches for speech emotion analysis [25], fast GPU processing of audio and video processing [26], and crowdsourcing and a semi-supervised active learning approach for automatically labeling large amounts of data [27], [28]. Some of these aspects have been deployed within the MixedEmotions Toolbox. Further, the MixedEmotions Toolbox can be easily deployed on one or more machines for distributed analysis and fast processing of large amount of data. A Visualization module is also included in the toolbox (Section III-D). Moreover, to investigate the Value of this MixedEmotions Toolbox, we designed three case studies on multimedia emotion processing which will be discussed in Section IV.

II. ARCHITECTURE OVERVIEW

The MixedEmotions Toolbox follows a microservice architecture in which the modules in the toolbox are independent of each other, so users need only the modules required for his/her analysis and can skip the others. The modules are containerized using Docker,[2] and therefore can be deployed without dependency restrictions, with the only requirement being a Docker server. Docker servers exist for all major operating systems, can be installed on small computers as well as in extensible cloud environments. As well as individual modules, users can also benefit from an orchestrator in the toolbox to enable big data operations sustained on horizontal scalability (using more machines). This orchestrator provides users an easy starting point to build applications as needed. In a nutshell, the orchestrator is an ETL[3] pipeline [29] adapted to the structure of the MixedEmotions, thus, it is suited to work with Docker containers deployed in Mesos[4] (Fig. 1), as well as external services as long as they have a REST API.[5] It is fully configurable with plain text configuration files, so a user does not need to have programming skills.

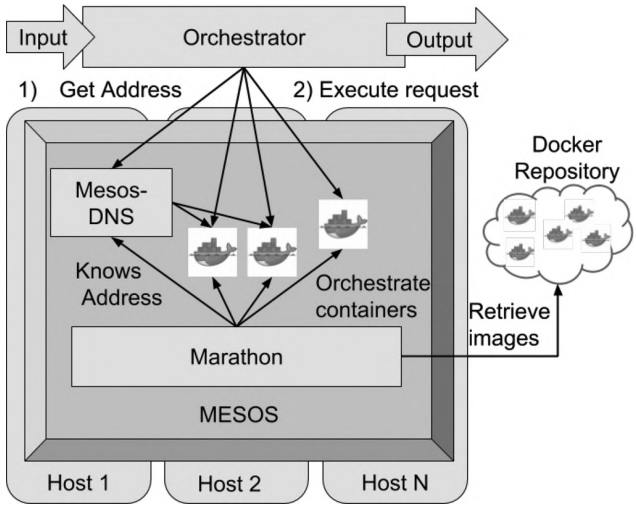

Fig. 1. Orchestrator within the MixedEmotions Toolbox.

Note that Docker Servers and Mesos Services can be deployed on multiple platforms, including, Linux, OS X, Windows and Windows Server, and making the MixedEmotions Toolbox platform independent.

Where to find the MixedEmotions Toolbox: The MixedEmotions platform is available online for demonstration and testing.[6] Open source and free for research purposes modules are located on GitHub[7] (source code and documentation), and ready-to-use modules can be found in the MixedEmotions docker repository.[8]

[2]www.docker.com
[3]Extract, Transform, Load.

[4]mesos.apache.org: The Mesos kernel runs on every machine and provides applications with APIs for resource management and scheduling across entire datacenter and cloud environments.
[5]REST = Representational State Transfer, API=Application Programming Interface. This is a simple and widely used standard for providing services over the internet.
[6]http://mixedemotions.insight-centre.org/
[7]https://github.com/MixedEmotions
[8]https://hub.docker.com/r/mixedemotions/

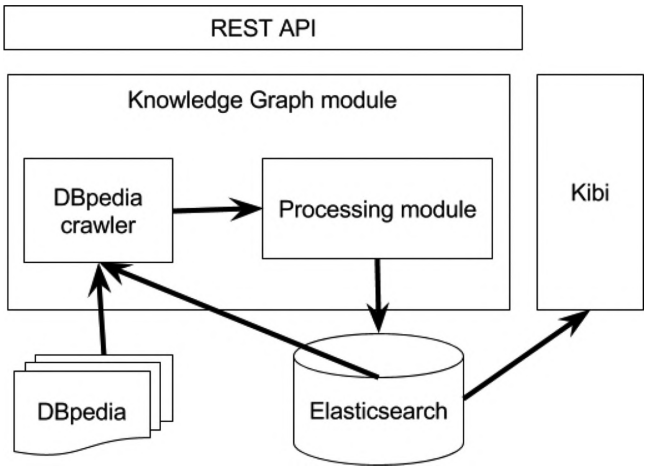

Fig. 2. Knowledge Graph module architecture.

III. Overview of Main Functionalities

In this section, we describe the modules in the MixedEmotions Toolbox for text, audio, and video processing with the focus of emotion recognition.

A. Text Processing

The toolbox includes the following modules for text processing: (1) several modules that implement recognition of affect expressed in text, (2) a module for the recognition of suggestions expressed in text, and (3) modules for semantic processing of text. While sentiment analysis (the recognition of positive/negative sentiment often directed at a particular entity) is an established field with many standard data sets and well developed methodologies (e.g., [30]), the recognition of more nuanced affect has received less attention, and in particular, there are very few gold standard annotated resources. This is also true for analysis of sentiment and emotion from many languages. To address this lack, two new resources for emotion detection from text were developed: (4a) a collection of tweets annotated with four emotion dimensions, and (4b) translations of WordNet into all official European languages, enabling the application of WordNet-based affective lexical resources (e.g., WordNet-Affect [31] and Senti-WordNet [32]) in those languages. Details of these modules and resources are as follows.

1) Sentiment and Emotion Recognition: Models for sentiment and emotion recognition from text across several languages and for general text and social media domains are included in the platform (see Tables II and III).

Several free and/or open source sentiment analysis tools are included in the toolbox. In addition, two Long-Short Term Memory (LSTM) [33] deep learning models trained on movie reviews [34] and tweets [35] are provided (see Table II). Evaluation of the English language sentiment models was performed on test tweets from the SemEval2015 task 10B [36] for tweet models, and movie reviews from [37] for general text models. F1 scores from the cross-validation analysis of the training data are also provided where appropriate.

The toolbox includes models for emotion detection from text for two emotion representation schemes: Ekman's six emotion categories [11] and the 4-dimensional Valence/Arousal/Dominance/Surprise representation scheme [16] (see Table III). These models fall into two broad categories: unsupervised lexicon based models, provided primarily as baseline systems, and supervised models trained on publicly available annotated data sets. The lexicon based models count word occurrences, summing associated emotions. Models built with WordNet-Affect [31] for Ekman emotions and Affective Norms for English Words (ANEW) [43], [44] for VAD are also provided. Two supervised Ekman models are included: one trained on tweet data utilising emotion hash tags as noisy emotion labels [45] and another from the recent WASSA shared task on emotion recognition [46], [47]. A final model trained on new VADS annotated data (see Section III-A4a) is also provided. F_1 and R^2 scores from the cross-validation analysis of the training data are provided where appropriate.

2) Suggestion Mining: Alongside requirements for the detection of sentiment and emotions in an opinionated text, another useful service which has been developed in the MixedEmotions Toolbox is the identification of suggestions and advice that may have been made in those texts. This will allow users and service providers to make more valuable decisions based on richer inferences on data (e.g., a brand reputation can be affected by positive and negative suggestions from the users alongside with their expressed sentiments).

Suggestion mining refers to the task of detection of such suggestions (advice, tips, recommendations, etc.) in the text obtained from social media. An example of suggestion in tweets can be: *"Dear Microsoft, release a new zune with your wp7 launch on the 11th. It would be smart"*. Since suggestion mining is a very recent area of research, our contribution also covers the creation of benchmark datasets to facilitate the development and evaluation of suggestion mining methods [49]. Currently, this module is only available for the English language.

The module utilises a Long Short Term Memory (LSTM) Neural Network to classify texts as suggestion or not suggestion. It is trained on suggestion mining datasets developed in-house, using crowdsourced annotations of hotel and electronics reviews [49]. This classifier yields a F_1 score of 0.64 and 0.67 over 10-fold cross-validation for hotel and electronics datasets respectively.

3) Semantic Analysis of Text: The toolbox includes modules for entity recognition in Spanish and English, both built on DBpedia[9] [50]. The English module issues queries to a Lucene[10] database containing all matching WikiPedia URIs, and entities are selected according to the score from the Lucene index. The DBpedia URI, the entity and its type are returned by the module.

The Spanish Entity recognition module is created using entities from DBpedia and their inlink count, which is the number of other entities related to it. Then, an *entities dictionary* is created using all the entities above a certain threshold. Given a text, the module will then extract all the phrases that can be found in the entities dictionary.

[9]DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web.

[10]Apache Lucene is an open-source search software: https://lucene.apache.org/

TABLE II
MODELS FOR SENTIMENT DETECTION FROM TEXT

Affect Representation	Lang	Domain	Algorithm	Train CV F_1	Test F_1	Reference
Sentiment (+, n, −)	EN	Text	LSTM	—	.76	[33] trained on [34]
Sentiment (++, +, n, −, –)	EN	Text	CoreNLP (NN)	—	.62	[38]
Sentiment (+, n, −)	EN	Text	LingPipe (SVM/NB)	—	.76	[39]
Sentiment (+, n, − / continuous)	EN	Text	VADER (Lexical + Rules)	—	.76	[40]
Sentiment (+, n, −)	EN	Tweets	LSTM	.48	.67	[33] trained on [35]
Sentiment (+, n, −)	EN, ES	Tweets	Sentiment140	.76	.79	[41]
Sentiment (+, n, −)	ES	Tweets	SVM (TASS2015)	.74	—	[42]
Sentiment (+, n, −)	CZ	Text	LingPipe (CZ reviews)	.86	—	[39]

Evaluation data: SemEval2015 task 10B [36] (tweet sentiment), movie reviews from [37] (text sentiment).

TABLE III
MODELS FOR EMOTION DETECTION FROM TEXT

Affect Representation	Lang	Domain	Algorithm	Train Eval.	Reference
Emotion (Ekman)	Multiple	Text	WordNet-Affect	—	[31]
Emotion (Ekman)	EN	Tweets	SVM (hashtags)	F_1: .37	[45]
Emotion (4 Ekman Intensities)	EN	Tweets	BLSTM + SVM	R^2: .45	[46] trained on [47]
Emotion (VAD)	EN, ES	Text	ANEW	—	[43], [44]
Emotion (VADS)	EN	Tweets	BLSTM	R^2: .24	[48] trained on new data (See 4a below)

4) New Resources for Affective Analysis of Text:

a) Emotion annotated text data (standard and new): There exists a limited number of publicly available emotion annotated text resources; these include: two thousand news headlines annotated with Ekman's six emotions [51], and several dimensionally annotated corpora: Affective Norms for English Texts [52] (a collection of 120 generic texts with VAD annotations), a collection of 2895 Facebook posts annotated by two annotators with Valence and Arousal dimensions [53], and the recent EMOBANK [54] (a collection of ten thousand texts from diverse sources but not including tweets). Moreover, Yu et al. [55] presented a collection of 2009 Chinese sentences from various online texts annotated with Valence and Arousal.

As a step towards addressing this limitation, we collected two new annotated tweet corpora: one containing 2019 generic tweets annotated with *Valence*, *Arousal*, *Dominance*, and *Surprise* (with annotator agreement of Krippendorffs' Alpha .42) [56], and another containing 360 tweets containing expressive emoji annotated with Ekman's six emotions [57] (with annotator agreement of Krippendorffs' Alpha .33).

b) Polylingual WordNet: The Princeton WordNet [58] is one of the most important resources for natural language processing, but is only available for English. Although it has been translated using the *expand* approach to many other languages [59]–[61], most of the WordNet resources resulting from these efforts have fewer synsets than the Princeton WordNet. Since manual translation and evaluation of WordNets is a very time consuming and expensive process, we apply Statistical Machine Translation (SMT)[11] to automatically translate WordNet entries. The biggest challenge in translating WordNets with an SMT system lies in the need to translate all senses of a word including low frequency senses. While an SMT system can only return the most frequent translation when given a term by itself, it has been observed that it provides strong word sense disambiguation when the word is given in a disambiguated context [17]. Therefore, we leverage existing translations of WordNet in other languages to identify contextual information for WordNet senses from a large set of generic parallel corpora. We used an approach to select the most relevant sentences from a parallel corpus based on the overlap with existing translations of WordNet in as many pivot languages as possible. The goal is to identify sentences that share the same semantic information with respect to the synset of the WordNet entry that we want to translate. This approach allows us to provide a large multilingual WordNet in 23 different European languages, which we call Polylingual WordNet.[12] As a result, the WordNet-Affect based emotion detection module is also applicable to those languages.

[11]The SMT models also exist as a MixedEmotions' module.

B. Audio Processing

This module recognizes emotions in terms of arousal and valence from speech signals.[13] It is based on the Bag-of-Audio-Words (BoAW) approach [62], trained on continuous emotionally labeled data (the *RECOLA* database [63]). RECOLA is an audio-visual database of 46 subjects during dyadic conversation in French. For each subject, a recording of 5 minutes length has been annotated time-continuously for Arousal and Valence dimensions by six different annotators (3 female, 3 male). From the 6 annotations, a single *gold standard* sequence has been computed for each dimension, using an *evaluator weighted estimator* [64].

[12]http://polylingwn.linguistic-lod.org/

[13]Although *audio* includes speech, music, and other acoustics, the module that we built within the MixedEmotions Toolbox is for speech. Other modules, such as music emotion may be added later to the toolbox.

BoAW originates from the bag-of-words approach in natural language processing. In this approach, word histogram vectors are used as a feature to classify text documents, e.g., in terms of *sentiment* or the author's *gender* [65]. For BoAW, the first step is the extraction of acoustic low-level descriptors (LLDs) from the raw waveform of the speech signal. audEERING's open-source toolkit openSMILE[14] [66] is used to extract *Mel-frequency cepstral coefficients (MFCCs)* and logarithmic energy over a short audio frame of 25 ms, with a step size of 10 ms. Each 13-dimensional LLD vector is then assigned to a so-called *audio word*, i.e., a template of an LLD vector. This is accomplished through a vector quantization step using a codebook which has been learned beforehand. A random sampling [67] of 200 LLDs from the training data has proven to be suitable for the task. In the vector quantization step, Euclidean distance is taken into account.

To make the power of the histogram independent from the duration of the input segment, a histogram normalization is performed. The whole BoAW-processing is accomplished by the open-source toolkit openXBOW[15] [68].

For decoding, a *support vector regressor* (SVR) with a linear kernel was trained [69]. All hyperparameters have been optimized systematically using a speaker-independent split of the database into training, validation, and test partitions [62]. The performance of arousal and valence recognition in terms of *Concordance Correlation Coefficients (CCC)*[16] [70] for the RECOLA and SEWA [71] datasets are summarized in Table IV.

TABLE IV
PERFORMANCE (CCC) OF THE AUDIO EMOTION RECOGNITION MODULE ON THE RECOLA DATABASE

Database	Partition	Arousal	Valence
RECOLA	Development	.797	.529
	Test	.722	.452
SEWA	Development	.359	.157

TABLE V
FACE VERIFICATION ACCURACY OF CNN [73] FINETUNED WITH AFFINE AND SIMILARITY GEOMETRIC ALIGNMENTS ON YOUTUBE FACES DATASET

Original	Affine alignment	Similarity alignment
.973	.974	.977

C. Video Processing

This module is responsible for emotion recognition (arousal/valence and Ekman's emotions) from facial gestures. The emotion recognition runs on top of face detection and tracking, facial landmark localization, head pose estimation, and face alignment. Face detection is based on a discriminatively trained deformable part model [72] which runs at approximately 8–16 fps on 720 p video. Faces are tracked in a video according to standard tracking by detection. To maintain identities across these partial tacks, visual fingerprints are extracted from individual frames, and clustered by hierarchical clustering using complete linkage and cosine distance. The features are then used as activations of a convolutional neural network (CNN) [73] which is fine-tuned on the Megaface dataset [74] for similarity transform facial alignment (See Table V for the effect of the fine-tuning).

Facial landmarks are localized by an ensemble of regression trees [75] which provides decent facial point localization at real-time speed even on a single core CPU. The faces are aligned using similarity transformation. Head orientation is estimated by Random Regression Forests [76] trained on AFLW dataset [77]. Body pose is tracked using Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields [78] which can run at 10 fps on an Nvidia GeForce GTX 1080 graphics card and can handle arbitrary poses, occlusion, and motion blur.

Facial expressions (sadness, happiness, surprise, disgust, anger) are estimated from aligned face regions using a CNN consisting of four convolution layers, two pooling layers, and three fully connected layers. The network achieves classification accuracy of 0.705 among five expression classes on the Facial Expression Recognition Challenge dataset [79]. More detailed facial information is extracted using the OpenFace toolkit [80] which implements a Constrained Local Neural Field (CLNF) deformable model for gaze tracking [81] and additional Support Vector Machine and Support Vector Regression models trained on the merged SEMAINE [82], DISFA [83], and BP4D [84] data for facial action unit detection.

Visual valence and arousal models were trained on the *RECOLA* database [63]. These models reuse activation features from the fully connected layers of the facial expression network (CNN-fc5 for the first fully connected layer and CNN-fc6 for the second). The per-frame features are compressed using PCA,[17] basic statistics are computed from a temporal window (mean, variance, minimum, maximum), and statistics from several neighboring frames are compressed again using PCA. The models built on these features are linear regressors trained with *Concordance Correlation Coefficients (CCC)*[18] objective function and weight decay. The results on the training and validation parts of the RECOLA database from AV + EC 2016 Challenge [85] are shown in Table VI.

D. Linked Data and Knowledge Graph

MixedEmotions Toolbox intends to exploit (emotion-related) information across different sources (i.e., emotion analysers for text, audio, video). To enable this capability, Linked Data principles have been investigated to define protocols and approaches to link the information of these sources to each other [86]. In the MixedEmotions Toolbox, the JSON Linked-Data (JSON-LD)

[14] opensmile.audeering.com
[15] https://github.com/openXBOW/openXBOW
[16] CCC is similar to the Correlation Coefficient, but it also considers the mean and variance of the two random variables.
[17] Principal Component Analysis.
[18] CCC is similar to the Correlation Coefficient, but it also considers the mean and variance of the two random variables [70].

TABLE VI
PERFORMANCE (CCC) OF THE VIDEO EMOTION RECOGNITION USING CNN ON AV + EC 2016 CHALLENGE DEVELOPMENT SET

	Valence	Arousal
Video Appearance	.474	.483
Video Geometric	.612	.379
CNN-fc5	.512	.532
CNN-fc6	.498	.585

Features video-appearance and video-geometric were provided as baselines by the challenge organizers.

format has been used for this task (Section III-D1). The use of linked data formats allows us to easily connect resources to common sense knowledge captured in knowledge graphs such as DBpedia [50] (Section III-D2).

1) Linked Data Representation: The MixedEmotions Toolbox follows a linked data approach in its services. The pillars of this approach are: (i) a representation model for all types of annotations covered by the toolbox (sentiments, emotions, suggestions), (ii) a means to uniquely identify annotations, (iii) a representation format to capture those annotations, (iv) a common interface for services within the toolkit to allow communication between them, and (v) a set of tools that unites all these aspects and enables the creation of new services. This section briefly covers these aspects, focusing on the representation.

The representation model includes all the concepts in the domain (*social post*, *entity*, *emotion*) and their properties or relationships (e.g., post *has* emotion, emotion *is of category* happy). Rather than creating an ad-hoc model for each domain, linked data principles encourage reusing already existing models. These models are also referred to as ontologies, vocabularies, or specifications. There are three vocabularies that are very relevant for sentiment and emotion annotation: Marl [87] (to annotate and describe subjective opinion), Onyx [88] (to annotate and describe emotions) with interoperability with Emotion Markup Language (EmotionML) [89] and NLP Interchange Format (NIF) 2.0 [90] (a semantic format and API for Natural Language Processing services). Moreover, the Onyx vocabulary provides a meta-model of emotions, i.e., instead of defining a set of categories or dimensions for emotions, it provides a meta-model so that different models can be defined and uniquely identified. It also contains definitions for the emotion models (vocabularies) in Emotion-ML and WordNet-Affect. Hence, annotators and service developers can be specific about what emotion models they are using (e.g., Ekman's big-6 categorical model,Russel's Circumplex model, etc).

Nevertheless, these models alone may not cover all the possible needs of possible use-cases for the MixedEmotions Toolbox. Therefore, additional concepts (e.g., suggestions, multi-results that include several entries and multimedia results) are defined, and the final proposed model (named the "MixedEmotions model") contains existing models (Marl, Onyx, NIF) and their extensions. This model uses NIF as the foundation for annotation of NLP results. NIF also provides different URI Schemes to identify text fragments inside a string, e.g., a scheme based on RFC5147 [91], and a custom scheme based on context. To this end, texts are converted to RDF[19] literals and a URI[20] is generated so that linked data annotations can be defined for that text. The same idea can also be applied to annotate multimedia [92]. The combination of Onyx's meta-models of emotion with the homogeneous multimedia annotation can be leveraged for automatic conversion and fusion of multimodal results [93].

To serialize these annotations, the MixedEmotions toolbox uses a common JSON-LD (JSON for Linked Data) schema. JSON-LD is a way of encoding Linked Data as JSON which provides a balance between semantic expression and ease of use for developers [94].

Moreover, this format is a good fit with the REST API that NIF defines for Natural Language Processing (NLP) services with standardized parameters. The MixedEmotions API adds several new concepts and parameters to those originally included in NIF, to cover the broad scope of the toolbox. It also establishes JSON-LD as its standard serialization format.

Lastly, these concepts are tightly integrated in the development kits and libraries provided by the MixedEmotions Toolkit. A notable example is Senpy,[21] a linked data framework for NLP services [95]. The aim of Senpy is to allow researchers to effortlessly turn their NLP analysis (e.g., sentiment and emotion analysis) into semantic web services. It also provides a series of common features that complement the services by leveraging their inherent semantics, such as automatic emotion model and format conversion, normalization of results and pipelining of several analysis. Senpy has been extensively used in the development of several modules of the MixedEmotions Toolbox.

Listing 1 (Supp. Materials: ex-linked-data.pdf) illustrates the semantic representation in a comprehensive example that includes multimodality (audio, video and text), fusion, and conversion of annotation. In particular, this example covers the analysis of the first two seconds of a video (located at `http://example.com/video.mp4`), and fusion of the three modalities. Since fusion requires all modalities to use the same dimensional emotion model, a conversion service exploits the semantic representation of emotion models in each annotation to find the appropriate conversion mechanism. As a result, the text results are converted form a categorical model to a dimensional one.

2) Knowledge Graph (KG): Knowledge graph theory uses graphs for the representation of concepts such as in medical and sociological texts [96]. The cumulation of such graphs can work as decision support system that can document the consequences of actions. The combination of knowledge graphs with concept models [97] led to the development of ontologies, which focused on the logical relations of concepts instead of words.

For a long time, knowledge graph theory was used for specific tasks: modeling of ecosystems or in linguistics for analyzing content of books [98]. Recently, the increasing popularity of linked data and the emergence of knowledge bases made large and general purpose knowledge graphs possible, such as Google's Knowledge Graph, which is a compilation of facts and

[19]Resource Description Framework
[20]Uniform Resource Identifier
[21]https://github.com/MixedEmotions/senpy

figures that provides contextual meaning to its searches [99]. In the MixedEmotions Toolbox, we provide a KG module that can be used to provide insights into relations between recognized entities using semantic knowledge from DBpedia [50]. The Entity Extraction and Linking module identifies entities mentioned in the analyzed resources, and then the KG module links them to the entities in DBpedia, so more specific information can be obtained about them (see Supp. Materials: ex-entities.png). Once the relations are extracted and filtered to keep the relevant ones only, they are stored in an Elasticsearch[22] database alongside other content metadata such as emotion annotations, where they can be readily visualized (e.g., using the Kibi graph browser, see below). The resulting KG contains the extracted entities, their specifications and related information, and the relations among them. The KG module is managed by a REST API, and needs an index in the Elasticsearch database that contains both the source text and the entities extracted. It can be queried by exploiting the REST API, so other modules can retrieve parts of the graph; in addition, it can be navigated through the Kibi graph browser.

The architecture of the KG module consists of five main parts: the Database, the DBpedia crawler, the Processing module, the Web server that exposes a REST interface, and the Kibi graph browser:

- *Database:* Elasticsearch repository stores information processed by other modules as well as KG module.
- *DBpedia crawler:* is responsible for crawling information from DBpedia, that is, related entities in the Database that are identified by the Entity Extraction and Linking module.
- *Processing module* filters the extracted information and splits it by type. The Entity Extraction and Linking module assigns one of the three types to the recognized entity: Person, Organization, and Location. Each type is processed separately so they can be stored in separate indexes. As the extracted information is not always 'clean' (it can wrongly be classified as a certain type of entity), the module applies customized filters for each type of entity to reduce the number of wrongly classified entities. Apart from writing the extracted information to the Database, the KG module automatically defines links between entities, adds the mapping of relations to Elasticsearch, and creates Kibi dashboards for each type of entity as well as a dashboard for the graph browser.
- *Web Server* allows monitoring and control of the KG module externally through a REST API.
- *Kibi graph browser* is a very powerful platform for interactive, exploratory big/streaming data discovery and alerting, with specific focus on exploration/leveraging of relationships across datasets. It performs 'on the fly' analytics on the collected entities and processed data stored in Elasticsearch. The Kibi graph browser provides the capability to visualize connections between entities and explore existing connections based on relations in DBpedia.

[22]Elasticsearch provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

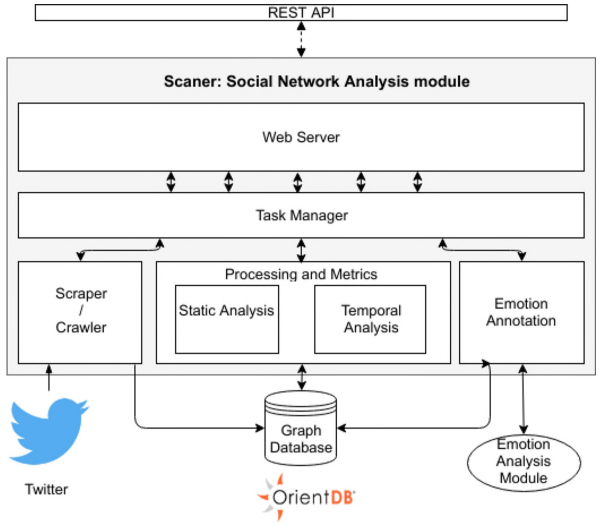

Fig. 3. Architecture of the Social Context Analysis module 'Scanner'.

E. Social Network Analysis

In general, sentiment and subjectivity are quite context-sensitive [100]; The meaning of a particular piece of content (e.g., a tweet, a Facebook status, or a blog post) may only be fully understood when its social context is taken into consideration. In fact, social context has an effect on the behaviour of users in social networks [101]. Recent work has demonstrated the existence of certain patterns in relationships in social media, which is explained by several social theories [102]. One notable example is social influence [103], which pertains to behavioural changes due to perceived relationships with other people, organizations and society in general.

Detecting and characterising social contexts and the emotions that are expressed therein has multiple applications. First, the detection of the most relevant shared content (e.g., tweets or posts), users (e.g., influencers), and groups of users (communities) provides a path for micro-analysis of opinions in brand monitoring [104] and content recommendation scenarios. Second, emotion propagation patterns can be used for both analysis and prediction of expected social influence of a message [105]–[108]. Those same patterns may also indicate false information or rumours [109]. Finally, social features can improve sentiment analysis and emotion detection [110], [111]. This can be specially relevant in microblogging based social networks such as Twitter, where the short length of the content makes the task very complex.

'Scanner' as a module in the MixedEmotions Toolbox that provides a standalone framework for crawling and analysing Twitter contents to perform social network and emotion analysis. It is capable of calculating different social metrics (e.g., content metrics, group metrics, temporal metrics, influence metrics). The architecture of this module is depicted in Fig. 3.

F. Decision Fusion

Since within the MixedEmotions Toolbox, emotions can be extracted from diverse modalities (video, audio, text) and sources, there is a need to combine extracted results and yield

a final (more reliable) estimate. For this, the decision fusion module accepts the outputs (in the MixedEmotions JSON-LD format) of modules that represent emotions in terms of continuous arousal and valence (irrespective of modality), and combines them by a weighted average of the values. The choice of classifier fusion (vs. feature fusion) is to keep modules independent of each other, and the choice of weighted average is because each modality may contribute differently to recognizing emotions (for example, it is known that valence can be recognized better via facial monitoring, while arousal can be recognized better via speech monitoring). Weights can be learned offline, set manually, or have the same values.

IV. Use Cases of the MixedEmotions Toolbox

The MixedEmotions Toolbox has been tested in the context of three concrete use cases (Emotion-driven Smart TV, Brand Reputation Analysis, Call Center Monitoring) to verify its usefulness.

A. Emotion-Driven Smart TV

In this use case, an emotion-driven recommendation engine is developed. The purpose of this engine is to use emotion signals to enhance traditional content- and user-based recommendations for TV programs. More specifically, the Apache Mahout open-source recommender in conjunction with video material published by the broadcaster Deutsche Welle is fed with emotion predictions of the MixedEmotions Toolbox. For each of Deutsche Welle's videos, the following contents are used for the emotion analyzer:

- the video's title and description text
- the transcription of the video's soundtrack[23]
- twitter messages relating to the video's topics
- the video's soundtrack itself

For each of these contents, the distribution of emotions was calculated using MixedEmotions emotion detection modules for the appropriate modality and fed into the recommendation engine. This was done alongside classical features such as keywords and the percentage of the video duration that the viewers actually watched.

The resulting recommendations were used to present viewers of Deutsche Welle's Apple TV application with suggestions of video contents to watch from two categories:

1) Eudaimonic content intriguing/challenging videos
2) Hedonic content joyful/entertaining videos.

These categories are based on recent research into media consumption [112], [113]. The idea is to give viewers the possibility to choose from these two distinct categories depending on their current mood, where they either prefer purely joyful content (e.g., travel and lifestyle) or more intriguing content (e.g., documentaries about conflicts or confrontational interviews). In the Supplementary Materials (ex-social-tv.png), a snapshot of the emotion analysis of videos of Deutsche Welle's programs after fusing transcription, audio, and tweet analysis is presented.

TABLE VII
The Average Percentage of the Suggested Videos Watched By Users

Mood Category	Without Emotions	With Emotions
Hedonic	88%	**99**%
Eudaimonic	86%	**92**%

In this case, the fusion is based on the collective histograms from different modalities. If the histogram is skewed toward positive valence, the content is Hedonic, and if it is skewed toward positive arousal, it is considered as Eudaimonic.

An A/B test is conducted to verify whether the addition of emotion signals helps to identify videos that a viewer is more likely to prefer and therefore watch to the end. 1227 users registered and 79 videos were selected as part of the experiment. After a user watches a video, a user has the option to classify it as Eudaimonic or Hedonic and the recommendation engine prepares two sets of videos based on their Eudaimonic and Hedonic contents. A 'hit' is counted when the user selects a video from the same emotional content as the previously-shown video. Overall, 9060 videos were watched. The results are presented in Table VII, and shows that, users tend to watch videos that were proposed by the emotion-driven recommendation engine to a fuller extent (99% of Hedonic and 92% of the Eudaimonic video's total duration was actually watched) compared to videos where the recommendation engine did not make use of emotion predictions (88% and 86% respectively). Moreover, the performance of classification (Eudaimonic or Hedonic) is 86%.

In another study, also we investigated if acoustic-based emotional features of a video can help to predict the popularity of that video [114]. We have used the 'Audio processing' module to extract acoustic features. We could achieve 70% accuracy on recognizing popular vs. non-popular content only using seven features. For more information please refer to [114].

B. Call Center Monitoring

Call center Monitoring is the second use-case, which mostly relies on emotion analysis from speech. Call centers offer a promising natural space for emotion mining and analysis. On a daily basis, each agent in a call center encounters customers with different emotions and moods. Recognition of these emotions will help to write better scripts for call center agents that can soothe negative emotions and lead to higher customer satisfaction.

To embed the emotion analysis functionality into this use case, three approaches were considered: (i) acoustic-based valence recognition with multilingual and Czech models,[24] (ii) analysis of the automatic transcription of the audio, based on the list of pre-defined positive and negative keywords and phrases, and (iii) sentiment recognition on the translation of the transcriptions using the statistical machine translation (SMT) module (Section III-A4b). This later approach is an extension of

[23]using https://github.com/MixedEmotions/MixedEmotions/wiki/m17.-Speech-to-text-by-Phonexia

[24]https://github.com/MixedEmotions/MixedEmotions/wiki/m23.-Audio-Emotion-extraction-by-Phonexia

TABLE VIII
PERFORMANCE OF DIFFERENT APPROACHES FOR SENTIMENT RECOGNITION (3-CLASS TASK) IN TERMS OF UNWEIGHTED AVERAGE RECALL (UAR), EVALUATED ON TWO CZECH CALL CENTERS

Method	language	Call Center 1	Call Center 2
(i) acoustic	Multi-lingual	.344	.431
(i) acoustic	Czech	.370	.449
(ii) keywords	Czech	.381	.359
(iii) sentiment	English	**.438**	**.496**

approach (ii) using methods of natural language processing that consider also the context of the utterance. In this case, we used the Phonexia sentiment analyzer, which is a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) fed with word2vec word embeddings.[25] The system produces the posterior probability of positive sentiment for each sentence, which is then mapped to one of the sentiment classes (positive, negative or neutral). In the Supplementary Materials (ex-call-center.png), we provided a snap shot of this tool.

Acoustic-, keyword- and sentiment-based systems were evaluated on Czech call center data. Transcriptions were automatically translated to English so that the above-mentioned English sentiment analyzer (which is trained on English corpora) can be applied. The results for 3-class sentiment recognition (positive, neutral, negative) are provided in Table VIII. As the results suggest, sentiment analysis on the translated transcriptions outperforms the acoustic- and keyword-based systems.

C. Brand Reputation Analysis

Brand Reputation analysis is the third use-case that uses the MixedEmotions Toolbox to implement an application for the assessment of the perceived reputation of a brand or product on the web. Its main objective is to mine selected sources of information and provide human interpretable results that can be investigated by the person in charge of the brand.

This use-case monitors Twitter and YouTube, and processes textual and audio contents to evaluate sentiments and emotions. Entities and the distribution of languages are also extracted. Human-readable results are visualized at real-time using Kibi to compare between different brands and to study emotions and sentiments regarding different dimensions such as hashtags, YouTube channels, or locations. A snapshot of the Kibi for emotion distribution for a Brand is provided in the Supp. materials (ex-brand-reputation.png).

V. CONCLUSION

In this paper, we introduced a free, open-source, and multimodal toolbox for emotion analysis: the 'MixedEmotions Toolbox'. The toolbox includes functionalities for text, audio, and video processing with the aim of emotion recognition. Three use cases were described: Emotion-driven Smart TV (emotion-based recommendation), Brand Reputation Analysis (monitoring reputation of a brand from tweets and YouTube videos), and Call Centre Monitoring (monitoring emotion of customers in a help-desk setting). In the future, we hope to see contributions to the release and will ourselves update further functionality aiming beyond improved robustness and increases in efficiency — multimedia data is often 'big', but it is always emotional!

[25] http://www.fit.vutbr.cz/∼imikolov/rnnlm/

REFERENCES

[1] K. Zhang *et al.*, "A probabilistic graphical model for brand reputation assessment in social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining*, Aug. 2013, pp. 223–230.

[2] L. Pang, S. Zhu, and C. W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, Nov. 2015.

[3] S. E. Shepstone, Z. H. Tan, and S. H. Jensen, "Using audio-derived affective offset to enhance tv recommendation," *IEEE Trans. Multimedia*, vol. 16, no. 7, pp. 1999–2010, Nov. 2014.

[4] I. Arapakis *et al.*, "Integrating facial expressions into user profiling for the improvement of a multimodal recommender system," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jun. 2009, pp. 1440–1443.

[5] F. Huang *et al.*, "Overlapping community detection for multimedia social networks," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1881–1893, Aug. 2017.

[6] R. A. Negoescu and D. Gatica-Perez, "Modeling Flickr communities through probabilistic topic-based analysis," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 399–416, Aug. 2010.

[7] M. Tkalčič, A. Odić, A. Košir, and J. Tasič, "Affective labeling in a content-based recommender system for images," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 391–400, Feb. 2013.

[8] J. Tarvainen, M. Sjöberg, S. Westman, J. Laaksonen, and P. Oittinen, "Content-based prediction of movie style, aesthetics, and affect: Data set and baseline experiments," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2085–2098, Dec. 2014.

[9] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[10] K. R. Scherer, "What are emotions? and how can they be measured?" *Soc. Sci. Inf.*, vol. 44, no. 4, pp. 695–729, Dec. 2005.

[11] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *J. Personality Social Psychol.*, vol. 17, no. 2, p. 124, 1971.

[12] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theories Emotion*, vol. 1, no. 3–31, pp. 3–33, 1980.

[13] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological Rev.*, vol. 110, no. 1, p. 145, 2003.

[14] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development Psychopathol.*, vol. 17, no. 03, pp. 715–734, 2005.

[15] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Urbana, IL, USA: Univ. Illinois Press, 1957.

[16] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2007.

[17] M. Arcan, J. P. McCrae, and P. Buitelaar, "Expanding WordNets to new languages with multilingual sense disambiguation," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 97–108.

[18] P. Vossen, F. Bond, and J. P. McCrae, "Toward a truly multilingual global wordnet grid," in *Proc. Global WordNet Conf.*, 2016, pp. 419–427.

[19] S. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Proc. 6th Biannual Conf. Affective Comput. Intell. Interaction*, Xi'an, China, Sep. 2015, pp. 125–131.

[20] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Proc. 41st Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 5800–5804.

[21] A. Popková *et al.*, "Investigation of bottle-neck features for emotion recognition," in *Proc. 19th Int. Conf. Text, Speech, Dialog.*, Sep. 2016, pp. 426–434.

[22] H. Sagha *et al.*, "Enhancing multilingual recognition of emotion in speech by language identification," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2016, pp. 2949–2953.

[23] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2016, pp. 3593–3597.

[24] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 314–327, Jul.–Sep. 2016.

[25] G. Trigeorgis *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. 41st Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 5200–5204.

[26] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT: The Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *J. Mach. Learn. Res.*, vol. 16, pp. 547–551, 2015.

[27] S. Hantke, E. Marchi, and B. Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification," in *Proc. 10th Lang. Resources Eval. Conf.*, May 2016, pp. 2156–2161.

[28] Z. Zhang *et al.*, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proc. 41st Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 5185–5189.

[29] P. Vassiliadis, "A survey of extract–transform–load technology," *Int. J. Data Warehousing Mining*, vol. 5, no. 3, pp. 1–27, 2009.

[30] H. Sagha, N. Cummins, and B. Schuller, "Stacked denoising autoencoders for sentiment analysis: A review," *Wiley Interdisciplinary Rev., Data Mining Knowl. Discovery*, vol. 7, no. 5, 2017, Art. no. e1212.

[31] C. Strapparava and A. Valitutti, "WordNet-Affect: An affective extension of WordNet," in *Proc. 4th Int. Conf. Lang. Resources Eval.*, vol. 4, 2004, pp. 1083–1086.

[32] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *Proc. 6th Int. Conf. Lang. Resources Eval.*, Genoa, Italy, 2006, pp. 417–422.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, vol. 1631, 2013, p. 1642.

[35] P. Nakov *et al.*, "Semeval-2013 task 2: Sentiment analysis in twitter," in *Proc. Joint Conf. Lexical Computational Semantics*, Atlanta, GA, USA, vol. 312, 2013.

[36] S. Rosenthal *et al.*, "Semeval-2015 task 10: Sentiment analysis in twitter," in *Proc. 9th Int. Workshop Semantic Eval., SemEval*, Denver, CO, USA, 2015, pp. 451–463.

[37] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168–177.

[38] C. D. Manning *et al.*, "The Stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist., System Demonstration*, 2014, pp. 55–60.

[39] Alias-i, "Lingpipe 4.1.0." Tech. Rep. 2016. [Online]. Available: http://alias-i.com/lingpipe

[40] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. AAAI Conf. Weblogs Soc. Media*, Oxford, U.K., 2014, pp. 216–225.

[41] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Rep., Stanford, CA, USA, vol. 1, p. 12, 2009.

[42] J. V. Romn *et al.*, "Overview of TASS 2015," in *Proc. Workshop Sentiment Anal. SEPLN*, vol. 1397, 2015, pp. 13–21.

[43] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Tech. Rep. C-1, The Center Res. Psychophysiol., Univ. Florida, Gainesville, FL, USA, 1999.

[44] J. Redondo, I. Fraga, I. Padrn, and M. Comesaa, "The Spanish adaptation of ANEW (Affective Norms for English Words)," *Behavior Res. Meth.*, vol. 39, no. 3, pp. 600–605, 2007.

[45] S. M. Mohammad, "# Emotional tweets," in *Proc. 1st Joint Conf. Lexical Comput. Semantics*, 2012, pp. 246–255.

[46] V. Andryushechkin, I. D. Wood, and J. O'Niel, "NUIG at EmoInt-2017: BiLSTM and SVR Ensemble to Detect Emotion Intensity" in *Proc. Workshop Comput. Approaches Subjectivity, Sentiment Soc. Media Anal.*, Copenhagen, Denmark, 2017, pp. 175–179.

[47] S. M. Mohammad and F. Bravo-Marquez, "WASSA-2017 shared task on emotion intensity," in *Proc. 8th Workshop Comput. Approaches Subjectivity, Sentiment Soc. Media Anal.*, Copenhagen, Denmark, 2017, pp. 34–49.

[48] A. Graves, S. Fernndez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. Int. Conf. Artif. Neural Netw., Formal Models Their Appl.*, 2005, pp. 799–804.

[49] S. Negi, K. Asooja, S. Mehrotra, and P. Buitelaar, "A study of suggestions in opinionated texts and their automatic detection," in *Proc. 5th Joint Conf. Lexical Comput. Semantics*, Aug. 2016, pp. 170–178.

[50] S. Auer *et al.*, "DBpedia: A nucleus for a web of open data," in *Proc. 6th Int. Semantic Web 2nd Asian Conf. Asian Semantic Web Conf.*, 2007, pp. 722–735.

[51] C. Strapparava and R. Mihalcea, "SemEval-2007 task 14: Affective text," in *Proc. 4th Int. Workshop Semantic Eval.*, 2007, pp. 70–74.

[52] M. M. Bradley and P. J. Lang, "Affective norms for English text (ANET): Affective ratings of texts and instruction manual," Univ. Florida, Gainesville, FL, USA, Tech. Rep. D-1, 2007.

[53] D. Preotiuc-Pietro *et al.*, "Modelling valence and arousal in Facebook posts," in *Proc. 15th Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, San Diego, CA, USA, 2016, pp. 9–15.

[54] S. Buechel and U. Hahn, "EMOBANK: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *Proc. Eur. Chapter Assoc. Comput. Linguist.*, 2017, p. 578.

[55] L.-C. Yu *et al.*, "Building Chinese affective resources in valence-arousal dimensions," in *Proc. 15th Annu. Conf. North Amer. Ch. Assoc. Comput. Linguistics, Human Lang. Technol.*, San Diego, CA, USA, 2016, pp. 540–545.

[56] I. D. Wood, J. P. McCrae, V. Andryushechkin, and P. Buitelaar, "A comparison of emotion annotation schemes and a new annotated data set," in *Proc. 11th Ed. Lang. Resources Eval. Conf.*, Miyazaki, Japan, 2018.

[57] I. Wood and S. Ruder, "Emoji as emotion tags for tweets," in *Proc. 10th Ed. Lang. Resources Eval. Conf.*, Portorož, Slovenia, 2016, pp. 76–79.

[58] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA:Bradford Books, 1998.

[59] P. Vossen, Ed., *EuroWordNet: A Multilingual Database With Lexical Semantic Networks*. Norwell, MA, USA: Kluwer, 1998.

[60] D. Tufiş, D. Cristea, and S. Stamou, "Balkanet: Aims, methods, results and perspectives. A general overview," *Romanian J. Inf. Sci. Technol.*, vol. 7, no. 1/2, pp. 9–43, 2004.

[61] E. Pianta, L. Bentivogli, and C. Girardi, "MultiWordNet: Developing an aligned multilingual database," in *Proc. 1st Int. Conf. Global WordNet*, Mysore, India, Jan. 2002, pp. 291–302.

[62] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc.*, Sep 2016, pp. 495–499.

[63] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 2nd Int. Conf. Workshop Emotion Representation, Anal. Synthesis Continuous Time Space*, Shanghai, China, 2013, pp. 1–8.

[64] B. Schuller, *Intelligent Audio Analysis*, (ser. Signals and Communication Technology). New York, NY, USA: Springer, 2013.

[65] B. Schuller, A. E.-D. Mousa, and V. Vasileios, "Sentiment analysis and opinion mining: On optimal parameters and performances," *WIREs Data Mining Knowl. Discovery*, vol. 5, pp. 255–263, Sep./Oct. 2015.

[66] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st Int. Conf. Multimedia.*, Oct 2013, pp. 835–838.

[67] S. Rawat *et al.*, "Robust audio-codebooks for large-scale event detection in consumer videos," in *Proc. 14th Int. Speech Commun. Assoc.*, 2013, pp. 2929–2933.

[68] M. Schmitt and B. W. Schuller, "openXBOW—Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *J. Mach. Learn. Res.*, vol. 18, pp. 96:1–96:5, 2016.

[69] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.

[70] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[71] F. Ringeval *et al.*, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, 2017, pp. 3–9.

[72] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–45, Sep. 2010.

[73] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp.1–12.

[74] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. 59th Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4873–4882.

[75] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. 57th Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[76] A. Pavelkov, A. Herout, and K. Behn, "Usability of pilot's gaze in aeronautic cockpit for safer aircraft," in *Proc. 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 1545–1550.

[77] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *Proc. Int. Conf. Comput. Vis. Workshops*, Barcelona, Spain, 2011, pp. 2144–2151.

[78] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. 60th Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[79] I. J. Goodfellow *et al.*, "Challenges in representation learning: A report on three machine learning contests," *Neural Netw.*, vol. 64, pp. 59–63, 2015.

[80] T. Baltruaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. Winter Conf. Appl. Comput. Vis.*, Mar. 2016, pp. 1–10.

[81] E. Wood *et al.*, "Rendering of eyes for eye-shape registration and gaze estimation," in *Proc. Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 3756–3764.

[82] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. Int. Conf. Multimedia Expo.*, Jul. 2010, pp. 1079–1084.

[83] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013.

[84] X. Zhang *et al.*, "BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database," *Image Vis. Comput.*, vol. 32, no. 10, pp. 692–706, 2014.

[85] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge.*, 2016, pp. 3–10.

[86] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Semantic Services, Interoperability Web Appl., Emerging Concepts*, vol. 5, pp. 205–227, 2009.

[87] A. Westerski, C. A. Iglesias, and F. Tapia, "Linked opinions: Describing sentiments on the structured web of data," in *Proc. 4th Int. Workshop Social Data Web.*, Oct. 2011, pp. 21–32.

[88] J. F. Sánchez-Rada and C. A. Iglesias, "Onyx: A linked data approach to emotion representation," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 99–114, 2016.

[89] M. Schrder *et al.*, "EmotionML—An upcoming standard for representing emotions and related states," in *Affective Computing and Intelligent Interaction*, (ser. Lect. Notes Comput. Sci.), S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Germany: Springer, 2011, vol. 6974, pp. 316–325.

[90] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer, "Integrating NLP using linked data," in *Proc. Int. Semantic Web. Conf.*, 2013, pp. 98–113.

[91] E. Wilde and M. Duerst, "URI fragment identifiers for the text/plain media type," Internet Engineering Task Force, Apr. 2008.

[92] J. F. Sánchez-Rada, C. A. Iglesias, and R. Gil, "A linked data model for multimodal sentiment and emotion analysis," in *Proc. 4th Workshop Linked Data Linguistics*, Beijing, China, Jul. 2015, pp. 111–116.

[93] J. F. Sánchez-Rada *et al.*, "Multimodal multimodel emotion analysis as linked data," in *Proc. ACII*, San Antonio, TX, USA, Oct. 2017.

[94] M. Lanthaler and C. Gütl, "On using JSON-LD to create evolvable RESTful services," in *Proc. 3rd Int. Workshop RESTful Des.*, 2012, pp. 25–32.

[95] J. F. Sánchez-Rada and C. A. Iglesias, "Senpy: A pragmatic linked sentiment analysis framework," in *Proc. Special Track Emotion Sentiment Intell. Syst. Big Soc. Data Anal.*, Oct. 2016, pp. 735–742.

[96] R. Bakker, "Knowledge graphs: Representation and structuring of scientific knowledge," Ph.D. dissertation, Univ. Twente, Enschede, The Netherlands, 1987.

[97] J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA, USA: Addison-Wesley, Jan. 1983.

[98] C. Hoede, "Modelling knowledge in electronic study books," *J. Comput. Assisted Learn.*, vol. 10, no. 2, pp. 104–112, 1994.

[99] A. Singhal, "Introducing the knowledge graph: Things, not strings," Official Google blog, 2012.

[100] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations Trends Inf. Retrieval*, vol. 2, no. 1/2, pp. 1–135, 2008.

[101] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proc. Nat. Acad. Sci.*, pp. 8788–8790, 2014.

[102] J. Tang, Y. Chang, and H. Liu, "Mining social media with social theories: A survey," *SIGKDD Explorations Newslett.*, vol. 15, no. 2, pp. 20–29, Jun. 2014.

[103] C. Tan *et al.*, "User-level sentiment analysis incorporating social networks," in *Proc. 17th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1397–1405.

[104] W. Deitrick and W. Hu, "Mutually enhancing community detection and sentiment analysis on twitter networks," *J. Data Anal. Inf. Process.*, vol. 1, no. 3, pp. 19–29, 2013.

[105] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proc. 4th Int. Conf. Web Search Data Mining*, 2011, pp. 177–186.

[106] Y. Artzi, P. Pantel, and M. Gamon, "Predicting responses to microblog posts," in *Proc. Conf. North Amer. Chap. Assoc. Comput. Linguist., Human Lang. Technol.*, 2012, pp. 602–606.

[107] S. Alhabash and A. R. McAlister, "Redefining virality in less broad strokes: Predicting viral behavioral intentions from motivations and uses of Facebook and Twitter," *New Media Soc.*, vol. 17, no. 8, pp. 1317–1339, 2015.

[108] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 925–936.

[109] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi, "The anatomy of a scientific rumor," *Sci. Rep.*, vol. 3, 2013, Art. no. 2980.

[110] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proc. Conf. Empirical Meth. Natural Lang. Process.*, 2011, pp. 53–56.

[111] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *Proc. 6th Int. Conf. Web Search Data Mining*, 2013, pp. 537–546.

[112] M. B. Oliver and A. A. Raney, " Entertainment as pleasurable and meaningful: Identifying hedonic and eudaimonic motivations for entertainment consumption," *J. Commun.*, vol. 61, no. 5, pp. 984–1004, 2011.

[113] R. J. Lewis, R. Tamborini, and R. Weber, " Testing a dual-process model of media enjoyment and appreciation," *J. Commun.*, vol. 64, no. 3, pp. 397–416, 2014.

[114] H. Sagha, M. Schmitt, F. Povolny, A. Giefer, and B. Schuller, "Predicting the popularity of a talk-show based on the emotionality of its speech content," in *Proc. 3rd Int. Workshop Affective Social Multimedia Comput./18th Annu. Conf. Int. Speech Commun. Assoc.*, Aug. 2017, p. 5.