A Content and Knowledge Management System Supporting Emotion Detection from Speech

Binh Vu, Mikel deVelasco, Paul Mc Kevitt, Raymond Bond, Robin Turkington, Frederick Booth, Maurice Mulvenna, Michael Fuchs, and Matthias Hemmje

Abstract Emotion recognition has recently attracted much attention in both industrial and academic research as it can be applied in many areas from education to national security. In healthcare, emotion detection has a key role as emotional state is an indicator of depression and mental disease. Much research in this area focuses on extracting emotion related features from images of the human face. Nevertheless, there are many other sources that can identify a person's emotion. In the context of MENHIR, an EU-funded R&D project that applies Affective Computing to support people in their mental health, a new emotion-recognition system based on speech is being developed. However, this system requires comprehensive data-management support in order to manage its input data and analysis results. As a result, a cloud-

Binh Vu
FernUniversität in Hagen, Hagen, Germany, e-mail: binh.vu@fernuni-hagen.de

Mikel deVelasco
Universidad del Pais Vasco UPV/EHU, Leioa, Spain, e-mail: mikel.develasco@ehu.eus

Paul Mc Kevitt
Ulster University, Derry/Londonderry, Northern Ireland, e-mail: p.mckevitt@ulster.ac.uk

Raymond Bond
Ulster University, Newtownabbey, Northern Ireland, e-mail: rb.bond@ulster.ac.uk

Robin Turkington
Ulster University, Newtownabbey, Northern Ireland, e-mail: turkington-r@ulster.ac.uk

Frederick Booth
Ulster University, Newtownabbey, Northern Ireland, e-mail: booth-f@ulster.ac.uk

Maurice Mulvenna
Ulster University, Newtownabbey, Northern Ireland, e-mail: md.mulvenna@ulster.ac.uk

Michael Fuchs
GLOBIT GmbH, Barsbüttel, Germany, e-mail: m.fuchs@globit.com

Matthias Hemmje
GLOBIT GmbH, Barsbüttel, Germany, e-mail: matthias.hemmje@globit.com

based, high-performance, scalable, and accessible ecosystem for supporting speech-based emotion detection is currently developed and discussed here.

1 Introduction and Motivation

Affective Computing is an emerging inter-disciplinary field developing technology that attempts to detect, analyse, process, and respond to important psychological traits such as emotions, feelings, or behaviours with the goal of improving human-computer interaction [1]. Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) is a 4-year project funded by the European Union (EU), that applies Affective Computing to enhance and advance future healthcare processes and systems, especially in providing assistance to people with dementia, medical professionals, and caregivers [2]. By gathering activity and related sensor data to infer the emotional state of the patient as a knowledge stream of emotional signals, SenseCare can provide a basis for enhanced care and can alert medics, professional carer, and family members to situations where intervention is required [3] [4].

One of the systems developed in SenseCare is a machine-learning-based emotion detection platform [5] which provides an early insight into the emotional state of an observed person. SenseCare can process a live video stream or a pre-recorded video which enables analysis to be completed on the fly or at a later stage. Similar to SenseCare, the MENtal Health monitoring through InteRactive conversations (MENHIR) is a EU-funded project that aims to support and improve the mental wellbeing of people by applying Affective Computing, especially conversational technologies, such as emotion recognition in speech, automatic conversation management (chatbots), and other multidisciplinary topics [6]. According to the World Health Organization (WHO), mental, neurological, and substance use disorders make up 10% of the global, and 30% of non-fatal, disease burden. The global economy loses about US$ 1 trillion per year in productivity due to depression and anxiety [7].

In MENHIR, new research assists people with improving their current state of emotion and provides a long-term overview of their state over time. A machine-learning-based emotion detection platform has been developed. Unlike SenseCare, where human emotions are extracted from a live video stream or a pre-recorded video, the MENHIR emotion detection platform identifies emotions from speech. The system relies on short-term features such as pitch, vocal tract features such as formants, prosodic features such as pitch loudness, as well as speaking rate to perform effectively. Furthermore, recurrent neural networks are applied to predict emotion in real-time using a 3D emotional model. This paper discusses the challenges of emotion detection based on speech and its corresponding transcription in the MENHIR project. Furthermore, it provides a solution to overcome these challenges. The architecture of the proposed system and its constituent components are described. Finally, we conclude and discuss future work.

2 Problem Statement

One of the goals of the MENHIR project is to further extend the results of earlier research work, expanding the set of identified depressive speech acoustic features and automating their detection so that depressed and anxious speech can be accurately distinguished from healthy speech [8]. To enable this, challenging scenarios need to be considered and overcome as discussed here.

After a series of human-to-human counselling conversations are recorded in a laboratory setting, a corpus of audio data of conversations is formed. Along with the audio files, their metadata, which consists of documents describing the conversations and spreadsheets describing the conversation results, are also provided for advanced annotation and analysis. All these data need to be stored in a high-performance repository where other analysis systems can connect to and download them when needed. Furthermore, multimedia objects usually take up a lot of storage space. This means the data repository also needs to be scalable to fulfil users' demands in the future.

In MENHIR, not only multimedia objects but also other kinds of scientific content, knowledge, and their metadata need to be imported, stored, and managed. Sharing and exchanging research results powers collaborative and co-creative networking among project participants. Therefore, a solution is needed to support the ingestion of scientific publications from different sources. Here, the imported content can be managed and transformed into learning materials. Similar to multimedia objects, scientific data content also needs high-performance, scalable, and fault-tolerant storage. Furthermore, a content management system will enable users to edit, share, and publish their content.

There are a number of collaborative services producing analysis results and generating observed subject and patient conversational behaviour, such as, authentication, authorization, speech analysis data services, big data speech analysis, collaboration and coordination services, psychological/affective analytics, reporting/result sharing and reproducibility services [8]. It is crucial to have an integration architecture for all the mental health services and applications employed in MENHIR. This architecture will provide a common platform for these systems to communicate in a predefined flow, where input data is received and results are stored.

For research results to make an impact, they need to be easily found and used. Meanwhile, related publications, datasets, and analysis results are distributed in different locations. Therefore, one needs to find a means to automatically gather and combine all these resources into scientific asset packages. Otherwise, users can only find fragments of related information. It will prevent them from having a complete overview of the research topic and discovering important relationships between factors. Organizing related information and data into scientific asset packages is a powerful method of systemizing results produced by conversational technologies.

Finally, classification helps to narrow the choices among content, information, and knowledge resources. By dividing the material into reduced subsets, classification can make content, information, and knowledge resource retrieval and access faster and more accurate [9]. In MENHIR, a considerable volume of subject data will be analysed by an emotion detection server and will be made available for use by, for example, chatbots. Furthermore, the analysis results and related scientific publications will also be generated and managed in MENHIR. Without organizing the content created into suitable categories, researchers will not have capacity for insight on the key data produced in MENHIR, discover connections between data or whether something is missing. Therefore, a system that allows the content, knowledge, analysis results, and datasets to be classified is critical for the success of MENHIR.

3 System Design

Based on these challenges, we have developed a system design to support MENHIR in the task of conversational technologies research and development. In this section, a cloud-based Content and Knowledge Management Ecosystem (KM-EP) for audio files and metadata persistence, human emotion detection, as well as asset packaging, classification, and management is introduced and described. Figure 1 illustrates the architecture of the system, which comprises the MENHIR KM-EP and supporting systems.

The MENHIR Content and Knowledge Management Ecosystem (KM-EP) provides a platform for managing scientific as well as educational content and knowledge resources. Furthermore, the KM-EP will act as a framework for researchers to deploy their work without spending time reimplementing basic functionalities, such as, e.g. user management and task scheduling. In Figure 1, four components of the MENHIR KM-EP, which are related and crucial for the tasks of audio data persistence, emotion detection, as well as asset packaging, classification, and management, are shown. The components are Media Archive (MA), Digital Library (DL), Taxonomy Manager (TM), and Asset Manager (AM).

The first component that is needed for the MENHIR KM-EP is the Media Archive (MA). The MA manages all multimedia objects in the KM-EP. The MA enables users to create, persist, manage, and classify different types of multimedia objects, such as, e.g. video, audio, images, presentation slides, along with their metadata. In MENHIR, audio files and their metadata need to be imported and stored together in the system. The initial audio files are, e.g., recordings of interviews, which are conducted in order to form a corpus of conversational audio data. This corpus will be used to validate the operation of the Emotion Detection Server. Their metadata consists, e.g., of documents describing the interviews and spreadsheets describing the interview results. Furthermore, audio files of conversations and interviews can also be uploaded and linked to user accounts automatically or manually. The KM-EP provides an interface where users can upload these files into the system and populate basic metadata information related to them, such as, e.g. title, description, authors, and creation date. The uploaded files are stored in a cloud storage service, which is fault-tolerant, flexible, scalable, and has high performance. This will enable users to have fast and stable access to the files worldwide. Furthermore, with

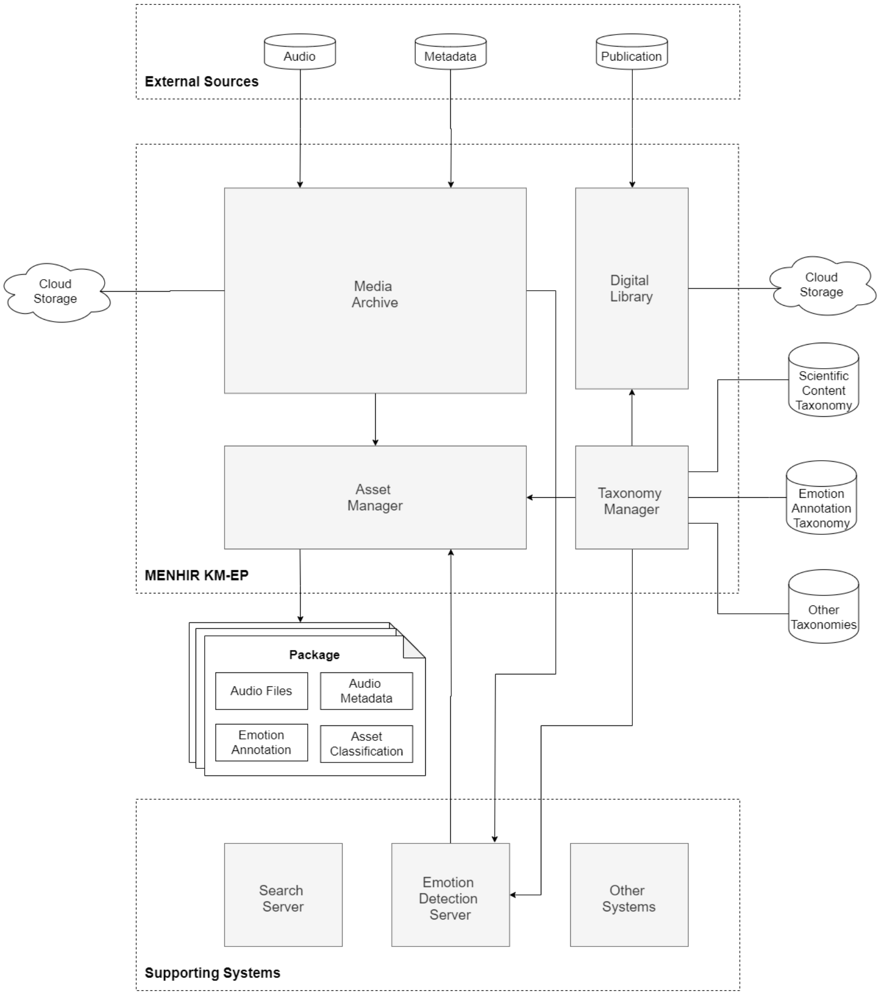

Fig. 1 Architecture of MENHIR Content and Knowledge Management Ecosystem (KM-EP)

the support of the TM component, multimedia objects can be classified into different categories. Classification enables objects to be searched and accessed easily and quickly by users.

The next component of the KM-EP is the Digital Library (DL). The DL enables users to import publications into the KM-EP, persist, and manage them. Using a Mediator-Wrapper Architecture, publications from different sources, such as, e.g. Mendeley [10], SlideShare [11], and in different formats, such as BibTex [12] and OAI-PMH [13], can be queried, uploaded, and integrated into the DL [14]. Similar to the MA, after importing or creating a new publication using the DL, users have the option to fill in its metadata, such as, e.g. title, abstract, conference, publisher, and upload the document. The uploaded files will also be stored in cloud storage to maintain their availability and scalability. By indexing file metadata and classifying publications into existing categories, they can be searched by users based on these criteria.

The Taxonomy Manager (TM) component supports the construction, collaboration, management, and evolution of taxonomies. The TM enables users to develop and manage their own taxonomies. With the support of its version control system, users can manage the changes of their taxonomies. Every modification is tracked and can be reverted. A complete history of changes helps users to compare different versions of a taxonomy. Furthermore, the branching feature enables users to create multiple versions of a taxonomy.

Multimedia objects, publications, and assets of the MENHIR KM-EP can be classified with support of the TM. As a result, users can search and browse contents quickly and easily. Classification also enables navigation inside the KM-EP. A persistent identifier introduced for each term in a taxonomy enables taxonomy evolution without affecting existing classifications. A rating system is implemented based on crowd voting to support the evaluation of taxonomies in the KM-EP. With the rating system, authors can improve the accessibility of their taxonomies, and users can also choose quickly more relevant taxonomies. A caching system enables thousands of taxonomies and terms to be retrieved and constructed in just a few milliseconds.

In MENHIR, the TM can not only be used to collect, classify, and provide access to audio materials from initial emotion analysis and results but can also support the emotion detection platform by providing an emotion annotation taxonomy. The machine learning platform can use this taxonomy to label its training and validation set. This creates a standard emotion classification that can be used for classifying results produced later by the platform. This process would be more costly without the classification, annotation, and access support of the TM in the MENHIR KM-EP supporting scientific research in the domain of Affective Computing.

The Asset Manager (AM) component is where related data, metadata, analysis results, and classification are gathered and combined into packages. In order to do this, a cronjob is developed and scheduled to run regularly after a given period of time. This cronjob has 3 tasks, which are: (1) searching for new audio files and their metadata and adding them into a new asset, (2) sending the new audio files and their metadata to the emotion detection server for analysis, and (3) receiving and adding analysis results into its package. This guarantees that new data will always be processed after it is uploaded to the MENHIR KM-EP.

After the cronjob has been started, the daemon searches for audio files along with their metadata in the MA. For each audio record found, the daemon will check if it belongs to a package in the AM or not. If it exists in a package, the emotion detection process, along with other processes, has been already completed for this audio record and the daemon continues to work with other audio records. Otherwise, the daemon needs to gather necessary data, such as uploaded files, documents describing the counseling interview where the audio file was recorded, and the spreadsheets describing the interview results. These files will be downloaded from the current cloud storage service to a temporary location in the local server. Next, the daemon

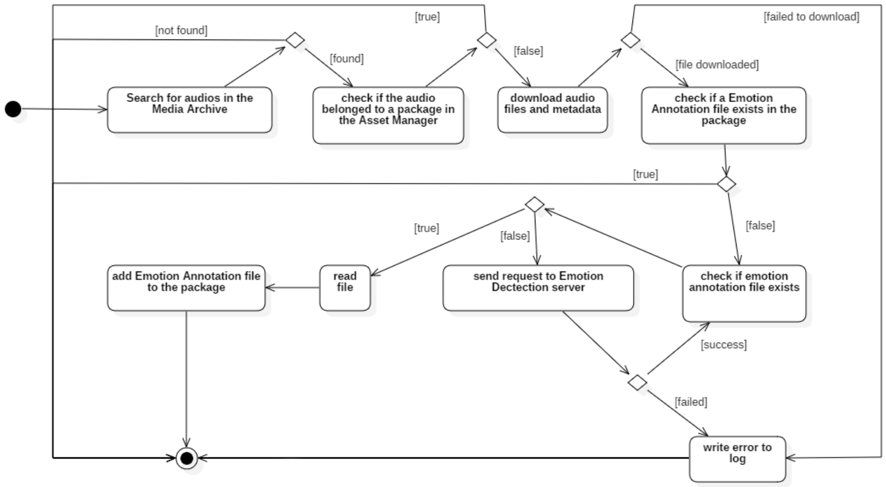
[not found]
[true]
[found]
[false]
[failed to download]
[file downloaded]
Search for audios in the Media Archive
check if the audio belonged to a package in the Asset Manager
download audio files and metadata
check if a Emotion Annotation file exists in the package
[true]
[true]
[false]
[false]
add Emotion Annotation file to the package
read file
send request to Emotion Dectection server
check if emotion annotation file exists
[success]
[failed]
write error to log

Fig. 2 Activity Diagram of the Asset Manager (AM) Cronjob

checks if an emotion annotation was produced for the audio record. If it has been produced, the daemon will go to the next step. Otherwise, it will search for the annotation file produced by the Emotion Detection server. If this file does not exist, the daemon will send a request to the server and let it process the downloaded audio file along with its metadata. An annotation file will then be produced by the server. The daemon reads the file and adds it to the new package. Errors that occur whilst the daemon is running will be written to the log file in order to enable the system administrator to debug them later. Figure 2 describes the activity flow of the cronjob.

After an emotion annotation file is added, users can use the Emotion Audio Player (EAP), which is an important feature of the AM, to play the audio files and discover the current emotional state of the subject in the audio. The emotion in the annotation file will be indexed based on its timestamp. When the audio playback reaches a timestamp, the player will display the emotion associated with it. Furthermore, with this annotation file, emotions of the subject in the audio file can be visualized with various visualization techniques. This enables users to have an overview of the current emotional state of the subject and provides an opportunity to explore hidden information behind human emotion.

Finally, the AM enables users to edit, delete and classify their packages through interacting with a user interface. Not only audio records and respective analysis results inside a package can be classified using the scientific content, emotion annotation, and other types of taxonomies, but the package itself can be classified into different categories using the TM. This classification will be stored in the package as well as indexed in a search server. With the AM, scientific content and respective analysis results can be managed in a central repository. This will reduce dramatically the effort to deploy, maintain, search, and reuse scientific data.

Supporting systems such as Emotion Detection Server and Search Server provide standalone, high-performance services that the MENHIR KM-EP can take advantage of. They provide interfaces, so the KM-EP can send requests and later receive results. In the context of this paper, we focus on the Emotion Detection Server, which is being developed in MENHIR.

The Emotion Detection Server detects human emotion from speech signals extracted from the audio files and their transcriptions. The files will be downloaded from the MA and sent to the processing server by the introduced cronjob from the AM component. The audio samples are processed by the server and their results are exported to annotation files and stored in the local server for the KM-EP to access and use. Automatic recognition of spontaneous emotions from speech is complex [15]. To overcome its challenges, two procedures have been conducted. The first one is the annotation task, that involves the segmentation of the audio samples in order to label them with emotions, and the second one is building a model that is able to distinguish between different emotional states.

In relation to the annotation, transcriptions are used to identify the spoken turns, and those turns have been split automatically into segments of between 2 and 5 seconds, because it is known that there is no emotion change in this time window. Subsequently, each segment is labelled by both professional and crowd annotators following the same questionnaire. The questionnaire includes both categorical and dimensional annotation (valence, arousal, and dominance). Using these annotations, we have experimented with the creation of a model capable of identifying the mood of the speaker through application of neural network algorithms. This model infers the subject's emotional state using both audio features (such as e.g. pitch, energy, Mel-Frequency Cepstral Coefficients (MFCCs)) or the spectrogram. With this model, an emotion detection server can be developed to provide the MENHIR KM-EP with emotion annotations from both acoustic signals and their corresponding transcription in real-time.

A high-performance search server is needed to index content objects in the MA and the DL, so they can be searched quickly by users. Furthermore, indexing classifications enables faceted search, which is a way to add specific, relevant options to the results pages, so that when users search for content, they can see where in the catalogue they've ended up [16]. With the faceted feature, users can have an overview of the classification of contents in real-time and quickly find results by selecting only relevant categories. Furthermore, faceted search enables navigation using taxonomy hierarchies, which are created and managed using the TM [17].

Besides a search server, other systems, such as, e.g. cache server, queuing system, are also important for the MENHIR KM-EP. Caching improves performance of the system by pre-processing and storing frequently used data in memory, each time it is required, it can be retrieved from there without requiring reconstruction resources. A queuing system enables the KM-EP to process data in an organized manner. Processing all data at once requires considerable computing power and resources. Therefore, organizing data into a queue and processing them accordingly would allow the resource to be distributed evenly and reduce stress on system components.

4 Conclusion and future work

The MENHIR project provides rapid intervention, appropriate feedback and overview on the state of development of subject mood and anxiety levels over time, by monitoring moods, behaviour, and symptoms of subjects in real time. The objective of the work reported in this paper is to develop an integration platform to support the ingestion and management of audio files and their metadata, results on human emotion detection from speech, and scientific asset packaging, classification, and management.

Here, we have described the challenges involved in the development and integration of such a platform. The content and knowledge management ecosystem (KM-EP) proposed here is a cloud-based, high-performance, scalable, and easy to use solution. By relying on its Media Archive and Digital Library, the KM-EP is able to ingest, modify, share, and preserve scientific publications and multimedia objects, such as audio files and their metadata. The Taxonomy Manager enables users to classify content and knowledge, which leads to better quality and faster exploration. Finally, the Asset Manager combines related scientific publications, multimedia objects, datasets, and analysis results into packages. With the Asset Manager, all the related data, information, and knowledge can be gathered and managed in one central repository, which is easier to maintain and reuse. The MENHIR KM-EP will provide a useful foundation for the development of conversational systems in mental health promotion and assistance.

The current emotion detection server uses a model, which needs to be trained offline by AI experts. This model also needs to be re-trained frequently with updated corpora to enhanced its accuracy. The MENHIR KM-EP can be extended in the future to use the uploaded audio records in the MA to form a new corpus. Then, the new model can be trained based on the new data corpus and replace the former model automatically. By doing this, the cost of developing an advanced emotion detection model can be reduced.

Acknowledgements This publication has been produced in the context of the MENHIR project. This project has received funding from the European Union's H2020 Programme under grant agreement No. 823907. However, this paper reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

1. M. Healy, R. Donovan, P. Walsh and H. Zheng, "A Machine Learning Emotion Detection Platform to Support Affective Well Being," in IEEE International Conference on Bioinformatics and Biomedicine, 2018.
2. "Sensor Enabled Affective Computing for Enhancing Medical Care," 19 April 2017. [Online]. Available: https://cordis.europa.eu/project/rcn/199563/factsheet/en. [Accessed 27 August 2019].

3. F. Engel, R. Bond, A. Keary, M. Mulvenna, P. Walsh, H. Zheng, H. Wang, U. Kowohl and M. Hemmje, “SenseCare: Towards an Experimental Platform for Home-Based, Visualisation of Emotional States of People with Dementia,” in Advanced Visual Interfaces. Supporting Big Data Applications, 2016.
4. M. Healy and P. Walsh, “Detecting demeanor for healthcare with machine learning,” in IEEE International Conference on Bioinformatics and Biomedicine, 2017.
5. R. Donovan, M. Healy, H. Zheng, F. Engel, B. Vu, M. Fuchs, P. Walsh, M. Hemmje and P. M. Kevitt, “SenseCare: Using Automatic Emotional Analysis to Provide Effective Tools for Supporting Wellbeing,” in IEEE International Conference on Bioinformatics and Biomedicine, 2018.
6. “MENHIR,” [Online]. Available: https://menhir-project.eu/. [Accessed 20 January 2020].
7. “Mental health,” WHO, 2 October 2019. [Online]. Available: https://www.who.int/newsroom/facts-in-pictures/detail/mental-health. [Accessed 20 January 2020].
8. M. Consortium, “MENHIR Proposal,” European Commission, 2018.
9. B. Vu, J. Mertens, K. Gaisbachgrabner, M. Fuchs and M. Hemmje, “Supporting Taxonomy Management and Evolution in a Web-based Knowledge Management System,” in HCI 2018, Belfast, UK, 2018.
10. “Mendeley,” [Online]. Available: https://www.mendeley.com/. [Accessed 27 January 2020].
11. “SlideShare,” [Online]. Available: https://de.slideshare.net. [Accessed 27 January 2020].
12. “Your BibTeX resource,” BibTeX, 2016. [Online]. Available: http://www.bibtex.org/. [Accessed 28 October 2019].
13. “Protocol for Metadata Harvesting,” Open Archives Initiative, [Online]. Available: https://www.openarchives.org/pmh/. [Accessed 28 October 2019].
14. B. Vu, Y. Wu, H. Afli, P. M. Kevitt, P. Walsh, F. Engel, M. Fuchs and M. Hemmje, “A Metagenomic Content and Knowledge Management Ecosystem Platform,” in IEEE International Conference on Bioinformatics and Biomedicine, San Diego, USA, 2019.
15. M. d. Vázquez, R. Justo, A. López Zorrilla and M. Inés Torres, “Can Spontaneous Emotions be Detected from Speech on TV Political Debates?,” in IEEE International Conference On Cognitive Infocommunications, 2019.
16. “What is faceted search and navigation?,” Loop54, 2016. [Online]. Available: https://www.loop54.com/knowledge-base/what-is-faceted-search-navigation. [Accessed April 2019].
17. B. Vu, R. Donovan, M. Healy, P. M. Kevitt, P. Walsh, F. Engel, M. Fuchs and M. Hemmje, “Using an Affective Computing Taxonomy Management System to Support Data Management in Personality Traits,” in IEEE International Conference on Bioinformatics and Biomedicine, San Diego, USA, 2019.