Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets

D1.10 MixedEmotions Data Management Plan, V2

Project ref. no	**H2020 644632**
Project acronym	**MixedEmotions**
Start date of project (dur.)	**01 April 2015 (24 Months)**
Document due Date	**31 May 2015 (Month 14)**
Responsible for deliverable	**Paradigma Tecnológico**
Reply to	**jruiz@paradigmatecnologico.com**
Document status	**Final**

Project reference no.	**H2020 644632**
Project working name	MixedEmotions
Project full name	Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets
Document name	MixedEmotions_D1.10_Data_Management_Plan_V2
Security (distribution level)	PU
Contractual delivery date	30 May 2016
Deliverable number	D1.10
Deliverable name	Data Management Plan, V2
Type	Other
Version	Final
WP / Task responsible	WP3 / Paradigma Tecnológico
Contributors	PT (José Ruiz Cristina), NUIG (Paul Buitelaar, Cécile Robin, Sapna Negi, Ian Wood), UPM (Carlos Ángel Iglesias, Fernando Sánchez), ST (Giovanni Tummarello), PX (Pavel Matejka, Áneta Cerná) BUT (Lubomir Otrusina), UP (Hesham Sagha), DW (Andy Giefer), ES (Vincenzo Masucci)
Project Officer	Martina.EYDNER@ec.europa.eu

Contents

3. 11

1. Introduction and scope

This report, the Data Management Plan (DMP) version 2, describes the data management life cycle for all data sets that have been or will be collected, processed or generated by the MixedEmotions project. It outlines how research data will be handled during the project, and after it is completed, describing what data is collected, processed or generated and what methodology and standards are followed, whether and how this data will be shared and/or made available, and how it will be curated and preserved.

As the DMP is not a fixed document, it evolves and gains more precision and substance during the lifespan of the project, therefore it will be necessarily incomplete. A final Data Management Report will be available by the end of the project.

2. Dataset identification and listing

To allow for more context and a better understanding of the purposes of the different data collecting, the datasets are listed categorized according to the consortium partner that collects the data.

2.1 Paradigma Tecnologico datasets

DW content (text)

Data set reference and name: DW texts and videos
Data set description: Texts and videos obtained from Deutsche Welle API regarding selected brands
Standards and metadata: Text, video, brand, date, language
Data sharing: Restricted availability through DW
Archiving and preservation (including storage and backup): Preserved in a “sources” index in the platform elasticSearch.

Contact: cnavarro@paradigmatecnologico.com

Twitter tweets (text)

Data set reference and name: Twitter tweets
Data set description: Tweets extracted from Twitter regarding selected brands
Standards and metadata: Text, brand, date, language, account.
Data sharing: None. There are legal issues sharing this data.
Archiving and preservation (including storage and backup): Preserved in a "sources" index in the platform elasticSearch.
Contact: cnavarro@paradigmatecnologico.com

Processed Results

Data set reference and name: Processed results
Data set description: Once input data is processed (eg. splitted and emotion, polarity and terms are added) the results are saved to be the base of the analytics.
Standards and metadata: Sentence, brand, date, language, account, original_text, emotions, polarity, concepts, topics, source, media.
Data sharing: No sharing, for commercial reasons.
Archiving and preservation (including storage and backup): Preserved in a "results" index in the platform elasticSearch.
Contact: cnavarro@paradigmatecnologico.com

2.2 NUIG datasets

Review Suggestion Dataset

Data set reference and name: Review Suggestion Dataset
Data set description: Manually labeled sentences from hotel and electronics reviews, which were in turn obtained from existing academic datasets. Each sentence is labeled as 'suggestion' or 'non-suggestion', depending on if the sentence conveys a suggestion. Data labelling is performed using paid crowdsourcing platforms.
Standards and metadata: sentiment polarity, review id, sentence id, tripadvisor hotel id
Data sharing: Publicly available.
Archiving and preservation (including storage and backup): TBD
Link: http://server1.nlp.insight-centre.org/sapnadatasets/EMNLP2015/
Contact: paul.buitelaar@insight-centre.org

Tweet Suggestion Dataset

Data set reference and name: Tweet Suggestion Dataset
Data set description: Manually labeled tweets, downloaded using twitter API. Each tweet is labeled as 'suggestion' or 'non-suggestion', depending on if it conveys a suggestion. Data labelling is performed using paid crowdsourcing platforms. Due to the restrictions imposed by twitter, only tweet id and manual label would be available in the downloadable version of the dataset.
Standards and metadata: tweet id
Data sharing: Publicly available.
Archiving and preservation (including storage and backup): TBD
Link: http://server1.nlp.insight-centre.org/sapnadatasets/starsem2016/tweets/
Contact: paul.buitelaar@insight-centre.org

Forum Suggestion Dataset

Data set reference and name: Forum Suggestion Dataset
Data set description: Manually labeled sentences of posts from a suggestion forum, scraped from the website *www.uservoice.com*. Each sentence is labeled as 'suggestion' or 'non-suggestion', depending on if it conveys a suggestion. Data labelling is performed by the project members.
Standards and metadata: Post id, sentence id, software name.
Data sharing: Publicly available.
Archiving and preservation (including storage and backup): TBD
Link: http://server1.nlp.insight-centre.org/sapnadatasets/starsem2016/SuggForum/
Contact: sapna.negi@insight-centre.org

VAPUI Annotated Tweets (crowd sourced)

Data set reference and name: VAPUI Annotated Tweets
Data set description: Planned data set containing manually labeled tweet comparisons. Tweets will be compared along up to 5 emotional dimensions: Valence (Pleasure / Positivity), Arousal (Activation), Potency (Dominance / Power), Unpredictability (Expectation / Novelty / Surprise) and emotional Intensity. Each annotation is a comparison between two tweets along one of the emotion dimensions. Annotators will be drawn from the CrowdFlower platform. Data on the time taken to perform the annotations will also be also collected. The data is expected to contain 10000 tweet comparisons over 2000 tweets.
Standards and metadata: tweet ids, data collection methodology
Data sharing: Publicly available only for academic research.
Archiving and preservation (including storage and backup): TBD
Contact: paul.buitelaar@insight-centre.org

VAPUI Annotated Tweets (pilot study)

Data set reference and name: VAPUI Annotated Tweets (pilot study data)
Data set description: Manually labeled tweet comparisons. Tweets were compared along each of 5 emotional dimensions: Valence (Pleasure / Positivity), Arousal (Activation), Potency (Dominance / Power), Unpredictability (Expectation / Novelty / Surprise) and emotional Intensity. Annotations were collected for each of two annotation schemes: comparing pairs of tweets and choosing the best/worst tweets from 4. Annotators were drawn from MixedEmotions collaborators and their contacts. Data on the time taken to perform the annotations was also collected. The data contains 30 annotated tweet pairs and 18 annotated tweet quads.
Standards and metadata: tweet ids, data collection methodology
Data sharing: Publicly available only for academic research.
Archiving and preservation (including storage and backup): TBD
Contact: paul.buitelaar@insight-centre.org

Ekman Annotated Emoji Tweets

Data set reference and name: Ekman Annotated Emoji Tweets
Data set description: Tweets containing emotive emoji labelled with Ekman's six basic emotions (Joy, Surprise, Sadness, Anger, Disgust, Fear). Emoji were removed from the tweets before annotation. Annotators were drawn from MixedEmotions collaborators and their contacts. Data on the time taken to perform the annotations was also collected. The data contains 366 annotated tweets.
Standards and metadata: tweet ids, selected emotive emoji, data collection methodology
Data sharing: Publicly available only for academic research.
Archiving and preservation (including storage and backup): TBD
Contact: paul.buitelaar@insight-centre.org

2.3 UPM datasets

Twitter relations

Data set reference and name: Twitter relations
Data set description: Relationships for Twitter accounts. That would be followers and followings of accounts that tweeted about our selected brands.
Standards and metadata: RDF.
Data sharing: No sharing. There are legal issues sharing this data.
Archiving and preservation (including storage and backup): In a graph database that could be Elasticsearch with the Siren plugin.
Contact: jfernando@dit.upm.es

2.4 ExpertSystem datasets

ES Dataset based on the enrichment of DW English Dataset

Data set reference and name: ES Dataset based on the enrichment of DW Dataset
Data set description: All articles published by Deutsche Welle over recent years in English. Metadata describing audio, video and image material published by Deutsche Welle of recent years in all DW languages. This dataset is semantically enriched by ES modules so the final result is a dataset with all the previous information, plus, for each article or A/V, a set of metadata (topic, main lemmas, people, and places)
Standards and metadata: IPTC topic, main lemmas, people, places
Data sharing: The data is available in the platform elasticSearch, access to which was described to the consortium in a separate document. The data is only to be used by consortium members but can be used for scientific publications with DW's permission. The reason is that the rights associated with DW's material vary from item to item, depending on the material's origin.
Archiving and preservation (including storage and backup): The data remains available on the ME Platform elasticSearch after the end of the project.
Contact: vmasucci@expertsystem.com

Twitter trend related to DW A/V

Data set reference and name: Twitter trend related to DW A/V
Data set description: Tweets extracted from Twitter selected through keywords related to DW A/V
Standards and metadata: IPTC topic, main lemmas, people, places, sentiment and emotions
Data sharing: The data is available in the platform elasticSearch, access to which was described to the consortium in a separate document.
Archiving and preservation (including storage and backup): Preserved in an index in the platform elasticSearch.
Contact: vmasucci@expertsystem.com

Twitter trend related to DW English's RSS feed

Data set reference and name: Twitter trend related to DW English's RSS feed
Data set description: Tweets extracted from Twitter selected through keywords related to DW English's RSS feed
Standards and metadata: IPTC topic, main lemmas, people, places, sentiment and emotions
Data sharing: The data is available in the platform elasticSearch, access to which was described to the consortium in a separate document.

Archiving and preservation (including storage and backup): Preserved in an index in the platform elasticSearch.
Contact: vmasucci@expertsystem.com

2.5 Phonexia datasets

CallCenter1

Data set reference and name: CallCenter1
Data set description: Czech telephone speech (PCM 16b linear, 8kHz wav) from a call center in an outbound campaign. Agent and client are recorded in separate channels. Important is the fact that the client's channel is available only. Speech is manually annotated with emotions on a segment level. Arousal and valence value of -1, 0 or 1 were assigned to every speech segment. These labels can be mapped to emotions 'anger', 'joy', 'sadness' or 'neutral'. For more details see the table below. This data is used for training of the emotion recognition system in Pilot 3.
Standards and metadata: call_id, segment_start, segment_end, emotion, arousal, valence
Data sharing: NDA does not allow to share this data or name the call center
Archiving and preservation: Phonexia servers.
Contact: pavel.matejka@phonexia.com

CallCenter2

Data set reference and name: CallCenter2
Data set description: Czech telephone speech (PCM 8b linear, 8kHz wav) from a call center in an outbound campaign. Both agent and client are recorded in a single channel. We manually tagged regions where the operator and client speak. Emotions annotation for client's segments was done in the same way as in the method from Call Center1. For more details see the table below. This data are used for training of the emotion recognition system in Pilot 3.
Standards and metadata: call_id, speaker_id, segment_start, segment_end, emotion, arousal, valence
Data sharing: NDA does not allow us to share this data or name the call center.
Archiving and preservation: Phonexia servers.
Contact: pavel.matejka@phonexia.com

name	duration [h:mm:ss]	arousal			valence		
		-1	0	1	-1	0	1
Call Center1	2:09:16	0:05:42	1:18:42	0:44:53	0:25:49	1:18:42	0:24:45
Call Center2	1:21:41	0:07:10	0:39:33	0:34:58	0:33:13	0:39:33	0:08:55

All	3:30:57	0:12:51	1:58:15	1:19:51	0:59:02	1:58:15	0:33:40

Table 1 *Distribution of arousal and valence values in used Czech Call Center data.*

2.6 DW datasets

DW Article Data and AV Metadata

Data set reference and name: DW Article Data and AV Metadata
Data set description: All articles published by Deutsche Welle over recent years in all DW languages. Metadata describing audio, video and image material published by Deutsche Welle of recent years in all DW languages. This data is mainly used for the recommendation engine and editorial dashboard developed in Pilot 1.
Standards and metadata: JSON format defined by Deutsche Welle.
Data sharing: The data is available via an API, access to which was described to the consortium in a separate document. The data is only to be used by consortium members but can be used for scientific publications with DW's permission. The reason is that the rights associated with DW's material vary from item to item, depending on the material's origin.
Archiving and preservation (including storage and backup): The data remains available through the API after the end of the project.
Contact: andreas.giefer@dw.com

2.7 BUT datasets

Brno Deceit dataset

Data set reference and name: Brno Deceit dataset
Data set description: The dataset will consist of recordings of interview-style sessions in which the interviewees provide true and deceitful statements based on preceding instructions. Part of the dataset is being recorded in a lab with a Kinect V2 RGB-D camera. Larger number of recordings will be recorded via a web application in unconstrained environments and with unconstrained equipment. Upper body video and audio is recorded in both instances. The Kinect V2 provides Full HD video, depth images and audio. The quality of the web application recordings varies due to the equipment used.
Standards and metadata: Truth/deceit labels for individual statements.
Data sharing: The dataset will be publicly available via http download for research purposes.
Archiving and preservation (including storage and backup): The data will remain stored and downloadable from BUT servers after the end of the project.
Contact: smrz@fit.vutbr.cz

2.8 UP datasets

AV+EC dataset

Data set reference and name: AVEC (or AV+EC)
Data set description: The dataset consists of continuous annotation of emotions from 27 participants, each 5 minutes of data recording. The recorded modalities are audio (speech), video, and physiological signals and data is useful for multimodal continuous emotion recognition. The annotations are in terms of arousal and valence. This database is used for the Audio Visual Emotion Challenge (AVEC) in 2015 and 2016. For more information please refer to http://arxiv.org/abs/1605.01600.
Standards and metadata: ARFF
Data sharing: As part of the challenge participants can download the data, however, not the annotations of the test partition.
Archiving and preservation (including storage and backup): Data is stored in a server at the University of Passau and it will stay there for the AVEC challenges of the next years.
Contact person: Fabien Ringeval (Fabien.Ringeval(at)univ-grenoble-alpes.fr)
Challenge URL: http://sspnet.eu/avec2016/
Contact: schuller@tum.de

3. Conclusions

We provided a summary of data sets collected, generated and/or enriched across modalities: DW news text and A/V data, call center audio data, twitter social media data, video data for deceit analysis and multimedia data collected and curated in the context of the AVEC challenge. These data sets will be further curated through automatic enrichment and manual annotation and will be made available publicly where possible and appropriate, as indicated in each section above.