MixedEmotions

Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets

D1.11 Data Management Report

Project ref. no	**H2020 644632**
Project acronym	**MixedEmotions**
Start date of project (dur.)	**01 April 2015 (24 Months)**
Document due Date	**31 March 2017 (Month 24)**
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_30_04_16_MixedEmotions_Data_Management_Plan_final_version_PT
Security (distribution level)	PU
Contractual delivery date	31 March 2017
Deliverable number	D1.11
Deliverable name	MixedEmotions Data Management Report
Type	Other
Version	Final
WP / Task responsible	WP3 / Paradigma Tecnológico
Contributors	PT(José Ruiz Cristina, Carlos Navarro), NUIG(Paul Buitelaar, Cécile Robin, Mihael Arcan, 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

Index

1. Introduction and scope

The Data Management Report is based on the previous Data Management Plans. It describes the data management life cycle for all data sets that have been collected, processed and generated all along MixedEmotions' project. It outlines how research data is handled during the project and what it will become after its completion. The document is describing what data has been collected, processed or generated, what methodology and standards have been followed, whether and how this data will be shared and/or made open, and how it will be curated and preserved.

2. Dataset identification and listing

The following datasets have been listed according to the consortium partner that collected the data.

2.1. Paradigma Tecnologico datasets

Twitter tweets (text)

Data set reference and name: Twitter tweets
Data set description: Tweets extracted from Twitter regarding selected brands. They have been collected using Twitter API to serve as an input to extract emotions and other interesting information about targeted brands in Pilot 2.
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): Not to be preserved for technical and legal reasons. Saved in a temporary folder.
Contact: cnavarro@paradigmatecnologico.com

Youtube videos (video)

Data set reference and name: Youtube videos
Data set description: Videos regarding selected brands, extracted from Youtube using Youtube API in order to serve as an input to extract emotions and other interesting information about targeted brands in Pilot 2.
Standards and metadata: Brand, date, language, account, video.
Data sharing: None. There are legal issues sharing this data.

Archiving and preservation (including storage and backup): Not to be preserved for technical reasons. Saved in a temporal folder.
Contact: cnavarro@paradigmatecnologico.com

Youtube metadata (text)

Data set reference and name: Youtube metadata
Data set description: Metadata from videos regarding selected brands, extracted from Youtube using Youtube API in order to serve as an input to extract emotions and other interesting information about targeted brands in Pilot 2.
Standards and metadata: Brand, date, language, account, video.
Data sharing: None. There are legal issues sharing this data.
Archiving and preservation (including storage and backup): Not to be preserved for technical and legal reasons. Saved in a temporary folder.
Contact: cnavarro@paradigmatecnologico.com

Processed Results

Data set reference and name: Processed results
Data set description: Data extracted using MixedEmotions platform stored for the final user so that he can access it via the visualization tools available in Pilot 2. Once input data is processed (eg. splitted and where 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. The reviews are obtained from existing academic datasets for Sentiment Analysis. Data labelling is performed using paid crowdsourcing platforms. The annotators were shown a sentences and were asked to choose one out of two labels which were *suggestion* and *non-suggestion*. A definition of suggestion was provided. Annotators were first tested on a set of 10 already labeled sentences, only those annotators were employed who provided 7 or more correct annotations. The data was used to train and evaluate machine learning based classifiers for the Suggestion Mining module.
Standards and metadata: The files are in csv format, with the following fields: sentence id, sentence text, sentiment polarity, and suggestion label.
Data sharing: Publicly available.

Archiving and preservation (including storage and backup): The dataset is available upon request to the authors of the relevant publications, which in this case is:
Towards the Extraction of Customer-to-Customer Suggestions from Reviews. Sapna Negi and Paul Buitelaar. Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)
URL: 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 as 'suggestion' or 'non-suggestion' downloaded using twitter API. Data labelling is performed using paid crowdsourcing platforms. The annotators were shown a tweet and were asked to choose one out of two labels which were *suggestion* and *non-suggestion*. A definition of suggestion was provided. Annotators were first tested on a set of 10 already labeled tweets, only those annotators were employed who provided 7 or more correct annotations. Due to the restrictions imposed by twitter, only tweet id and manual label would be available in the downloadable version of the dataset. The data was used to train and evaluate machine learning based classifiers for the Suggestion Mining module.
Standards and metadata: The files are in csv format, with the following fields: tweet id, and suggestion label.
Data sharing: Publicly available only for academic research.
Archiving and preservation (including storage and backup):
The dataset is available upon request to the author's of the relevant publications, which in this case is:
A Study of Suggestions in Opinionated Text and their Automatic Detection. Sapna Negi, Kartik Asooja, Shubham Mehrotra, Paul Buitelaar. *Sem 2016, Co-located with ACL 2016
Contact: sapna.negi@insight-centre.org
Url: http://server1.nlp.insight-centre.org/sapnadatasets/starsem2016/tweets/

Forum Suggestion Dataset

Data set reference and name: Suggestion Forum Dataset
Data set description: Manually labeled sentences of posts from a suggestion forum. Each sentence is labeled as 'suggestion' or 'non-suggestion', depending on if it conveys a suggestion. The posts are scraped from the website *www.uservoice.com*. Posts are automatically split into sentences using NLTK. Each sentence is labeled as 'suggestion' or 'non-suggestion', depending on if it conveys a suggestion. Data labelling is performed by the project members, and the dataset was used to train and evaluate machine learning based classifiers for the Suggestion Mining module.
Standards and metadata: Post id, sentence id, software name.
Data sharing: Publicly available only for academic research.
Archiving and preservation (including storage and backup): The dataset is available upon request to the authors of the relevant publications, which in this case is:

A Study of Suggestions in Opinionated Text and their Automatic Detection. Sapna Negi, Kartik Asooja, Shubham Mehrotra, Paul Buitelaar. *Sem 2016, Co-located with ACL 2016
URL: http://server1.nlp.insight-centre.org/sapnadatasets/starsem2016/SuggForum/
Contact: paul.buitelaar@insight-centre.org

VAPU Annotated Tweets (crowd sourced)

Data set reference and name: VAPU Annotated Tweets
Data set description: Data set containing manually labeled tweets and tweet comparisons. Tweets were annotated along 4 emotional dimensions: Valence (Pleasure / Positivity), Arousal (Activation), Potency (Dominance / Power) and Unpredictability (Expectation / Novelty / Surprise).
Two annotation schemes were used: comparison between two tweets along one of the emotion dimensions and ranking single tweets on a 5-point scale along one of the emotion dimensions. Annotators were native english speakers drawn from the CrowdFlower platform. Each emotion dimension + annotation scheme was performed as a separate CrowdFlower task. Data on the time taken to perform the annotations is also included. The data contains comparisons of 2019 tweet pairs with 4-6 annotations for each pair and 2019 individual tweets with 3-4 annotations for each tweet. The tweet pairs are drawn from the same 2019 tweets as individual annotations, with each of the tweets present in at least one tweet pair.
The data was collected to be used as evaluation and/or training data for emotion detection models from social media text with a dimensional emotion representation scheme.
Standards and metadata: tweet ids, data collection methodology
Data sharing: Publicly available.
Archiving and preservation (including storage and backup): Available on request to ian.wood@insight-centre.org or other member of the Unit for Natural Language Processing at the Insight Centre in NUIG under the Creative Commons Attribution 3.0 Unported (CC BY 3.0) licence.

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.
It was collected to be used as evaluation and/or training data for emotion detection models from social media text with a dimensional emotion representation scheme.
Standards and metadata: tweet ids, data collection methodology
Data sharing: Publicly available.

Archiving and preservation (including storage and backup): Available on request to ian.wood@insight-centre.org or other member of the Unit for Natural Language Processing at the Insight Centre in NUIG under the Creative Commons Attribution 4.0 (CC BY 4.0) licence.

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). The data contains 366 annotated tweets. 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 was used to evaluate the role of emoji in emotion expression in Twitter.
Standards and metadata: tweet ids, selected emotive emoji, data collection methodology
Data sharing: Publicly available.
Archiving and preservation (including storage and backup): Available on request to ian.wood@insight-centre.org or other member of the Unit for Natural Language Processing at the Insight Centre in NUIG under the Creative Commons Attribution 4.0 (CC BY 4.0) licence.

Polylingual WordNet

Data set reference and name: Polylingual WordNet
Data set description: Polylingual WordNet is an extension of Princeton WordNet, providing it in 23 languages by automatic translation. The data has been created through automatic translation utilising multilingual text corpora for word sense alignment.
The aim was to provide the WordNet lexical resource in multiple languages, and to showcase the methodology for the automatic translation of lexical resources.
Standards and metadata: Released as both OntoLex JSON-LD as well as in the Global WordNet LMF Format.
Data sharing: Publicly available under http://polylingwn.linguistic-lod.org/
Archiving and preservation (including storage and backup): This resource is available for re-use under the Creative Commons Attribution 4.0 License.
Expanding wordnets to new languages with multilingual sense disambiguation. Mihael Arcan, John P. McCrae and Paul Buitelaar, Proceedings of The 26th International Conference on Computational Linguistics, (2016).
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. The dataset is a collection of tweets crawled with Twitter API, from followers and followings of accounts that tweeted about our selected brands. The goal was to showcase relationships between users, and the propagation of emotions relevant to targeted keywords. It is used to test the Scaner module and for Pilot 2.
Standards and metadata: RDF.
Data sharing: No sharing. There are legal issues sharing this data.
Archiving and preservation (including storage and backup): Both in Elasticsearch and OrientDB.
Contact: jfernando@dit.upm.es

2.4 ExpertSystem datasets

ES Dataset based on the enrichment of DW English Dataset

Data set reference and name: DW data enrichment.
Data set description: All articles published by Deutsche Welle over recent years in English, fetched from DW's repository. 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 data (topic, main lemmas, people, places, organizations). It was used to create the tagged knowledge base related to DW's data.
Standards and metadata: IPTC topic, main lemmas, people, places, organizations
Data sharing: The data is stored in and available through Elasticsearch. 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 will be available in Elasticsearch after the end of the project.

Twitter trend related to DW's A/V

Data set reference and name: Twitter trend on DW's A/V
Data set description: Tweets selected through keywords related to DW A/V descriptions, extracted from Twitter with the Twitter API. The dataset was used to link DW's A/V to the sentiments and emotions of the related trend tweets.
Standards and metadata: Sentiment and emotions tags
Data sharing: The data is stored in and available through Elasticsearch.
Archiving and preservation (including storage and backup): Preserved in an index of Elasticsearch.

Contact: vmasucci@expertsystem.com

Twitter trend related to DW's English RSS feed

Data set reference and name: Twitter trend on DW RSS feed
Data set description: Tweets selected through keywords related to DW's English RSS feed, and extracted from Twitter using the Twitter API. It was used to link DW's English RSS feed to the sentiments and emotions of the related trend tweets
Standards and metadata: Sentiment and emotions
Data sharing: The data is stored in and available through Elasticsearch.
Archiving and preservation (including storage and backup): Preserved in an index of 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) collected from a call center in an outbound campaign. Agent and client are recorded in separate channels. It is important to note that only the client's channel is available. The 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 the training and evaluation 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 a 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) collected 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. The emotions annotation for the 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 is used for training 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

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

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

CallCenter3

Data set reference and name: CallCenter3
Data set description: English telephone speech (PCM 16b linear, 8kHz wav) collected from the feedback line of a rent-to-own company. Clients rate the customer experience in a few sentences. Manual speech transcription is annotated with sentiment on a sentence level. One of the 5 sentiment labels (very negative, negative, neutral, positive or very positive) is assigned to each sentence. These labels can be mapped to sentiment values in interval <-1, 1>. For more details see the table below. This data is used for evaluation and comparison of different sentiment recognition systems in Pilot 3.
Standards and metadata: call_id, sentence_start, sentence_end, sentiment, transcription
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

CallCenter4

Data set reference and name: CallCenter4
Data set description: English telephone speech (PCM 16b linear, 8kHz wav) collected from the helpdesk of an accommodation company. Clients and agent are recorded in separate channels. Calls are manually transcribed and the client's channel is annotated in the same way as in the method from CallCenter3. For more details see the table below. This data is used for evaluation and comparison of different sentiment recognition systems in Pilot 3.
Standards and metadata: call_id, channel_id, sentence_start, sentence_end, sentiment, transcription
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

CallCenter5

Data set reference and name: CallCenter5
Data set description: English telephone speech (PCM 16b linear, 8kHz wav) collected from the feedback line of tax-preparation company. Clients rated the customer experience in a few sentences. Calls are manually transcribed and annotated the same way as in the method from CallCenter3. For more details see the table below. This data are used for evaluation and comparison of different sentiment recognition systems in Pilot 3.

Standards and metadata: call_id, sentence_start, sentence_end, sentiment, transcription
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

Distribution of sentiment classes in English call center data.

name	duration [h:mm:ss]	sentiment		
		negative	neutral	positive
Call Center3	2:29:41	1:00:07	0:34:04	0:55:31
Call Center4	4:59:26	0:31:31	4:04:07	0:23:48
Call Center4	1:16:42	0:23:18	0:12:07	0:41:17
All	8:45:50	1:54:55	4:50:18	2:00:36

2.6 DW datasets

Comment [1]: +andreas.giefer@dw.com _Assigned to andreas.giefer_

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 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 you can download data, however, not the annotations of the test partition.
Archiving and preservation (including storage and backup): Data is stored in a server in 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/

3 Conclusion

In this document the datasets collected during the project are categorized and characterized. Each partner has detailed the characteristics of their datasets, including their purpose, method of collection and distribution option.