D4.5 FINAL VERSION OF THE TEXT ANALYSIS COMPONENT

Grant Agreement nr	611057
Project acronym	EUMSSI
Start date of project (dur.)	December 1st 2013 (36 months)
Document due Date :	July 31st 2016 (+60 days) (32 months)
Actual date of delivery	August 30th
Leader	GFAI
Reply to	susannep@iai.uni-sb.de
Document status	Submitted

Project co-funded by ICT-7th Framework Programme from the European Commission

Project ref. no.	611057
Project acronym	EUMSSI
Project full title	Event Understanding through Multimodal Social Stream Interpretation
Document name	
Security (distribution level)	PU – Public
Contractual date of delivery	July 31st 2016 (+60 days)
Actual date of delivery	August 30th
Deliverable name	Final version of the text analysis component, including: NER, event detection and sentiment analysis
Type	P – Prototype
Status	Submitted
Version number	1
Number of pages	5
WP /Task responsible	GFAI / GFAI & UPF
Author(s)	Susanne Preuss (GFAI), Maite Melero (UPF), Jens Grivolla (UPF)
Other contributors	
EC Project Officer	Mrs. Alina Lupu Alina.LUPU@ec.europa.eu
Abstract	The deliverable should report on the finishing of the text analysis components, modifications and adaptations based on demonstrator needs and user feedback, and the integration into the platform, but its content has been merged with D4.7.
Keywords	Text analysis component, Named Entity Recognition, Named Entity Linking, Keyphrase Extraction, QuoteFinder, Normalization of ASR-ed text, Sentiment Analysis
Circulated to partners	N/A
Peer review completed	N/A
Peer-reviewed by	N/A
Coordinator approval	Yes

Table of Contents

1. BACKGROUND

Deliverable D4.5 is described as follows, in the DoW:

Final version of the text analysis component, including: NER, event detection and sentiment analysis: Final version of the text analysis component for integration in the multimodal platform, including: (1) a NER system that is capable of detecting names of all sorts in all text sorts and all languages the project deals with, (2) a component detecting / extracting events, relations and topics, for four languages and (3) a sentiment analysis module, for four languages.

D4.5 is the final prototype of the text analysis component integrated in the multimodal EUMSSI platform. This prototype is the result of the work performed within tasks 4.1, 4.2, 4.3 and 4.6 during the first 32 months.

According to the DoW, Task 4.1 aims at *implementing and testing a NER system that is capable of detecting names in free text in English, French, Spanish and German.* The goal of task 4.2 is the *implementation of a processing component capable of detecting and extracting events, through the detection and extraction of topic, key phrases and relations.* Task 4.3 aims at *implementing and testing a sentiment analysis module*. And Task 4.6 *takes care of the integration of the SM and text analysis modules with the crossmodal platform, including fine-tuning of the modules for the implementation of the use-case demonstrators.*

Within the EUMSSI project, the output of the text analysis component (WP4) feeds into the cross-modal semantic representation framework (WP5; LUH, VSN, UPF) which serves as the basis for the contextualization and recommendation tools (WP6;UPF) whose specifications are guided by the user specifications developed in WP2.1 (DW, LUH, VSN). The text analysis component takes as input news-related data such as news articles provided by Deutsche Welle (DW), or news articles crawled from the web by LUH (WP5). Other text types such as the output of OCR (Optical Character Recognition) and ASR (Automatic Speech Recognition) are provided by IDIAP and LIUM respectively (WP3). Processing pipelines and annotation data formats are determined in consultation with all partners (WP2.2 System architecture, WP2.3 Data infrastructure and representation definition).

D4.5 is preceded by D4.1 -*Preliminary version of the text analysis component*- delivered in month 12, and D4.3 -*First functional version of the text analysis component*- delivered in month 24.

During the third year of the project, WP4 efforts have focused on those functionalities deemed more relevant by the user evaluation of the two demonstrators, particularly by Deutsche Welle journalists providing feedback on the contextualization tool, as well as on the full integration of the text analysis component with the EUMSSI platform. In summary, most of the work on the text analysis component has concentrated on the following points:

1. Implementation of on-demand text analysis for the Contextualising Tool.
2. Finalizing the QuoteFinder.
3. Tailoring the keyphrase extraction to the needs of the demonstrators and user feedback.
4. Make more efficient the integration of sentence segmentation and punctuation of ASR transcripts.
5. Finalizing name normalization and name enrichment
6. Extend sentiment analysis to quotes.
7. Integration of the text analysis pipelines into the EUMSSI platform
8. Evaluation of all the text analysis components

D4.5 is followed by **D4.7 in month 34, which is a deliverable of type Report** that contains a description and evaluation report of the social and text analysis components.

D4.5 being of type Prototype, and being followed just two months later by D4.7 of type Report, we have decided to include all the descriptive part in the latter for clarity purposes and to avoid unnecessary repetitions. Therefore, a description of all the points enumerated above together with the results of the evaluation of the respective parts can be found in D4.7. The reader is referred to D4.7 for the final documentation of the social and text analysis components of the EUMSSI platform.