MULTISENSOR

Mining and Understanding of multilinguaL contenT for Intelligent Sentiment Enriched coNtext and Social Oriented inteRpretation

FP7-610411

D4.1

Multimedia indexing and topic-based classification

Dissemination level:	Public
Contractual date of delivery:	Month 12, 31/10/2014
Actual date of delivery:	Month 12, 31/10/2014
Workpackage:	WP4 Multidimensional content integration and retrieval
Task:	T4.1 Topic-based modelling T4.4 Multimodal indexing and retrieval
Type:	Report
Approval Status:	Final Draft
Version:	1.0
Number of pages:	62
Filename:	D4.1_MultimediaIndexing_2014-10-31_v1.0.pdf

Abstract

The document describes the techniques for topic-based classification using supervised machine learning on top of a vector-based content representation. It also includes the presentation of a unified model for the representation of socially enriched multimedia.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

Co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	11/09/2014	Document structure	A. Moumtzidou, S. Vrochidis (CERTH)
0.2	18/10/2014	Contributions	D. Liparas, T. Tsikrika, A. Moumtzidou (CERTH)
0.3	24/10/2014	Integrated document	A. Moumtzidou, (CERTH)
0.4	24/10/2014	Review of CERTH contribution	I. Kompatsiaris (CERTH)
0.5	24/10/2014	Internal Review of the whole document	I. Arapakis (BM-Y!)
1.0	31/10/2014	Final version	A. Moumtzidou (CERTH)

Author list

Organization	Name	Contact Information
CERTH	Anastasia Moumtzidou	moumtzid@iti.gr
CERTH	Dimitris Liparas	dliparas@iti.gr
CERTH	Stefanos Vrochidis	stefanos@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr
CERTH	Theodora Tsikrika	theodora.tsikrika@iti.gr

Executive Summary

This Deliverable presents the first version of the topic-based modeling and the multimodal indexing and retrieval modules.

Specifically, D4.1 presents the general architecture of MULTISENSOR and show how the WP4 modules fit in it, while also showing which user requirements are covered by them. Initially, we present the development of a multimedia data representation framework that will support multimedia indexing and retrieval. The requirements, characteristics of web sites, social media hosting and sharing platforms are used for defining the features that the proposed model should satisfy. Then the proposed model, named SIMMO, is presented extensively, and a comparison with existing approaches is performed. In addition, the Deliverable also presents the study and experiments conducted on category-based classification that involves the classification of News Items into a predefined set of generic topics (i.e. politics, finance, and lifestyle). Initially, a study with the existing approaches is presented, which is followed by the presentation of the proposed approach that uses Random Forests (RF). Finally, a set of experiments is conducted that use several modalities including textual and visual information. The results of these experiments are presented and conclusions are drawn from them.

Abbreviations and Acronyms

AFP	Agence France-Presse
ASR	Automatic Speech Recognition
AP	Associated Press
C	Characteristics
CEP	Content Extraction pipeline
CHMM	Conditioned Hidden Markov Model
DBM	Deep Boltzmann Machine
DC	Document Classification
DPA	Deutsche Presse-Agentur
FOAF	Friend Of A Friend
IPTC	International Press Telecommunications Council
JSON	JavaScript Object Notation
KNN	K Nearest Neighbour
OCR	Optical Character Recognition
OOB	Out-Of-Bag
PA	Press Association
R	Requirements
RF	Random Forests
RBMs	Restricted Boltzmann Machines
SIFT	Scale-invariant feature transform
SIMMO	Socially Interconnected MultiMedia-enriched Objects
SIOC	Socially-Interlinked Online Communities
SVM	Support Vector Machine
T	Tasks
UML	Unified Modeling Language

Table of Contents

1 INTRODUCTION

This Deliverable targets two of the research challenges of MULTISENSOR which are: a) the proposal and implementation of a multimodal indexing structure that effectively captures enriched multimedia content, and b) the topic-based classification that involves the categorization of the news items into predefined generic categories such as sports, finance, and lifestyle.

To achieve the first goal, MULTISENSOR will develop a multimedia data representation framework that will support multimedia indexing and retrieval. The model will need to capture a number of characteristics inherent in the online social multimedia, and to support tasks commonly performed in multimedia information processing such as search and clustering. Finally, the proposed model will be compared with existing approaches in order to draw conclusions on the efficiency of the model and the capturing of the aforementioned characteristics and tasks.

Regarding the second goal, MULTISENSOR will develop a topic-based classification technique that will classify the News Items stored in the News Repository into generic categories. The multimedia data and their characteristics described earlier will be used as input for this module.

The tasks that target the aforementioned goals are Tasks 4.1 and 4.4 and they report the techniques for topic-based classification using supervised machine learning on top of a content representation model, and the model itself respectively. The timeline of the tasks along the project's lifetime is the following:

ACTIVITY	Y1	Y2	Y3
WP4	D4.1		D4.4
A.4.1 Topic-based		M3 – M33	
Category-based	M3-M20		
Topic-event detection			M18-M33
A.4.4Multimodal		M1 – M33	
Model development	M1-M16		
Indexing Structure		M14-M33	

The document is structured as follows: In Section 2 a short overview of MULTISENSOR's architecture is realized with a focus on how the aforementioned modules fit in it. In Section 3 the multimodal indexing and retrieval model is presented. Section 4 contains the topic-based modelling techniques used and finally Section 5 concludes the Deliverable.

2 ARCHITECTURE

This section provides a short overview of the general architecture of MULTISENSOR and how the modules, which are under discussion in the current Deliverable, fit in it.

2.1 Description

According to Deliverable 7.2 ("Technical requirements and architecture design"), the main process applied to all News Items harvested by MULTISENSOR platform is the Content Extraction Pipeline (CEP) (Figure 1). This pipeline includes all the data processing modules involved in MULTISENSOR such as language analysis, syntactical analysis, sentiment analysis, and video and image analysis of the contents. The aforementioned processing modules are applied to each one of the News Items stored in the News Repository, and the results retrieved, along with the original content, are stored in an indexing structure that handles them effectively. Next, the topic-based classification module is applied that attaches a topic, selected from a predefined list of topics, to each News Item.

Finally, the data stored inside this structure can be used in several types of retrieval techniques, such as similarity search based on text or image, and clustering, and the retrieval can be performed either on the full dataset stored in the indexing structure or on part of it.

Figure 1 depicts this pipeline, while the modules that are discussed in the current Deliverable are highlighted.

2.2 WP4 – related modules

The modules of the architecture, as depicted in Figure 1, that are related to the current Deliverable are the Indexing and the Classification modules.

The "Indexing module" deals with the development of a structure that holds the multimodal information produced during the processing of the data found in the News Repository and it can be broken down to the following subcomponents: a) Model development that involves the specification of a representation model for holding several dimensions (i.e. textual, visual, contextual, sentiment, location and time) of the multimedia information; the model that will be developed draws upon several existing models and combines them in order to achieve a more complete description of an object, and b) Indexing structure for holding and retrieving efficiently the multimodal entities of multimedia information will be developed. Each modality will be treated differently during the indexing and retrieval procedure.

As far as the "Classification module" is concerned, it deals with the classification of the content stored at the News Repository into categories, by using the multimodal features (i.e. textual and visual concepts, events, contextual and sentiment information) and stored in the system. The categories used are identified as generic and they are retrieved from widely used taxonomies, and they are also reviewed by expert end-users. The classification step involves the construction of a training set, the training of one or more models depending on the technique used, and finally the testing of the models using real data coming from the News Repository. Finally, according to Deliverable 8.2 ("User requirements, specification of pilot use cases and validation plan"), which describes the user requirements for MULTISENSOR use cases, the "Classification module" was considered of interest only for the following two use cases:

- Journalism use case scenario
- Commercial media monitoring use case scenario

The users in the aforementioned scenarios will be aided by a high-level labeling of the News Items to some general categories that will allow them to easier browse through the big selection of articles/tweets etc. that are stored inside the database.

More information on the input, output, the programming languages or tools used, the dependencies and the critical factors of the aforementioned modules can be found in Deliverable 7.1 ("Roadmap towards the implementation of MULTISENSOR platform").

Finally, Figure 2 depicts in more detail the type of data that will be stored inside the indexing structure (i.e. original data, and textual, contextual, visual information produced after applying processing techniques) and also, the fact that the "Classification module" uses all these channels for adding labels to the News Item objects. It should be noted that the result produced during the "Classification module" is also stored in the indexing structure.

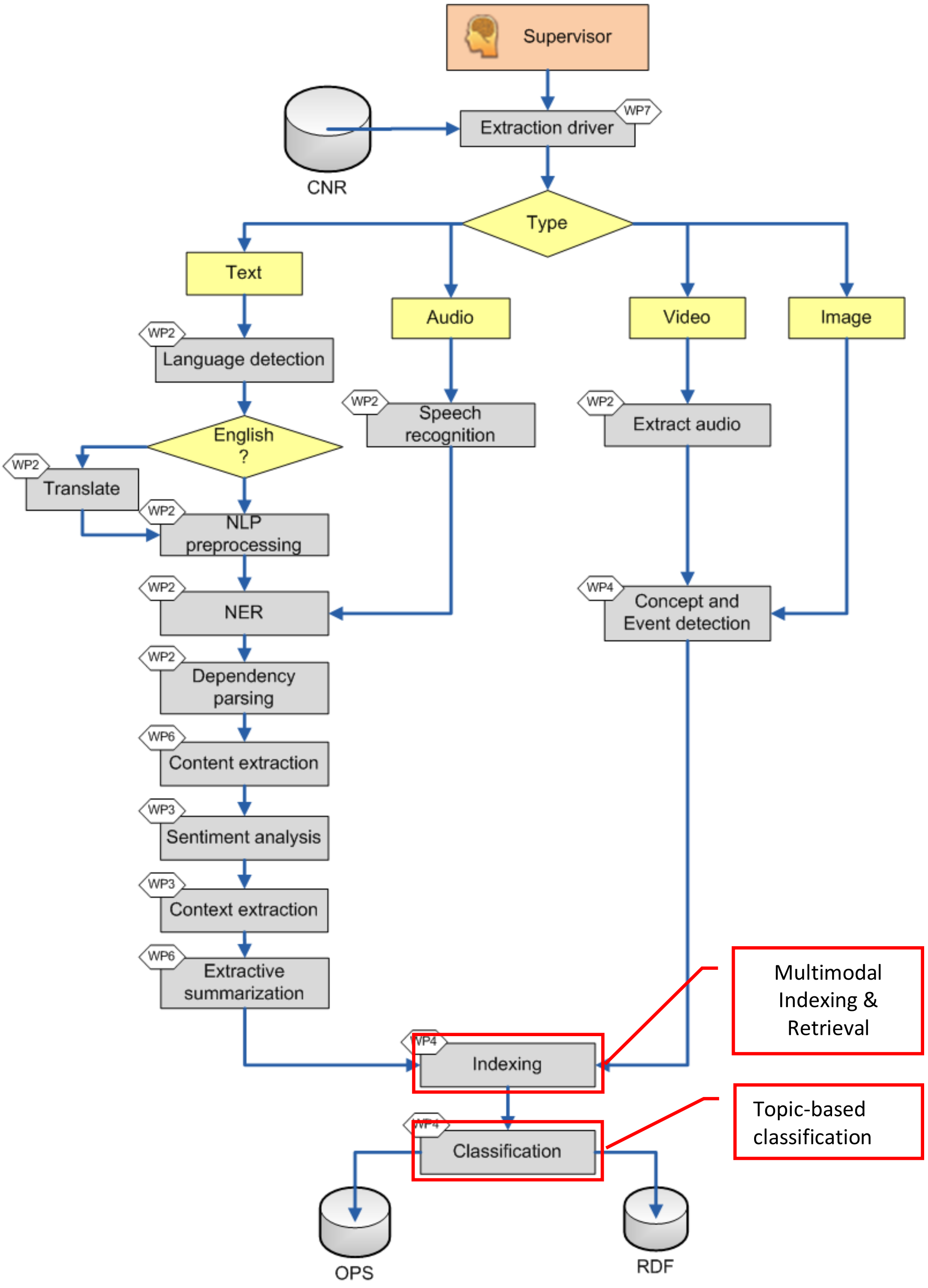


Figure 1: Content Extraction pipeline with the modules discussed in this Deliverable highlighted.

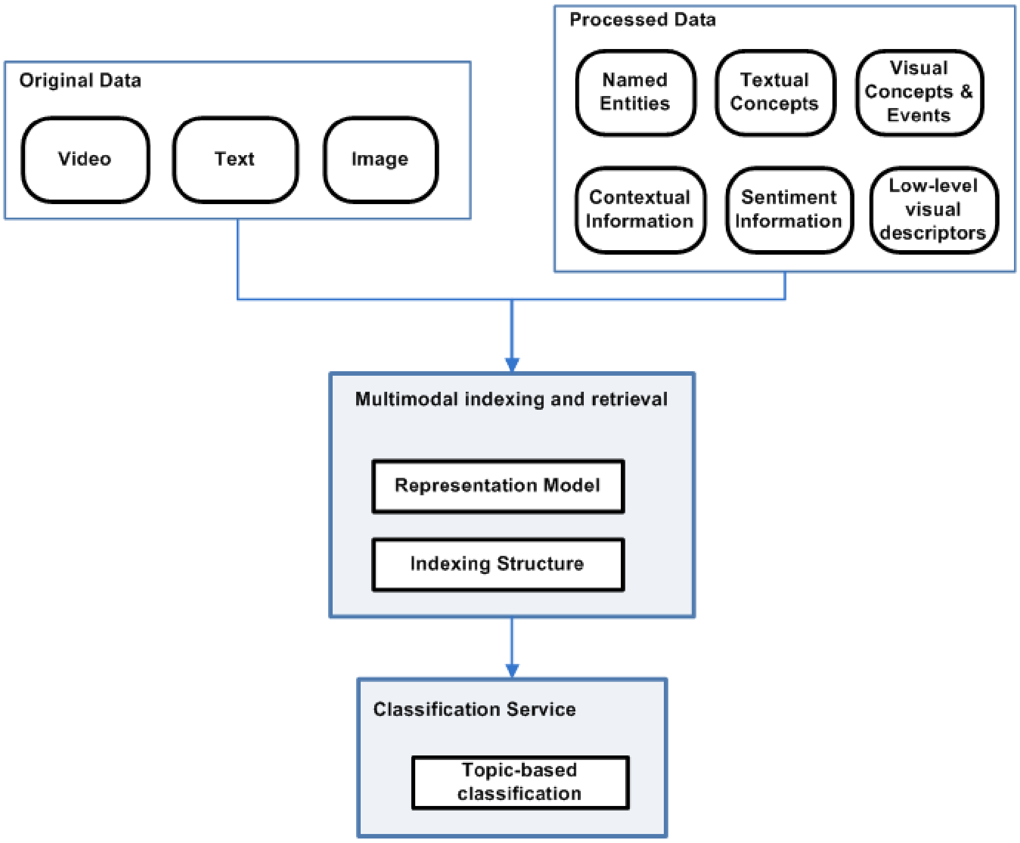


Figure 2: Multimodal indexing/ retrieval and classification services

3 MULTIMODAL INDEXING AND RETRIEVAL

The multimodal indexing and retrieval module involves the development of a multimedia data representation framework that allows for the efficient storage and retrieval of socially connected multimedia objects. The module involves the following two activities:

1. Model development: In this activity, a representation for holding several dimensions (i.e. textual, visual, contextual, sentiment, location and time) of the multimedia information is specified.
2. Indexing structure: An indexing structure for holding and retrieving efficiently the multimodal entities of multimedia information will be developed.

In this Deliverable, we focus only on the first activity that is the development of a data representation model that captures a broad range of the characteristics of web sites, social media hosting and sharing platforms and ensures interoperability across diverse multimedia objects, hosting services, and tasks. Moreover, in order to satisfy the information needs of the diverse users in such online social multimedia environments, it is necessary to develop effective multimedia information processing, analysis, and access applications that support tasks such as clustering, classification, summarisation, search, recommendation, and retrieval.

In general, a considerable number of models have been proposed for handling the aforementioned needs. However, their focus is usually limited on part of the characteristics, e.g. multimedia content or social characteristics, and thus they are unable to fully capture all the content information. For example, MPEG-7 (Chang et al., 2001), RUCoD (Daras et al., 2011), and WebLab (Giroux et al., 2008) models capture very well the description and structure of multimedia content but they do not consider the social characteristics and interconnections in current web settings, since most were developed prior to the advent of social Web or/and for different purposes. On the other hand, more recent approaches, like the SIOC (Bojars et al., 2008) and FOAF (Brickley and Miller, 2012) ontologies, capture such social aspects but they do not account for the potential multimodality of online content and the variety of its extracted annotations.

The model that is proposed within the context of MULTISENSOR aims at bridging this gap by fully covering the characteristics of interconnected multimedia objects and at the same time avoid the complexity of previous models (e.g., (Chang et al., 2001)). The proposed model is called SIMMO, which stands for Socially Interconnected MultiMedia-enriched Objects. SIMMO definition has been recently accepted for publication in the 21st International Conference on MultiMedia Modelling (MMM2015) (Tsikrika et al., 2015).

3.1 Data representation requirements

Before proceeding with the presentation of the proposed model, it is necessary to analyse the characteristics stemming from the nature of online multimedia taking into account their social dimension, with particular focus on those that are typically considered in the context of multimedia information processing, analysis, and access applications. These characteristics should guide the structure and elements definitions of the proposed model. The most salient such characteristics according to (Papadopoulos and Kompatsiaris, 2014; Ramzan et al., 2013) are the following:

C1 **Host heterogeneity and fragmentation:** The number of online services hosting and sharing media content, ranging from Web sites hosted on dedicated servers to social media sharing platforms, such as Flickr, Instagram, YouTube, is rapidly growing. These services differ significantly in several aspects such as the attributes their structure and their attributes. It is possible though to identify cross-platform/services mappings for several of these attributes, such as among those conveying endorsement, e.g., likes in Facebook and favourites in Twitter.

C2 **Media objects diversity:** Online media content is expressed in a variety of modalities (such as text, images, video, and audio) and contained within diverse media objects, ranging from simple media items (e.g., an online image or video file) to complex multimedia documents (e.g., Web pages and social media posts) consisting of heterogenous media items.

C3 **Online links and relations:** Media objects are usually connected to each other with various relations. As mentioned above, multimedia documents can contain media items (e.g., YouTube videos can be embedded in Web pages or be shared through tweets), while they can also be interconnected with other media objects (e.g., Web pages or social media posts can contain links to other Web pages).

C4 **Social links and interactions:** The users of social media sharing and networking platforms are connected with each other through explicit links (e.g., followship, friendship) and interact with the posted content and with each other (often using content), e.g., they like Facebook posts, comment on YouTube videos, add Vines to replies in Twitter, etc. Such social user behaviour is also supported outside the context of such platforms by several Web sites that allow social interactions with their Web pages through, e.g., posting comments on them or sharing them on Twitter or Facebook.

C5 **Dynamic content:** Multimedia documents can also be classified based on their relationship with time. Static multimedia documents do not have a temporal dimension, whereas dynamic Web pages change over time, e.g., through comments being continuously posted on them.

C6 **Automatically generated metadata:** The digital devices currently used for generating media items (e.g., images) have the capability to automatically create a considerable number of metadata for annotating them, such as the geographical identification metadata. Such automatically generated metadata typically accompany the raw content, but social media sharing platforms may replace them with explicit metadata fields or even completely remove them in some cases.

Apart from the characteristics related to the nature of online social multimedia, there is the need of supporting a number of tasks that are commonly performed in multimedia information processing, analysis, and access applications (Papadopoulos and Kompatsiaris, 2014; Ramzan et al. 2013), such as search, clustering, and summarisation. Such tasks are also based on the user requirements of MULTISENSOR project (Deliverable 8.2) and other relevant European projects such as SocialSensor[1] and REVEAL[2]. Typical such tasks in on-line

[1] http://socialsensor.eu/

[2] http://revealproject.eu/

social multimedia settings include those listed below; their characteristics are influenced, to a large extent, by the properties of such settings outlined above:

T1 **Cross-host search:** End users are interested in retrieving media content in response to their information needs irrespective of the environment hosting the relevant media objects (see also C1), e.g., both Web pages and tweets relevant to a submitted query. Establishing the relevance and importance of media objects hosted in multiple and widely different environments is particularly challenging given their heterogeneity.

T2 **Multimodal search:** End users are interested in retrieving relevant information irrespective of the media in which it is encoded, while also having the freedom to express their queries in whichever media they find intuitive, e.g., using similar images when searching for an image or keywords to find their favourite song, and combinations thereof. Enabling unified retrieval that is transparent to users given queries expressed in any number of modalities is a difficult task given also the heterogeneity of available media objects (see also C2) and annotations (as discussed next in T3).

T3 **Layered annotation:** Multimedia content can be currently described in a multitude of ways and at different levels of abstraction, including descriptive metadata (e.g., creation date) (see also C6), textual annotations (e.g., keywords), low-level features (e.g., visual features such as SIFT), high-level features (e.g., concepts), and events. Many such annotations are interdependent, e.g., high-level features are generated based on the extracted low-level features, while events may be determined using the identified concepts. Establishing relations among annotations (e.g., determining which visual features were used for the concept annotation process) is important in many settings, particularly when end users are search professionals or researchers.

T4 **Varied granularity access:** In many cases, end users are interested in accessing media content at a granularity level different to that of a multimedia object. When searching for information, for instance, retrieval of only the specific media segments that contain relevant information, instead of entire multimedia objects, reduces users' cognitive load and increases their satisfaction. Such focussed retrieval applications include finding only the shots in a news video relevant to a story or only the image segments where e.g., a specific logo appears. Furthermore, representation at higher levels of granularity, e.g., multimedia collections, is also useful in many contexts. For instance, an aggregated view created by summarising a set of social media posts on the same subject or story provides a snapshot of public opinion on that topic.

T5 **Content provenance:** In several applications, it is important to track the original source of a content item posted online, e.g., to establish whether an image has been previously published in a different context. The ease with which media content is embedded within multimedia documents and shared across diverse platforms (see also C3 and C4) indicates the significance, but also the difficulty of this task. This is further the case when online content undergoes manipulations and alterations, and is subsequently reposted for entertainment (e.g., memes) or malicious (e.g., propaganda) purposes.

T6 **Influentials identification:** When researching a particular story or topic, several types of users (e.g., journalists, analysts, etc.) are interested in identifying influential and relevant content and also, particularly in the case of social media, the content contributors who publish such content. As this is typically achieved by analysing the Web and social link structures, it is paramount to model such relations between multimedia objects, between users and multimedia objects, and also between users (see also C3 and C4).

Based on the above characteristics and tasks, we recognized the requirements that should be satisfied for having an effective data representation model that would enable multimedia information processing, analysis, and access applications in online socially interconnected environments. Without claiming that the above lists are exhaustive, they do cover both the principal aspects of online multimedia settings and their social features (see also Section 3.5). Therefore, our model should represent (in brackets the relevant items from the above lists giving rise to each requirement):

R1 media content of various modalities (C2, T2),

R2 diverse media objects, ranging from mono-modal media items to composite multimedia documents (C2, C3, T2),

R3 media objects across heterogeneous hosting environments in a unified manner (C1, T1),

R4 online relations between media objects (C2, C3, T5, T6),

R5 social interactions between users and media objects (C3, C4, C5, T5, T6),

R6 content contributors, their relations and interactions, as expressed through their accounts in social Web platforms (C4, T6),

R7 granularity at various levels, ranging from media segments to multimedia collections (T4),

R8 rich heterogeneous annotations describing media objects of any granularity, and the relationships between such annotations (T2, T4, T3), and

R9 descriptive metadata as attributes of media objects (C6, T3).

In the sequel, we present the SIMMO model that was developed by considering the aforementioned requirements that are related to the characteristics and tasks of online social multimedia.

3.2 SIMMO description

This Section presents the proposed framework for the unified representation of Socially Interconnected and MultiMedia-enriched Objects (SIMMO) available in web environments. SIMMO consists of a number of core entities and their sub-classes, attributes, and relations that have been determined based on the requirements (R1-R9) identified in the previous Section. While similar entities can be also found, at least in part, in other models (e.g., (Chang et al., 2001; Daras et al., 2011; Bojars et al., 2008)) that were part of our inspiration (see Section 3.5), it is the interconnections among SIMMO elements and the novel approach of bridging the gap between multimedia and social features that make SIMMO unique in its ability to support a wide range of applications.

Figure 3 presents a conceptual model of SIMMO with the following core entities and their sub-classes:

- **Object** is a generic entity representing media content ranging from mono-modal **Items** to multimedia **Documents**. Each Item represents the actual media content consisting of a single modality, such as **Text**, **Image**, **Video**, or **Audio**, whereas

Documents may be viewed as container objects consisting of potentially multiple such Items, and thus modalities. The most common instantiations of Web Documents are **Webpages** (e.g., pages in news sites, in entertainment portals, etc.) or Posts in media sharing platforms with social characteristics (e.g., Facebook posts, tweets, etc.). There are also cases of Webpages consisting of **Posts**; a forum page, for instance, can be viewed as a container object consisting of posts on the same topic. The **Media** entity is introduced as an abstraction of Image, Video, and Audio so as to represent their common characteristics, such as the fact that they all may be associated with a Text item modelling the text associated with them (e.g., a caption) or extracted from them through e.g., ASR (Automatic Speech Recognition) for Video and Audio, and OCR (Optical Character Recognition) for Image and Video. Finally, further media (e.g., 3D objects) may be added as Item instantiations depending on the requirements of the particular application.

- **Source** is a generic entity representing media content contributors. This includes **UserAccounts** representing users generating content, mainly posts in social media sharing platforms where they hold accounts, and **WebDomains** representing the Web sites hosting media content generated by their contributors. WebDomains are viewed as content contributors, even though they do not correspond to the actual person who contributed the content, given that in many cases the information regarding such people may not be available, or may be of much lesser importance in this specific context.
- **Segment** locates the media content of Items at a finer level of granularity (e.g., a passage in text, a region in an image, or a portion of a video) by including positional information as attributes. Instantiations of Segments (not depicted in Figure 3) include **LinearSegments** (e.g., with start/end positions as attributes for referring to text parts), **SpatialSegments** (e.g., with (x, y) pairs as attributes for referring to image regions), **TemporalSegments** (e.g., with start/end times as attributes for referring to video frames/shots/scenes), and **SpatioTemporalSegments**. A **SegmentGroup** represents a collection of Segments; it is also modelled as a sub-class of Segment, thus allowing it to contain both Segments and other SegmentGroups. Figure 4 depicts the Segment instantiations.
- **Collection** models aggregates of Objects (i.e., higher levels of granularity), such as corpora of Web documents, sets of tweets, and image collections.
- **Annotation** is a generic entity representing together with its sub-classes a wide range of descriptions extracted from media content. These include annotations typically extracted from text (e.g., keywords, named entities, summaries, categories, etc.), media content features (e.g., low level descriptors, concepts and events), affective descriptions (e.g., sentiment and polarity), veracity scores reflecting the reliability of information and thus the trust that should be placed on it, and many others. Figure 5 depicts the Annotation instantiations.
- **Topic** refers to any subject of interest in the context of an information processing, analysis, or access applications that users would like to keep track of. Its explicit representation allows to support a broad range of tasks, such as information filtering, topic tracking, and classification.

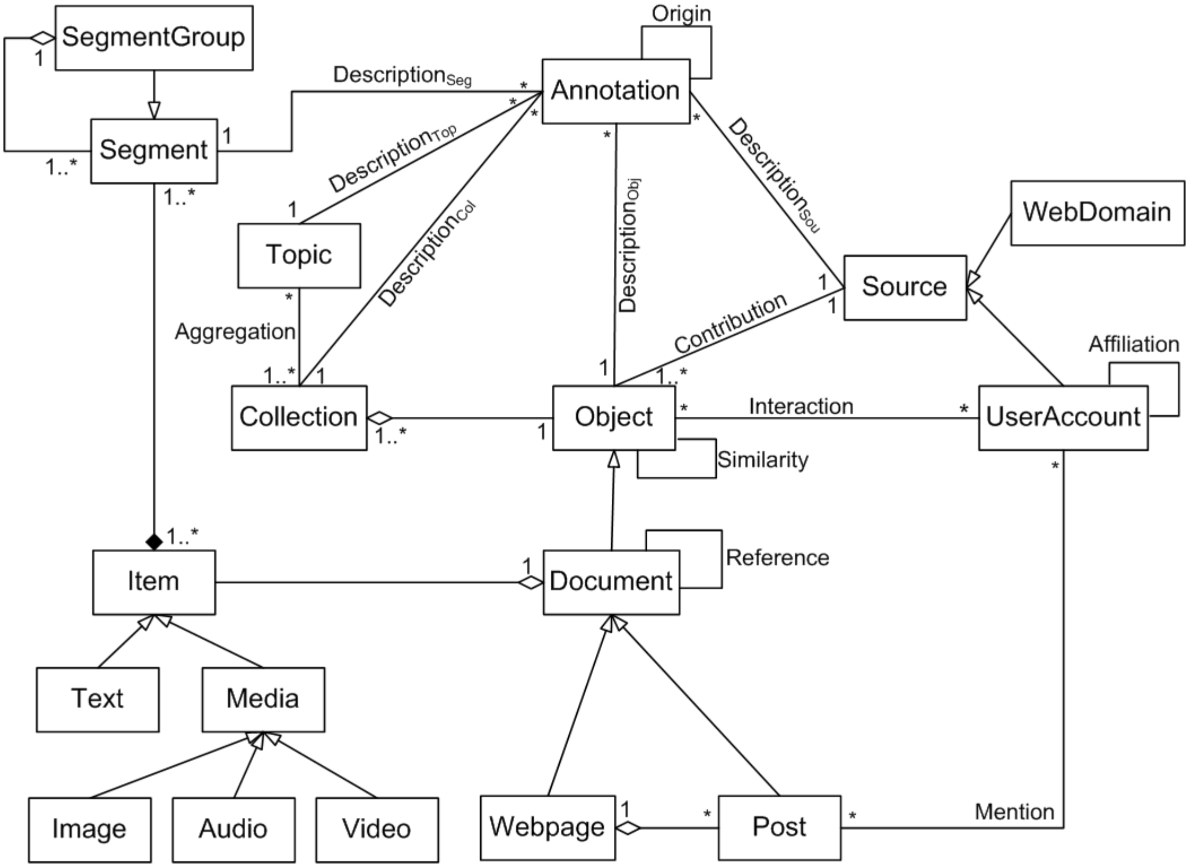


Figure 3: SIMMO conceptual model presenting its elements and their relations.

The main relations between these SIMMO elements, excluding the generalisation and aggregation/composition relations already discussed, are:

- The generation of media objects is modelled through a **Contribution** association between **Source** and **Object**.
- Explicit relations between Documents are modelled as **Reference** associations, with attributes such as the type of the relation. By considering that a Document may Reference another Document, we also consider (through inheritance) that a Webpage may Reference another Webpage (e.g., link to it) and a Post may Reference another Post (e.g., reply to it or comment on it). We consider that this association is also able to model the References to Webpages from Posts (e.g., the Web links embedded in tweets) and to Posts from Webpages (e.g., to the comments dynamically posted on a Webpage).
- Objects may also be implicitly related to other Objects, e.g., through a computation of their similarity. Such **Similarity** relations are modelled as recursive associations between Objects, with attributes such as the type of the relation and the similarity score. This is useful in several applications and tasks, including clustering and verification of content provenance.
- A UserAccount may be involved in several relations, e.g., (i) be mentioned in a Post, (ii) be affiliated with (be friends with, follow etc.) another UserAccount, or (iii) interact with an Object (through likes, shares, views, etc.); the latter is more common for Posts, but users also interact with (e.g., like) whole Webpages. These three relations are modelled through the **Mention**, **Affiliation**, and **Interaction**

assosiations, respectively, with attributes, such as the type of relation and the date it was established. As mentioned above, commenting is not modelled as a relation between Documents and UserAccounts, but rather as a Reference between two Documents (e.g., two Posts).

- All types of entities (i.e., Objects, Segments, Collections, Sources, and Topics) and their sub-classes may be associated with Annotations that are used for describing them. Such **Description** relations represent, for instance, the annotation of an Image with the SIFT features extracted from it, a TemporalSegment of a Video (such as a shot) with Concepts, or a UserAccount with Keywords reflecting the users' profile. Furthermore, links between different annotations (e.g., low-level descriptors and the concepts obtained from them) are modelled through the reflexive relation **Origin** between Annotations to denote the provenance of one with respect to the other.
- Each Topic is associated with a Collection of Objects on the particular subject of interest and may also be annotated itself. For instance, the Topic “Tour de France 2014" bicycle race would be associated with a Collection of Documents, such as Webpages and tweets on the subject, and could be annotated with the concepts “cycling" and “yellow jersey", the entity “Union Cycliste Internationale", and extracted locations, such as “Grenoble, France".

SIMMO elements and their relations also have several attributes representing their properties. For example, each Object is associated with a URI, creation date, and crawl date. Text is described by its format (e.g., HTML), an Image by its size, EXIF data, and associated thumbnail, a Video by its duration, number of frames, and associated thumbnail, and an Audio by its duration. Documents also have attributes related to the statistics regarding their social interactions, e.g., numbers of likes, comments, views, etc. The properties of a UserAccount include a stream ID denoting the platform hosting the account, the user's name, and the number of followers/following/friends. A complete list of the available attributes can be found in our implementation of SIMMO, discussed next.

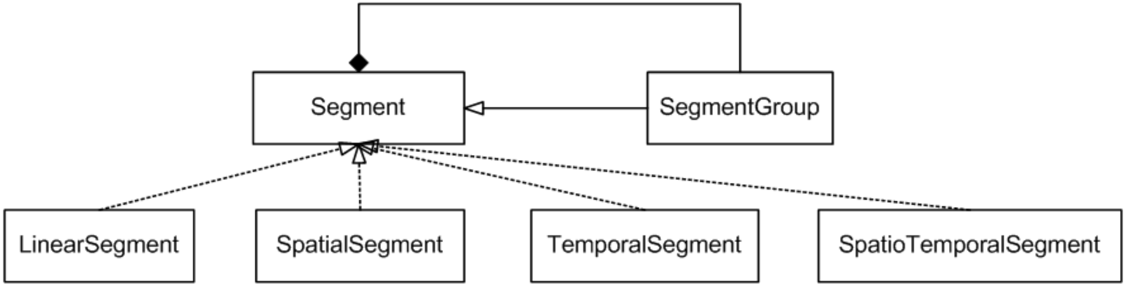


Figure 4: SIMMO Segment instantiations.

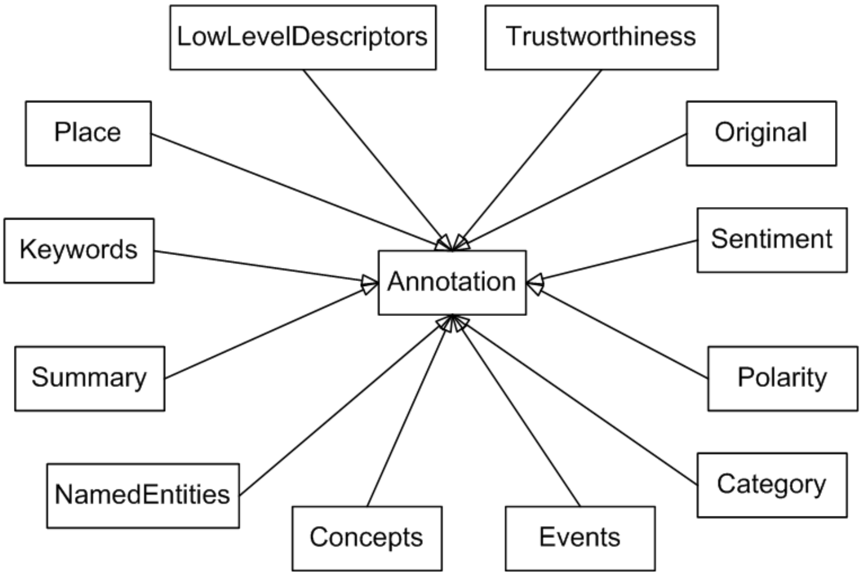


Figure 5: SIMMO Annotation instantiations.

3.3 Implementation of SIMMO

The SIMMO framework is implemented in Java 1.7. Specifically, we have used Apache Maven[3] for controlling the project's build process, unit testing, and documentation creation, and the Google GSON library[4] for converting Java objects into their JSON representation.

As far as Apache Maven is concerned, it is a software project management and comprehension tool. Maven is a build automation tool used primarily for Java projects and it addresses two aspects of building software: a) it describes how software is built, and second, it describes its dependencies. Moreover, it uses an XML file for describing the software project being built, its dependencies on other external modules and components, the build order, directories, and required plug-ins.

Regarding the Google GSON library, it is an open source Java library for serializing and deserializing Java objects to (and from) JSON. It uses reflection, so it does not require additional modifications to classes of (de)serialized objects. It can also handle collections, generic types and nested classes. When deserializing, GSON is navigating the type tree of the object, which is being deserialized. Of course, the use of Google GSON is just an option among the existing JSON libraries or any other serialisation method.

Finally, the SIMMO framework is open-source, released under the Apache License v2, and available at: https://github.com/MKLab-ITI/simmo. Figures 6, 7, and 8 depict the UML (Unified Modeling Language) diagrams of the SIMMO model in general, the classes related to Segment and the ones related to Annotation class. The diagrams capture apart from the relations among the classes, the public variables of the classes.

[3] http://maven.apache.org/

[4] https://code.google.com/p/google-gson/

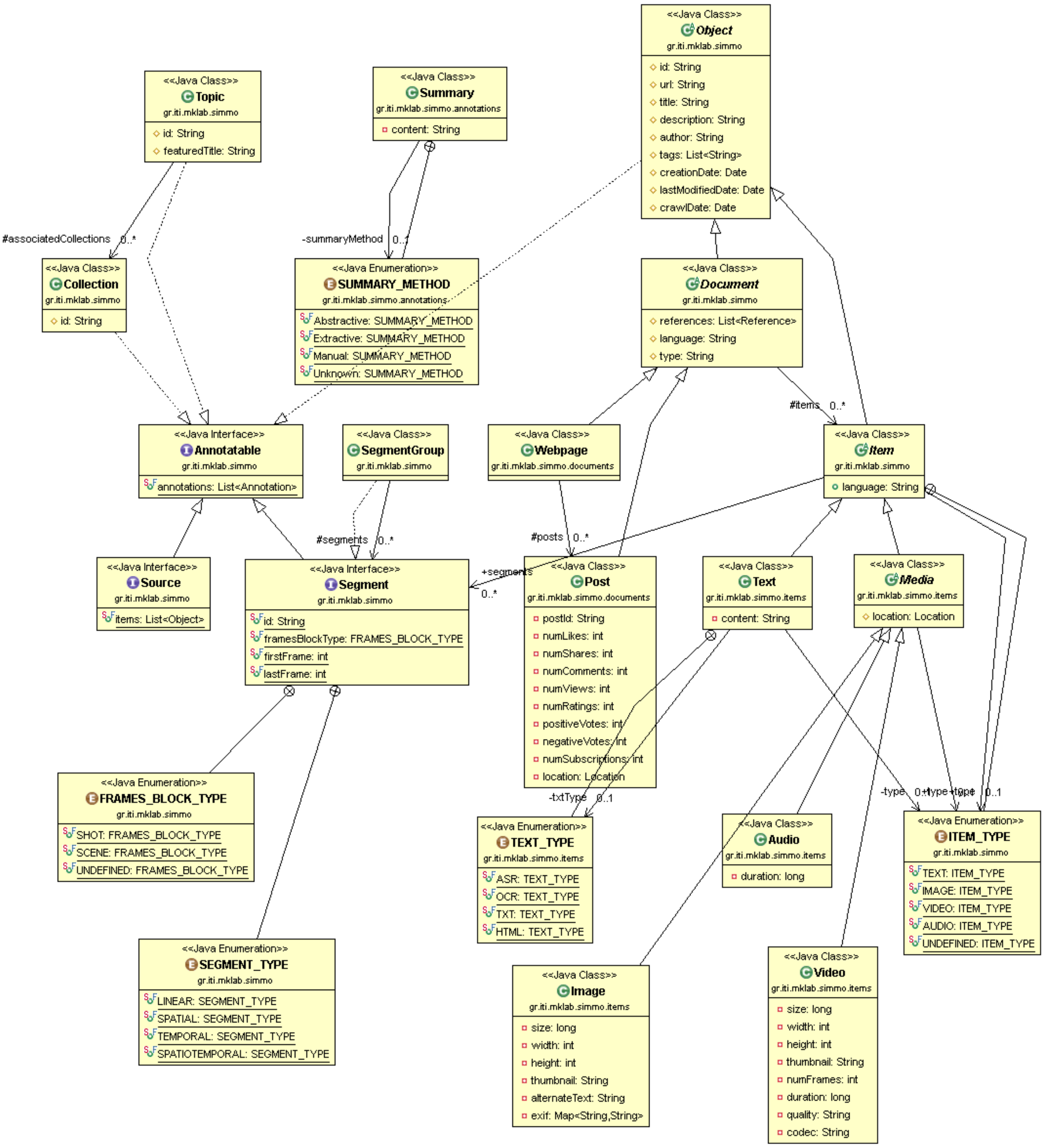


Figure 6: SIMMO UML diagram capturing the variables and the relations among classes.

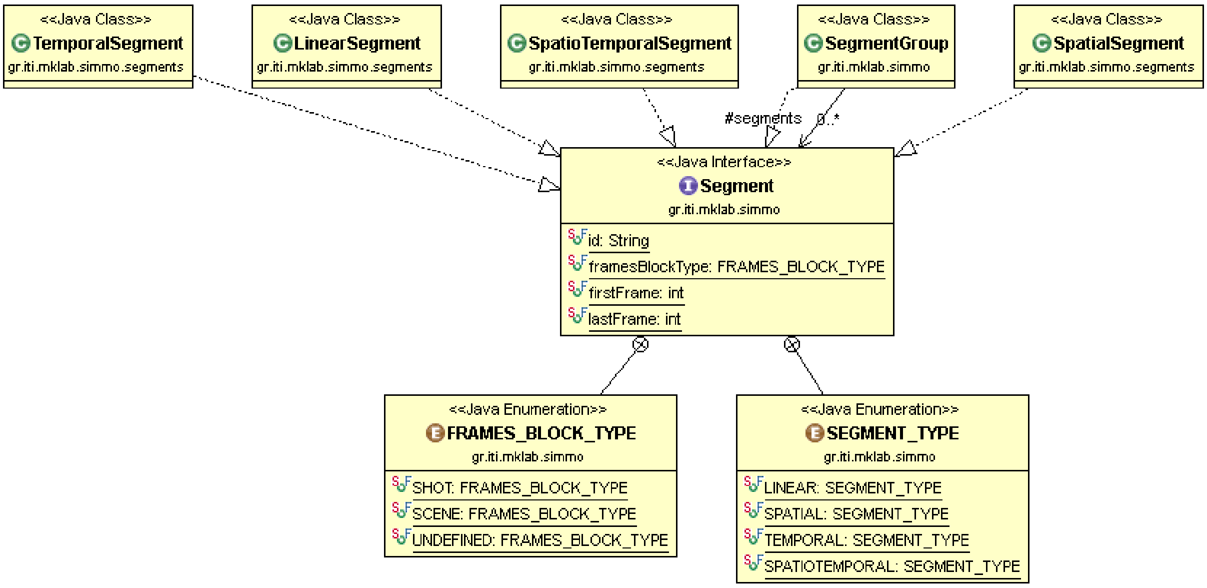


Figure 7: Segment UML diagram capturing the variables and the relations among related classes.

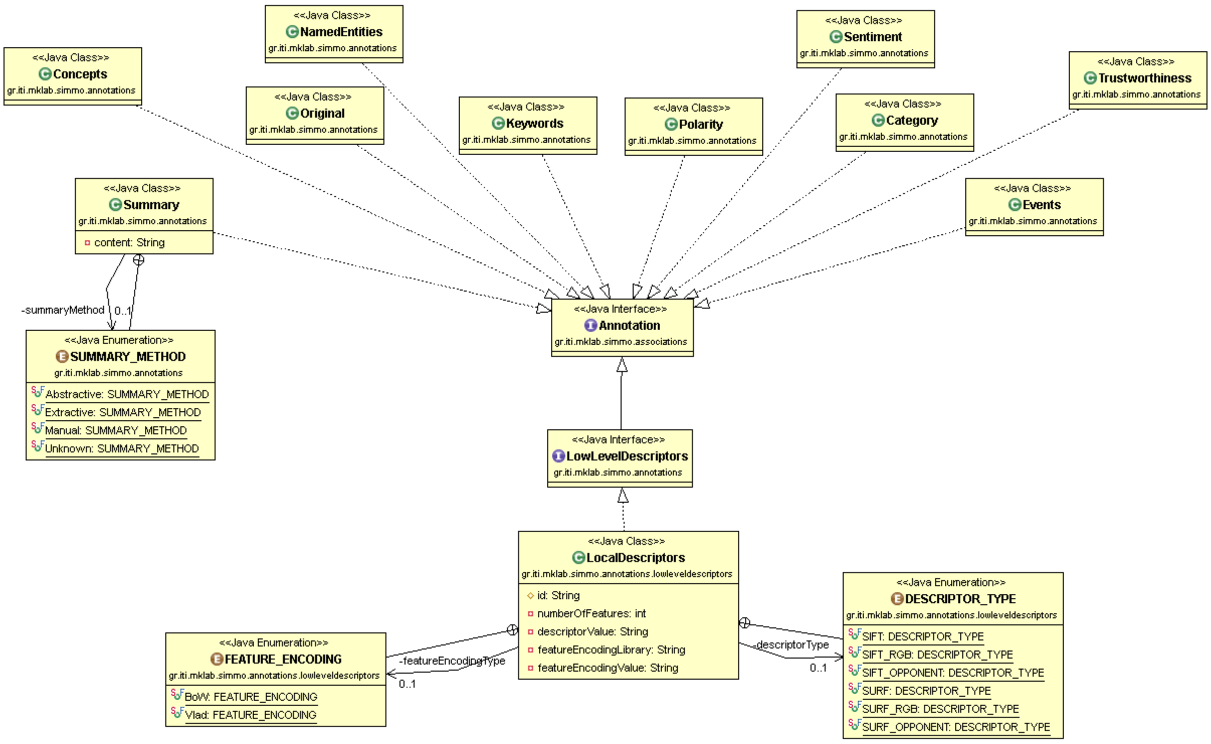

<<Java Class>>
Concepts
gr.iti.mklab.simmo.annotations
<<Java Class>>
NamedEntities
gr.iti.mklab.simmo.annotations
<<Java Class>>
Original
gr.iti.mklab.simmo.annotations
<<Java Class>>
Keywords
gr.iti.mklab.simmo.annotations
<<Java Class>>
Polarity
gr.iti.mklab.simmo.annotations
<<Java Class>>
Sentiment
gr.iti.mklab.simmo.annotations
<<Java Class>>
Category
gr.iti.mklab.simmo.annotations
<<Java Class>>
Trustworthiness
gr.iti.mklab.simmo.annotations
<<Java Class>>
Summary
gr.iti.mklab.simmo.annotations
content: String
<<Java Class>>
Events
gr.iti.mklab.simmo.annotations
-summaryMethod
0..1
<<Java Enumeration>>
SUMMARY_METHOD
gr.iti.mklab.simmo.annotations
Abstractive: SUMMARY_METHOD
Extractive: SUMMARY_METHOD
Manual: SUMMARY_METHOD
Unknown: SUMMARY_METHOD
<<Java Interface>>
Annotation
gr.iti.mklab.simmo.associations
<<Java Interface>>
LowLevelDescriptors
gr.iti.mklab.simmo.annotations
<<Java Class>>
LocalDescriptors
gr.iti.mklab.simmo.annotations.lowleveldescriptors
id: String
numberOfFeatures: int
descriptorValue: String
featureEncodingLibrary: String
featureEncodingValue: String
<<Java Enumeration>>
FEATURE_ENCODING
gr.iti.mklab.simmo.annotations.lowleveldescriptors
BoW: FEATURE_ENCODING
Vlad: FEATURE_ENCODING
-featureEncodingType
0..1
-descriptorType
0..1
<<Java Enumeration>>
DESCRIPTOR_TYPE
gr.iti.mklab.simmo.annotations.lowleveldescriptors
SIFT: DESCRIPTOR_TYPE
SIFT_RGB: DESCRIPTOR_TYPE
SIFT_OPPONENT: DESCRIPTOR_TYPE
SURF: DESCRIPTOR_TYPE
SURF_RGB: DESCRIPTOR_TYPE
SURF_OPPONENT: DESCRIPTOR_TYPE

Figure 8: Annotation UML diagram capturing the variables and the relations among related classes.

Finally, for storing the SIMMO objects, we plan on using the MongoDB[5] database. MongoDB (from "humongous") is a cross-platform document-oriented database that is classified as a NoSQL database. MongoDB eschews the traditional table-based relational database

[5] http://www.mongodb.org/

structure in favour of JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster. More information on the database and the functions used for inserting and retrieving records from it will be presented on the next Deliverable, D4.3.

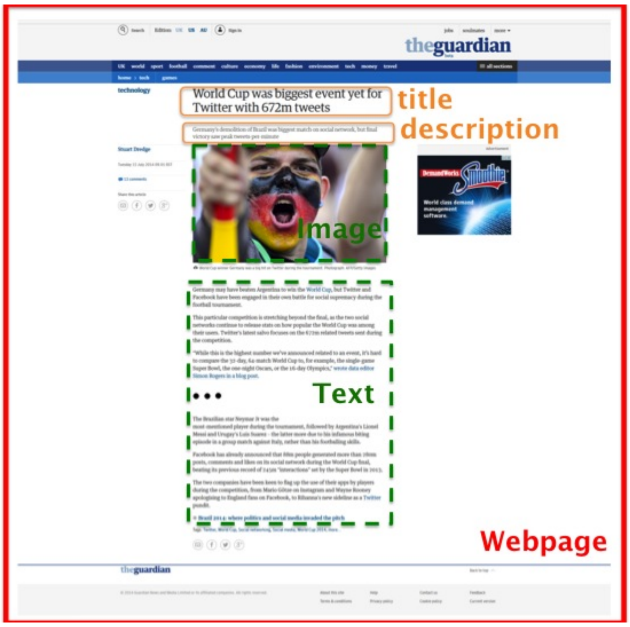
Based on this implementation, examples of commonly encountered online multimedia objects are presented next.

3.4 Application of SIMMO in MULTISENSOR

This section presents some examples of the JSON code used based on the SIMMO specification, for capturing the content of different online multimedia objects that will be dealt in MULTISENSOR. The goal of these examples is to illustrate the flexibility and expressive power of the proposed framework.

3.4.1 SIMMO Examples

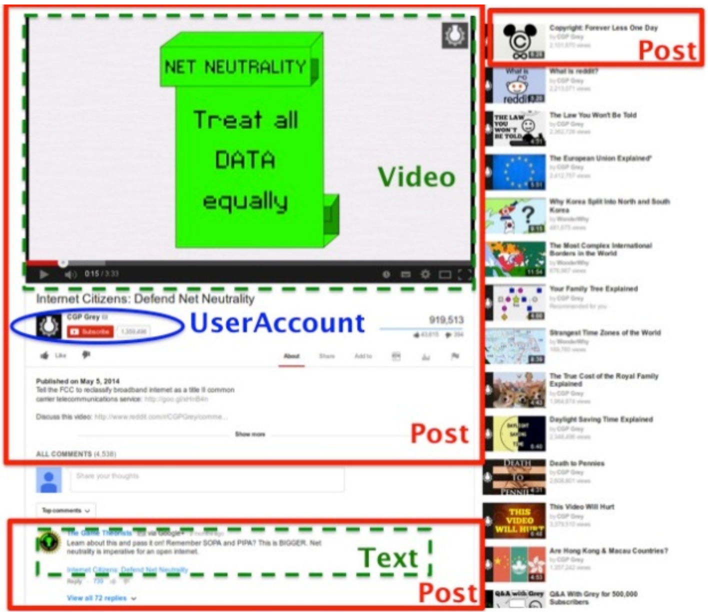
The examples presented in this section capture three typical, but of different type, online multimedia objects with social interconnections: (i) a Web page with content in different modalities and various annotations, (ii) a YouTube video with comments by several users, and (iii) a tweet with social interactions.



World Cup was biggest event yet for Twitter with 672m tweets

Germany's demolition of Brazil was biggest match on social network, but final victory saw peak tweets-per-minute

Figure 10: Example of multimedia document showing a YouTube video with comments.



Internet Citizens: Defend Net Neutrality

CGP Grey

919,513

Published on May 5, 2014
Tell the FCC to reclassify broadband internet as a title II common carrier telecommunications service: http://goo.gl/nHnB4n

Discuss this video: http://www.reddit.com/r/CGPGrey/comme...

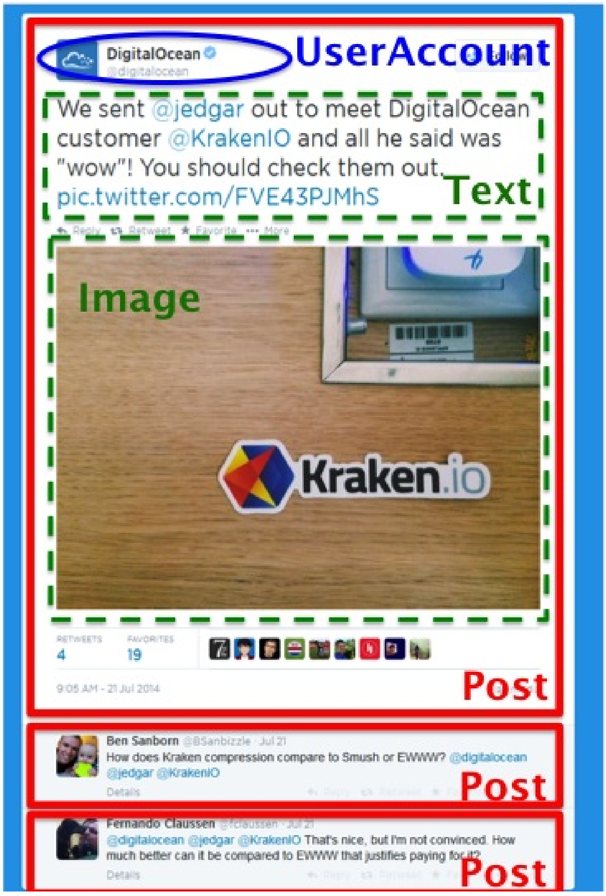
ALL COMMENTS (4,538)

The Game Theorists via Google+
Learn about this and pass it on! Remember SOPA and PIPA? This is BIGGER. Net neutrality is imperative for an open internet.
Internet Citizens: Defend Net Neutrality

View all 72 replies

Copyright: Forever Less One Day
What is reddit?
The Law You Won't Be Told
The European Union Explained*
Why Korea Split Into North and South Korea
The Most Complex International Borders in the World
Your Family Tree Explained
Strangest Time Zones of the World
The True Cost of the Royal Family Explained
Daylight Saving Time Explained
Death to Pennies
This Video Will Hurt
Are Hong Kong & Macau Countries?
Q&A With Grey for 500,000 Subscribers

Figure 11: Example of multimedia document showing a tweet with replies.



DigitalOcean
@digitalocean

We sent @jedgar out to meet DigitalOcean customer @KrakenIO and all he said was "wow"! You should check them out. pic.twitter.com/FVE43PJMhS

RETWEETS 4 FAVORITES 19

9:05 AM - 21 Jul 2014

Ben Sanborn @BSanbizzle · Jul 21
How does Kraken compression compare to Smush or EWWW? @digitalocean @jedgar @KrakenIO

Fernando Claussen @fclaussen · Jul 21
@digitalocean @jedgar @KrakenIO That's nice, but I'm not convinced. How much better can it be compared to EWWW that justifies paying for it?

Figure 12: Example of multimedia document showing a Web page in the news domain.

Consider, for instance, a Web page from an online newspaper discussing the recent World Cup 2014 final (Figure 10). SIMMO models this as a Webpage corresponding to the following JSON (some URLs have been shortened):

```
<!-- Webpage.json -->
{ /* Webpage */
```

```
  "id":"180444840287",
  "url":"http://goo.gl/5JRsHi",
  "title":"World Cup was biggest event yet for Twitter with 672m tweets",
  "description":"Germany's demolition of Brazil ... peak tweets-per-minute",
  "tags":["Twitter", "World Cup", "Social networking", ..., "Digital media",
"Internet" ],
  "creationDate":"Jul 15, 2014 8:01:20 AM",
  "items":[
    {
      "type": "TEXT",
      "textFormat":"HTML",
      "content":"<p>Germany may have beaten Argentina to win the World Cup,..."
    },
    {
      "type":"IMAGE",
      "url":"http://goo.gl/Uh4okO",
      "width":620,
      "height":372
      "describedBy":[{"type":"LOWLEVELDESCRIPTOR", "annotationId":"A8423"},...],
    }
  ],
  "references":[ { "type":"LINK", "referencedDocumentId":"180444840289"},...],
  "describedBy":[ { "type":"SUMMARY", "annotationId":"A9765" }, .... ]
}
```

The Webpage has particular attributes, such as title and description, and contains HTML Text and an Image, each with its own properties. Both the Webpage and its constituent Items may be annotated (e.g., the Webpage with a summary and the Image with visual features, listed below as separate JSON entries). The Webpage also connects to other Webpages through References of type "LINK".

```
{ /* Summary */
  "id":"A9765"
  "summaryMethod":"Manual",
  "content":"Germany may have beaten Argentina to win the World Cup, ..."
}

{ /* LowLevelDescriptor */
  "id":"A8423"
  "descriptorType":"SURF",
  "numberOfFeatures":128,
  "descriptorValue":"128 1035 <CIRCLE 470 276 1 0 0>; 0.000537204 0.000681045
... 0.00020111"
}
```

The next example corresponds to a YouTube video (Figure 11) contributed by a UserAccount and modelled as a Post consisting of the actual video content and References to its comments, each also modelled as a Post. Several social interaction features are also modelled as attributes, such as the number of subscriptions to the UserAccount and the number of views of the video.

```
<!-- Post.json -->

{
  "id":"wtt2aSV8wdw",
  "url":"https://www.youtube.com/watch?v=wtt2aSV8wdw",
  "title":"Internet Citizens: Defend Net Neutrality",
  "description":"Tell the FCC to reclassify broadband internet ...",
  "creationDate":"May 5, 2014 4:07:17 PM",
  "createdBy":"acc98754",
  "items":[
    {
      "type":"VIDEO",
      "url":"https://www.youtube.com/v/wtt2aSV8wdw",
      "width":1280,
      "height":720,
     "duration":213,
    }
  ],
  "numComments":4538,
  "numViews":919353,
  "positiveVotes":43615,
  "negativeVotes":394,
  "references":[ { "type":"COMMENT", "referencedDocumentId":"409sfh" }, ... ]
}

{ /* UserAccount */
  "id":"acc98754",
  "name":"CGP Grey",
  "numSubscriptions":1361024,
  "avatarSmall":"http://goo.gl/YJS4PG"
}

{ /* Post */
  "id":"409sfh",
  "createdBy":"acc74528"
  "items":[
    {
      "type":"TEXT"
      "textFormat":"HTML",
      "content":"<div class="Ct">Learn about this and pass it on! ... </div>",
    }
  ],
  "numComments":72,
  "positiveVotes":739,
  "negativeVotes":0
}
```

The final example corresponds to a tweet (Figure 12) modelled as a Post that contains both Text and an Image, together with Mentions to specific UserAccounts, while statistics of social interactions are represented by attributes. Replies to the tweet are also modelled as Posts, which are not listed here for simplicity.

```
 <!-- Post.json -->

 {
   "id": "491252639225901056",
   "createdBy": "digitalocean"
   "creationDate": "Jul 21, 2014 4:05:30 PM",
   "items": [
     {
       "type": "TEXT"
       "textFormat": "HTML",
       "content": "We sent @jedgar out to meet DigitalOcean customer @KrakenIO
 and all ..."
     }
     {
       "type": "IMAGE"
       "url": "http://pbs.twimg.com/media/BtFHq9ZCUAAhyho.jpg:large",
       "height": 768,
       "width": 1024,
     }
   ],
   "mentions":[ { "mentioned":"jedgar" }, ... ],
   "numShares": 4,
   "positiveVotes": 19,
   "negativeVotes": 0,
   "references":[ { "type":"REPLY", "referencedDocumentId":"491255375912370176"
 },...]
 }
```

3.5 Comparison with existing approaches

To assess the expressive power of SIMMO, we compare it to other multimedia data representation models. Initially, we present the existing approaches together with a short presentation of the modeling abilities, which is followed by a comparison which is performed on the basis of the requirements identified in Section 3.1.

There were some first, early attempts to describe the content and structure of multimedia data such as the one proposed by (Caetano and Guimaraes, 1998), but these were soon superseded by the MPEG-7 standard (Chang et al., 2001). The MPEG-7 standard is a generic, but complex, framework that enables highly structural, detailed descriptions of multimedia content at different granularity levels. It relies on: (i) Descriptors (D) defining the syntax and semantics of diverse features, (ii) Description Schemes (DS) describing the structure and semantics of relations among D or DS, (iii) a Description Definition Language allowing the creation and modification of DS and D, and (iv) Systems Tools, supporting various tasks, e.g., synchronisation of descriptions with content.

Later, the MPEG-21 (Burnett et al., 2003) followed as an open framework for multimedia delivery and consumption, focussing on how the elements of a multimedia application infrastructure should relate, integrate, and interact. MPEG-21 centres around the concept of Digital Items, i.e., structured objects with multimedia content and metadata, and Users interacting with them; it also puts particular emphasis on Intellectual Property issues and mechanisms for the management of rights.

More recently, the Rich Unified Content Description (RUCoD) (Daras et al., 2011) framework was introduced, within the context of the European project I-SEARCH[6], for representing intrinsic properties of multimedia Content Objects, enhanced with real-world information (e.g., geo-location) and affective descriptions (e.g., in the valence/arousal 2D space). Each RUCoD consists of: (i) a header containing descriptive metadata (e.g., id and creation date) together with information about the media it contains and their descriptors, (ii) low-level descriptors, (iii) real-world descriptors, and (iv) user-related descriptors.

Another related platform was WebLab[7] that was developed by WebLab (Giroux et al., 2008). The basic development of WebLab was realized during the EU project VITALAS[8], but extensions of the platform were realized during other projects as well such as VIRTUOSO[9], WebContent[10], AXES[11] and TWIRL[12]. The WebLab platform (Giroux et al., 2008) which was developed for integrating multimedia information processing components also defined a common exchange format to support the communication between such components. This exchange format, in essence a multimedia data representation model, centres on the notion

[6] http://www.isearch-project.eu/isearch/

[7] http://weblab-project.org/index.php?title=WebLab

[8] http://vitalas.ercim.eu/

[9] http://www.virtuoso.eu/

[10] http://www.webcontent.fr/scripts/home/publigen/content/templates/show.asp?L=EN&P=55

[11] http://www.axes-project.eu/

[12] http://twirl-project.eu/

of Resource that models several types of entities, including content in various modalities, multimedia documents and their segments, and diverse annotations.

The models discussed above focus on the description of multimedia content and thus satisfy requirements R1, R2, R7, R8 and R9, listed in Section 3.1; see also Table 1 for an overview of the requirements satisfied by each model. Given though that most were developed prior to the explosion of social media, they do not take into account the social characteristics and interconnections in current web environments (requirement R5). Such aspects have been addressed by ontologies, such as SIOC (Bojars et al., 2008) and FOAF (Brickley and Miller, 2012). SIOC (Socially-Interlinked Online Communities) captures the nature, structure, and content of online communities (such as forums) through the representation of Users creating Posts organised in Forums that are hosted on Sites; modelled as sub-classes of the generic concepts Item, Container, and Space, respectively. SIOC is commonly used in conjunction with the FOAF (Friend Of A Friend) vocabulary to express users' personal information and social networking interactions. These approaches are not concerned though with the potential multimodality of Posts/Items and the annotations extracted from such multimedia content (requirements R2 and R8), that are of paramount importance in information processing, analysis, and access tasks.

SIMMO bridges the gap between these perspectives by modelling both multimedia content (and its descriptions) and also users' social interactions with such content and with each other; see Table 1. To this end, SIMMO has borrowed several elements from the aforementioned approaches, while it has also introduced new aspects to support the emerging needs and requirements. For instance, the SIMMO multimedia content description draws many ideas from MPEG standards, but eschews their complexity, while SIMMO Annotations instantiated as LowLevelDescriptors could be mapped to standardised MPEG-7 Descriptors. Modelling granularity at the Segment level has been inspired by WebLab, while RUCoD has motivated the incorporation of affective and real-world features. The concept of UserAccount has been borrowed by FOAF, while the Post and Forum SIOC elements could be mapped to the Post and Webpage SIMMO components. Finally, many of the attributes of media Objects, Documents, and Items are equivalent to those proposed by Dublin Core (http://www.dublincore.org).

Requirement: brief description	MPEG-7	RUCoD	WebLab	SIOC+FOAF	SIMMO
R1: multiple modalities	✓	✓	✓		✓
R2: diverse media objects	✓	✓	✓		✓
R3: heterogeneous hosts	✓	✓	✓	✓	✓
R4: online links		✓		✓	✓
R5: social interactions				✓	✓
R6: contributors (description+relations)	~			✓	✓
R7: granularity at different levels	✓	✓	✓		✓
R8: various annotations	✓	✓	✓		✓
R9: descriptive metadata	✓	✓	✓	✓	✓

Table 1: Comparison of different models w.r.t. the requirements identified in Section 3.1 (✓= requirement is satisfied; ~ = requirement is partly satisfied)

4 TOPIC-BASED CLASSIFICATION

This activity deals with the classification of News Items retrieved from the News Repository into categories and it involves the following sub-activities:

- Classification of news items by using their multimodal features extracted in other WPs (WP2 and WP3) and supervised learning techniques such as Support Vector Machines (SVM) and Random Forests (RF).
- Rule-based classification of the news items by exploiting the ontology framework developed in WP5.

Therefore, category-based classification component receives as input the multimodal features existing in the indexing structure developed in T4.4 and outputs the degree of confidence of each category for the specific News Item. Although the proposed classification framework is currently developed in the statistical programming language R, the final component will be developed in Java.

The category-based classification component uses the multimodal features created in WP2 (textual and visual concepts) that are stored in the indexing structure developed in WP4 (T4.4) for the classification. In future implementations, it will use also the ontologies developed in WP5 for realizing the rule-based classification. In the current Deliverable we didn’t address rule-based classification, since the content has not yet been transformed to triplet format (WP2) in the knowledge base (WP5) and therefore the rule-based classification will be investigated in year 2 and reported in D4.3.

4.1 Relevant work on category-based classification

In this Section the relevant work on category-based classification is presented. Since the task is related to Document classification (DC) in general and to News Items classification in particular, previous work in these two fields is provided separately. Furthermore, related studies regarding classification using multimodal features are described in a separate Subsection.

4.1.1 Document classification

Document classification (DC) deals with the task of assigning a document to one or more predefined categories (Sebastiani, 2002). Over the years, numerous supervised machine learning methods have been applied to DC. Among others, Naïve Bayes (Ting et al., 2011), Rocchio (Zeng and Huang, 2012), Self-Organizing Maps (Saarikoski et al., 2011), SVM (Ho et al., 2013) and RF (Klassen and Paturi, 2010) can be listed. It must be noted that all these studies make use of textual features. Other studies investigate text classification by using semi-supervised machine learning methods (see for example (Braga et al., 2009) and (Shi et al., 2011)).

A number of DC-related studies deal specifically with documents in web page format. For instance, (Selamat and Omatu, 2004) apply Neural Networks and Principal Component Analysis for web page feature selection and classification. Furthermore, (Aung and Hla, 2009) employ a RF classifier for multi-category web page classification. Finally, (Xu et al., 2011) present a classification algorithm for web pages, Link Information Categorization (LIC), which is based on the K Nearest Neighbour (KNN) method and makes full use of the characteristics of Web information. A detailed review of the algorithms used for web page classification can be found in (Qi and Davison, 2009).

In addition to the aforementioned studies, which only use textual features, there have also been some studies for DC that make use of visual features. For example, (Shin et al., 2001) apply decision tree and self-organizing map classifiers to categorize document page images, using image features that express “visual similarity” of layout structure. In another work, (Chen et al., 2006) explore image clustering as a basis for constructing visual words for representing documents. Then they apply the bag-of-words representation and standard classification methods to train an image-based classifier.

4.1.2 **News items classification**

In the relevant literature, there are several research studies that examine news items classification specifically. For example, (Gamon et al., 2008) apply a maximum entropy classifier on unigram features to detect emotional charge around links to news articles in posts from political weblogs. In another work, (Lin et al., 2008) automatically classify documents into reader-emotion categories and investigate how different feature settings affect classification performance. They demonstrate that certain feature combinations achieve good accuracy. Furthermore, (Bandari et al., 2012) employ both regression and classification algorithms that make use of social interaction features to the task of predicting the popularity of news items on the social web. The results of the study show that one of the most important predictors of popularity is the source of the article. However, in a recent study, (Arapakis et al., 2014) show that predicting the popularity of news articles based on features available at cold start (such as the ones used by (Bandari et al., 2012)) is not a feasible task, given the existing techniques. Specifically, they demonstrate that the methods used are biased to predict the most frequent class of unpopular articles with more accuracy than the less frequent class of popular articles. Finally, (Swezey et al., 2012) utilize news articles streams and Bayesian text classification in order to classify contents of interest related to geographic communities.

4.1.3 **Multimodal classification**

Classification approaches that make use of features from multiple modalities have been investigated in various studies. (Chen and Hauptmann, 2004) propose an approach to build a robust news video classifier that integrates content from multiple media. More specifically, they apply Fisher’s Linear Discriminant for feature selection, they concatenate the projections from multiple sources into a single feature vector as the combination strategy (early fusion), and apply classification using this representation. In another study, (Montagnuolo and Messina, 2009) present a methodology for genre classification of television programmes. They extract feature sets from four informative sources, they train a separate neural network classifier for each feature set and they combine the outputs of the individual classifiers by an ensemble method to get the final classification. Finally, (Glodek et al., 2011) employ a Conditioned Hidden Markov Model (CHMM) to the problem of naturalistic multiparty dialogue detection. They utilize two modalities in their data set and also focus on late classifier fusion techniques.

In addition, a number of studies have incorporated deep learning methods in the task of multimodal classification. (Ngiam et al., 2011) introduce a deep networks-based methodology for multimodal learning. They utilize an extension of Restricted Boltzmann Machines (RBMs) with sparsity as a building block, in order to construct models for learning single modality representations (audio and video). The combination of these models leads to shared representation learning settings with better results. They employ the proposed models for unsupervised feature learning on a number of data sets and they test their performance through supervised classification (by means of a linear SVM). Moreover, (Srivastava and Salakhutdinov, 2012) propose a Deep Boltzmann Machine (DBM) model for learning multimodal data representations. The model's key concept is learning a probability density over the space of multimodal inputs. In addition, the model is able to extract a representation even when some of the modalities are absent. The authors demonstrate the effectiveness of the methodology in classification and information retrieval tasks. Regarding classification, the model is trained using both unlabeled and labeled bimodal data (images and text) and the extracted fused representation is used for classification by logistic regression. In the information retrieval task, DBM is compared to other deep learning methods by using both unimodal and multimodal queries and the results indicate that DBM performs the best among the compared models.

4.2 Category-based classification based on multimodal features

4.2.1 Theoretical background – Proposed classification framework

Theoretical background

The machine learning method chosen for the proposed classification framework is RF (Breiman, 2001). The main reason for selecting RF is the fact that it can handle multiclass classification tasks directly. Therefore, the need to develop a classification model for each class/category separately is eliminated. Additionally, an important motivation for using RF was the application of late fusion strategies based on the operational capabilities of the method. A brief description of the theoretical background of RF is provided below.

Random Forests (RF) is an ensemble learning method for classification and regression (Breiman, 2001). The basic notion of the methodology is the construction of a multitude of decision trees. Within RF's operational procedures, two sources of randomness are employed:

1. A different bootstrap sample, drawn randomly from the training data, is used for the construction of each decision tree.
2. At each node split during the construction of a decision tree, a random subset of *p* variables is selected from the original variable set and the best split based on these *p* variables is used.

For the prediction of an unknown case, the outputs of the trees that are constructed by RF are aggregated (majority voting for classification / averaging for regression). For a model consisting of *T* trees, the following equation is used for predicting the class label l of a case *y* through majority voting:

$$l(y) = argmax_c(\sum_{n=1}^{T} I_{h_n(y)=c}) \tag{1}$$

where I the indicator function and *h_n* the *nth* tree of the RF.

RF has the ability to provide an estimation of its generalization error through an internal mechanism called Out-Of-Bag (OOB) error estimate. During the construction of each tree, only 2/3 of the original data's cases are used in that particular bootstrap sample. The rest 1/3 of the instances (OOB data) are predicted by the constructed tree and thus, used for testing its performance. The OOB error estimate is the averaged prediction error for each training case *y*, using only the predictions of the trees that do not contain *y* in their bootstrap sample. In general, it is not considered necessary to perform cross-validation during the training of a RF model. This is because of the fact that the OOB error estimate is actually an indicative performance score based on cases that do not take part in the training procedure of RF (the OOB data). Furthermore, RF can supply a matrix that contains proximities between the training cases. This is achieved by putting all the training cases down each tree and based on the frequency that pairs of cases end up in the same terminal nodes, this proximity matrix is computed.

Apart from the RF-related studies mentioned in Section 4.1, there are many successful applications of RF to a wide range of disciplines in the relevant literature. Among others, image classification (Xu et al., 2012), web spam classification (Erdélyi et al., 2011), network intrusion detection (Li and Meng, 2013) and neuroimaging (Gray et al., 2013) can be listed. Moreover, several modifications and improvements of the RF algorithm have been proposed, such as (Robnik-Šikonja, 2004).

Proposed classification framework

In Figure 13, the flowchart of the proposed classification framework (training phase) is illustrated. Next, the different steps and notions of the framework are described in detail.

First of all, it is assumed that each News Item is represented by a number of modalities (for instance, textual modality, visual modality etc.). By applying certain procedures, we extract a number of features from the raw data of each modality, thus formulating the corresponding feature vectors that serve as input for the construction of the classification models. At this point, it should be noted that we chose to follow the approach of treating each modality's features separately, instead of concatenating all the features into one large vector. In this way, we are able to exploit the representation and the information contained in each modality in an independent manner.

As a result of the aforementioned approach, in the training phase a separate RF model is trained for each modality. In order to formulate a final fused RF model, we apply a late fusion strategy by computing weights for each modality's RF outputs. For the computation of the modality weights, three different methods that exploit the operational procedures of RF are applied:

OOB error estimate: The underlying notion here is that if a RF model is able to predict the OOB cases for one or more classes efficiently, it is expected to perform equally well on unknown cases. Therefore, from each modality's RF model, the corresponding OOB accuracy values are computed. This is done for each class separately. Then, the accuracy values are normalized (by dividing them by their sum) and serve as weights for the RF models' outputs, e.g. for class l:

- ${}_{acc}\mathrm{OOB}_{li}$: OOB accuracy value for class *l* for modality i (i=1…N, N=number of modalities)
- W_{li}: $\frac{\mathrm{accOOB}li}{\sum_{j=1}^{N}\mathrm{accOOB}lj}$ (weight for class *l* for modality i) (2)

Proximity ratio: For the second weighting strategy, the proximity matrix of a RF model is taken into consideration. First, for each RF the proximity matrix between all pairs of data cases *P*={p_{ij},*i*,*j* =1, …,*w*} (*w*=number of data cases) is constructed. Next, the proximity ratio values between the inner-class and the intra-class proximities (for each class) are computed (Zhou et al., 2010) as in the following equation:

$$R = \frac{P_{inner}}{P_{intra}} \tag{3}$$

where

$$P_{inner} = \sum_{i,j=1}^{w} p_{ij}\ (if\ l_i = l_j) \tag{4}$$

$$P_{intra} = \sum_{i,j=1}^{w} p_{ij}\ (if\ l_i \neq l_j) \tag{5}$$

and l_i, l_j the class labels of cases *i* and *j*, respectively. Finally, for each modality and for each class, the proximity ratio values are first averaged and then normalized (by dividing them by their sum), in order to be used as modality weights for the RF models, e.g. for class l:

- ${}_{mean}\mathrm{R}_{li}$: Averaged proximity ratio value for class *l* for modality i (i=1…N, N=number of modalities)
- W_{li}: $\frac{\mathrm{meanR}li}{\sum_{j=1}^{N}\mathrm{meanR}lj}$ (weight for class *l* for modality i) (6)

A large proximity ratio value for a class is an indication that the cases of that class are encountered frequently in the terminal nodes of a RF model's trees (inner-class proximity) and are not intermixed with cases from other classes (intra-class proximity). Thus, the larger the proximity ratio value for a class, the better the performance of the RF model for that class can be considered.

Adjusted proximity ratio: This approach takes into account the two aforementioned weighting strategies (OOB error estimate and proximity ratio). It is used for adjusting the proximity ratio values, in cases where one or more classes for a modality's RF model exhibit high averaged proximity ratio values but disproportionally low OOB accuracy values. As a result, the weights assigned to these classes will be biased towards the "worse" modality (in terms of accuracy performance) and this will affect the late fused RF outputs. To overcome this, for each class and for each modality, the averaged proximity ratio values are multiplied by the corresponding OOB accuracy values, in order to formulate the adjusted proximity ratio values as in the following equation:

$$R_{adjusted} = R * OOB_{accuracy} \tag{7}$$

After the computation of the adjusted proximity ratio values, the same normalization procedure (as in the other two weighting strategies) is applied, e.g. for class l:

- $_{\text{mean}}$Radj$_{li}$: Averaged adjusted proximity ratio value for class *l* for modality i (i=1…N, N=number of modalities)
- W$_{li}$: $\frac{\text{meanRadj}li}{\sum_{j=1}^{N}\text{meanRadj}lj}$ (weight for class *l* for modality i) (8)

During the testing phase, for the prediction of an unknown case, RF outputs probability estimates per class for that case. The probability outputs P_1, P_2, …, P_N (*N*=number of modalities) from the RF models are multiplied by their corresponding modality weights W_1, W_2, …, W_N and summed to produce the final RF predictions, e.g. for class l:

$$P_l^{fused} = W_{l1}P_{l1} + W_{l2}P_{l2}+\ldots+W_{lN}P_{lN} \tag{9}$$

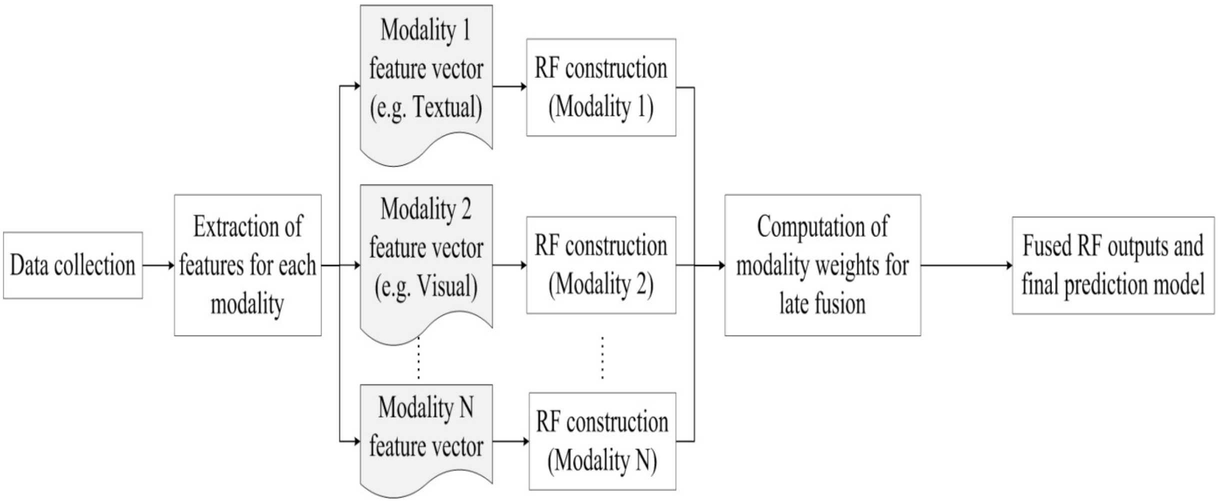

Data collection
Extraction of features for each modality
Modality 1 feature vector (e.g. Textual)
Modality 2 feature vector (e.g. Visual)
Modality N feature vector
RF construction (Modality 1)
RF construction (Modality 2)
RF construction (Modality N)
Computation of modality weights for late fusion
Fused RF outputs and final prediction model

Figure 13: Proposed classification framework (training phase).

4.3 IPTC news codes taxonomy

In this Section the rationale for defining the list of topic categories for the conducted experiments is presented.

The specification of the topic categories was based on the International Press Telecommunications Council (IPTC) news codes taxonomy. The definition of IPTC (as provided in http://en.wikipedia.org/wiki/International_Press_Telecommunications_Council) is the following:

The International Press Telecommunications Council (IPTC) is a consortium of the world's major news agencies, other news providers and news industry vendors and acts as the global standards body of the news media. Currently more than 50 companies and organizations from the news industry are members of the IPTC, including global players like Associated Press (AP), Agence France-Presse (AFP), Deutsche Presse-Agentur (DPA), BBC, Getty Images, Press Association (PA), Reuters and The New York Times.

IPTC aims at simplifying the distribution of information. To achieve this technical standards are developed to improve the management and exchange of information between content providers, intermediaries and consumers. IPTC is committed to open standards and makes all standards freely available to its members and the wider community. Some examples of the technical standards developed are:

- Photo metadata
- IPTC NewsML-G2-Standards Family
- NewsCodes
- NITF - News Industry Text Format
- RightsML
- rNews
- SportsML
- IPTC 7901

In this case, we are interested in the NewsCodes standard. The NewsCodes includes metadata taxonomies for News. Specifically, IPTC creates and maintains sets of concepts - called a controlled vocabulary or taxonomy - to be assigned as metadata values to news objects like text, photographs, graphics, audio - and video files and streams. This allows for a consistent coding of news metadata across news providers and over the course of time (that is why they are called IPTC NewsCodes).

For an easy overview the NewsCodes are organized into the following groups:

- Descriptive NewsCodes to categorize news content, including Media Topics, Subject Codes, Genres, Scene Codes, and World Regions.
- Administrative NewsCodes to properly administrate news items, including Audio- and Video-Codecs, Colorspace, News Product and News Provider.
- NewsML-G2 NewsCodes for the specific use with the NewsML-G2 news exchange format.
- NewsML 1 NewsCodes for the specific use with the NewsML 1 news exchange format.
- Photo Metadata NewsCodes to be used with corresponding fields of photo metadata panels: Subject Codes, Scene Codes and Digital Source Type.

We are considering the Descriptive NewsCodes that contain the following categories:

- **Genre**: Indicates a nature, journalistic or intellectual characteristic of an item.
- **Media Topic**: Indicates a subject of an item.
- **Scene**: Indicates a type of scene covered by an item.
- **Subject Code**: Indicates a subject of an item.
- **Subject Qualifier**: Indicates a narrower attribute-like context for a Subject Code, e.g. for sports: the gender of participants, indoor/outdoor competition etc.
- **World Region**: Indicates a region of the world.

We are using the Media Topic NewsCodes, which contains 1100 terms with a focus on text. It has 17 top level terms and the depth of the tree is up to 5 levels. Figure 14 depicts the first level of the taxonomy that shows the upper/ high level categories.

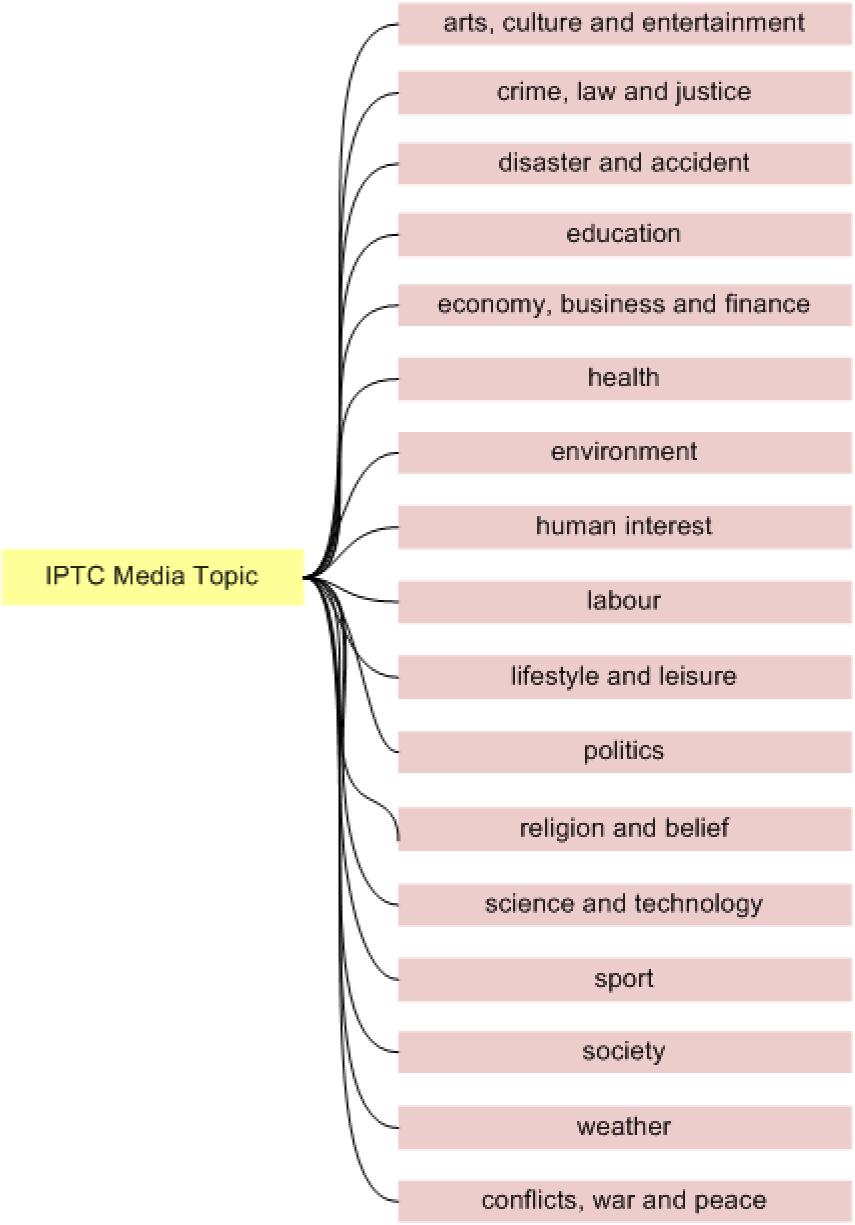


Figure 14: Top level of IPTC Media Topic NewsCodes taxonomy

Within MULTISENSOR and after an extensive discussion with the user experts, a set of topics that are of concern for each use case were defined.

4.4 Experiments

In this Section the results from the application of the proposed classification framework to two datasets, namely the News sites dataset and the MULTISENSOR dataset, are provided and described. It must be noted that part of the results from the News sites dataset experiments have been accepted for presentation at the 7th Information Retrieval Facility Conference (IRFC2014) (Liparas et al., 2014).

4.4.1 News sites dataset

Dataset description

The dataset contains web pages from three well known News Web Sites, namely BBC, The Guardian and Reuter. Overall, 651, 556 and 360 web pages were retrieved from each site, respectively. It was deemed necessary to annotate the web pages manually, regardless of the fact that in the three News Web Sites descriptions about the topic of each web page are provided, since in many cases the descriptions are inconsistent with the content of the web pages. The manual annotation was realized for a subset of the topics recognized by the IPTC news codes taxonomy[13], which is the global standards body of the news media. Specifically, the most important topics were selected with the guidance of media monitoring experts and journalists. Table 2 contains a detailed description of the dataset[14] and the topics considered.

Topics **News Sites**	Business, finance	Lifestyle, leisure	Science, technology	Sports	Num. of documents per site
BBC	102	68	75	202	447
The Guardian	67	59	116	96	338
Reuter	165	7	29	57	258
Num. of documents per topic	334	134	220	355	1043

Table 2: Details of news sites dataset

Feature extraction

In this study, it is assumed that each article is represented by two modalities: a) the textual description and b) the images. First, N-grams are extracted (globally and not per category) from the textual description. N-grams were chosen as the textual features because they were found as relatively easy to compute and effective for various classification tasks (for example (HaCohen-Kerner et al., 2008) and (Braga et al., 2009)). Then, the biggest image of each article was selected and visual features were extracted. In this case it was assumed that the biggest image was the representative one. At this point it should be noted that the possibility of the biggest image being a banner was checked by the following procedure: the length-to-height ratio was computed and if the value was disproportionally low or high, then it was considered to be a banner and was rejected. Otherwise, if the ratio value was balanced, it was considered to be an image. This is due to the structure of a banner (having a large length and a small height or vice versa). On the other hand, in most cases the dimensions of an image are close to a square.

[13] http://www.iptc.org/site/Home/

[14] The dataset is publicly available at: http://mklab.iti.gr/files/ArticlesNewsSitesData.7z

N-gram textual features

Given the fact that the concept extraction module (WP2) was not mature during this period in order to include a concept-based representation, we extracted N-gram features with the support of Yaakov HaCohen-Kerner from the Department of Computer Science, Jerusalem College of Technology – Lev Academic Centre, who is a member of the MULTISENSOR User Group.

For the extraction of the textual features from a news article web document, the following procedure was applied:

1. All appearances of 421 stopwords for general texts in English were deleted (Fox, 1989).
2. All possible continuous N-gram words (for N =1, 2, 3, 4) were created, provided that the all the words in a certain N-gram were in the same sentence.
3. The frequency of each N-gram feature in the corpora was counted.
4. The unigram, bigram, trigram and four-gram (each group alone) features were sorted in descending order.

To avoid unnecessarily large number of N-grams, only a subset of the most frequent features from each group was selected. More specifically, 195 of the most frequent N-gram features were selected as follows: a) 100 most frequent unigrams; b) 50 most frequent bigrams; c) 30 most frequent trigrams; d) 15 most frequent four-grams. The motivation for these numbers is as follows: The larger the value of N is, the smaller the number of relatively frequent N-grams in the corpus is. In this case, the reduction factor was determined to be approximately 2.

Visual features

The low-level visual features that were extracted in order to capture the characteristics of images are the MPEG-7 visual descriptors. The MPEG-7 standard specifies a set of descriptors, each defining the syntax and the semantics of an elementary visual low-level feature. Each descriptor aims at capturing different aspects of human perception (i.e., color, texture and shape). In this work, five MPEG-7 visual descriptors capturing color and texture aspects of human perception were extracted (Sikora, 2001):

1. **Color Layout Descriptor**: captures the spatial distribution of color or an arbitrary-shaped region.
2. **Color Structure Descriptor**: is based on color histograms, but aims at identifying localized color distributions.
3. **Scalable Color Descriptor**: is a Haar-transform based encoding scheme that measures color distribution over an entire image.
4. **Edge Histogram Descriptor**: captures the spatial distribution of edges and it involves division of image into 16 non-overlapping blocks. Edge information is then calculated for each block.
5. **Homogenous Texture Descriptor**: is based on a filter bank approach employing scale and orientation sensitive filters.

Then, an early fusion approach was applied, which involved the concatenation of all the aforementioned descriptors into a single feature vector. In this case, 320 visual features were extracted in total. The number of features/dimensions that were created from each descriptor are the following: a) Color Layout Descriptor: 18 features/dimensions; b) Color Structure Descriptor: 32 features/dimensions; c) Scalable Color Descriptor: 128 features/dimensions; d) Edge Histogram Descriptor: 80 features/dimensions; e) Homogeneous Texture Descriptor: 62 features/dimensions.

Experimental setup

For the experiments, the dataset was randomly split into training and test sets. Approximately 2/3 of the cases were kept for training purposes, whereas the rest (1/3) were used as test set, in order to estimate the classification scheme's performance.

Regarding the RF parameters that were used in the experiments, the following setting was applied: The number of trees for the construction of each RF was set based on the OOB error estimate. After conducting several experiments with a gradually increasing number of trees, the OOB error estimate was stabilized after using 1000 trees and no longer improved. Thus, the number of trees was set to *T*=1000. For each node split during the growing of a tree, the number of the subset of variables used to determine the best split was set to $p = \sqrt{\mathrm{k}}$ (according to (Breiman, 2001)), where *k* is the total number of features of the dataset. Specifically, in this study, taking into consideration the dimensionality of each modality, *p* was set to 14 for the textual modality, while for the visual modality *p* was set to 18.

Finally, for the evaluation of the performance of the proposed methodology, the precision, recall and F-score measures for each topic/category were computed, along with their corresponding macro-averaged values, as well as the accuracy on the entire test set (all categories included).

Results

The test set results from the application of RF to each modality separately are summarized in Table 3. We are mainly interested in the values of F-score, since it considers both precision and recall. We notice that the textual modality outperforms the visual in all measures, both regarding each topic and the macro-averaged scores. This indicates that textual data is a more reliable and solid source of information, in comparison to the visual data. More specifically:

- The RF trained with the textual data achieves a macro-averaged F-score value of 83.2%, compared to 45.5% for the visual modality
- The accuracy for the textual modality RF is 84.4%, while the visual modality RF achieves only 53%
- The worst results for the visual data RF are attained for the topics "Lifestyle-Leisure" (recall 12% and F-score 20.7%) and "Science-Technology" (precision 45.3%, recall 38.7% and F-score 41.7%). However, the results regarding the topic "Sports" are considered satisfactory. A possible explanation for this is the fact that the images from the "Lifestyle-Leisure" web pages depict diverse topics and therefore their visual appearance strongly varies. On the other hand, the images regarding the topic "Sports" contain rather specific information such as football stadiums (a characteristic example is depicted in Figure 15).

Modality / Topics	Textual			Visual		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Business-Finance	80.0%	87.3%	83.5%	56.3%	57.3%	56.8%
Lifestyle-Leisure	86.7%	78.0%	82.1%	75%	12%	20.7%
Science-Technology	79.1%	70.7%	74.6%	45.3%	38.7%	41.7%
Sports	91.3%	93.8%	92.5%	52.8%	76.8%	62.6%
Macro-average	**84.3%**	**82.5%**	**83.2%**	**57.4%**	**46.2%**	**45.5%**
Accuracy	**84.4%**			**53.0%**		

Table 3: Test set results from the application of RF to each modality

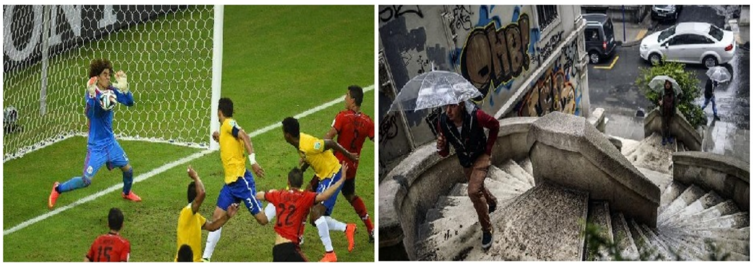


Figure 15: Characteristic image from a "Sports" web page (left)[15], along with an image regarding a web page from the "Lifestyle-Leisure" topic (right)[16].

In Table 4 the test set results from the application of the late fusion strategy to RF are provided, using the three different weighting methods described in Section 4.2.1 (OOB error/Proximity ratio/Adjusted proximity ratio). The weighting methods regarding the

[15] http://www.bbc.com/sport/0/football/27897075-"_75602744_ochoa.jpg"

[16] http://www.bbc.com/travel/feature/20140710-living-in-istanbul-"p022ktsw.jpg"

proximity ratio and the adjusted proximity ratio yield better performance results than the corresponding method for the OOB error. More specifically:

- The accuracy of Textual + Visual (Adjusted proximity ratio) is slightly better than the corresponding accuracy of Textual + Visual (Proximity ratio) (86.4% compared to 86.2%), while both of them are better than the accuracy of Textual + Visual (OOB error) (85.9%)
- The three weighted RFs achieve almost equal macro-averaged precision values (87.1% for Adjusted proximity ratio, 86.9% for OOB error and 86.8% for Proximity ratio), while regarding the macro-averaged F-score results, Textual + Visual (Adjusted proximity ratio) and Textual + Visual (Proximity ratio) both outperform Textual + Visual (OOB error) (85.4% for the Adjusted proximity ratio and 85.3% for the Proximity ratio to 84.3% for OOB error)

For comparison purposes, we also constructed a fused RF model, where equal weights were assigned to each modality. We note that after following this weighting approach (i.e. with equal weights), the performance of RF diminished in all aspects. More specifically, the macro-averaged F-score value dropped down to 78.9% and the accuracy value down to 80.4%. In Figure 16 the macro-averaged F-score values of all 6 RF models constructed in this study are sorted in ascending order. We observe that Textual + Visual (Adjusted proximity ratio) and Textual + Visual (Proximity ratio) are the best performing models among all cases.

The test set confusion matrices from all RF applications can be found in Appendix A.

Weighting method / **Topics**	**Textual + Visual** (Weighting based on OOB error per topic)			**Textual + Visual** (Weighting based on proximity ratio per topic)			**Textual + Visual** (Weighting based on adjusted proximity ratio per topic)		
	Prec.	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**	**Prec.**	**Rec.**	**F-score**
Business-Finance	80.3%	92.7%	86.1%	82.4%	89.1%	85.6%	82%	90.9%	86.2%
Lifestyle-Leisure	92.5%	74.0%	82.2%	92.9%	78.0%	84.8%	92.9%	78.0%	84.8%
Science-Technology	83.9%	69.3%	75.9%	81.4%	76.0%	78.6%	82.4%	74.7%	78.3%
Sports	90.7%	95.5%	93%	90.5%	93.8%	92.1%	91.3%	93.8%	92.5%
Macro-average	**86.9%**	**82.9%**	**84.3%**	**86.8%**	**84.2%**	**85.3%**	**87.1%**	**84.3%**	**85.4%**
Accuracy	**85.9%**			**86.2%**			**86.4%**		

Table 4: Test set results after the late fusion of RF regarding three different weighting schemes

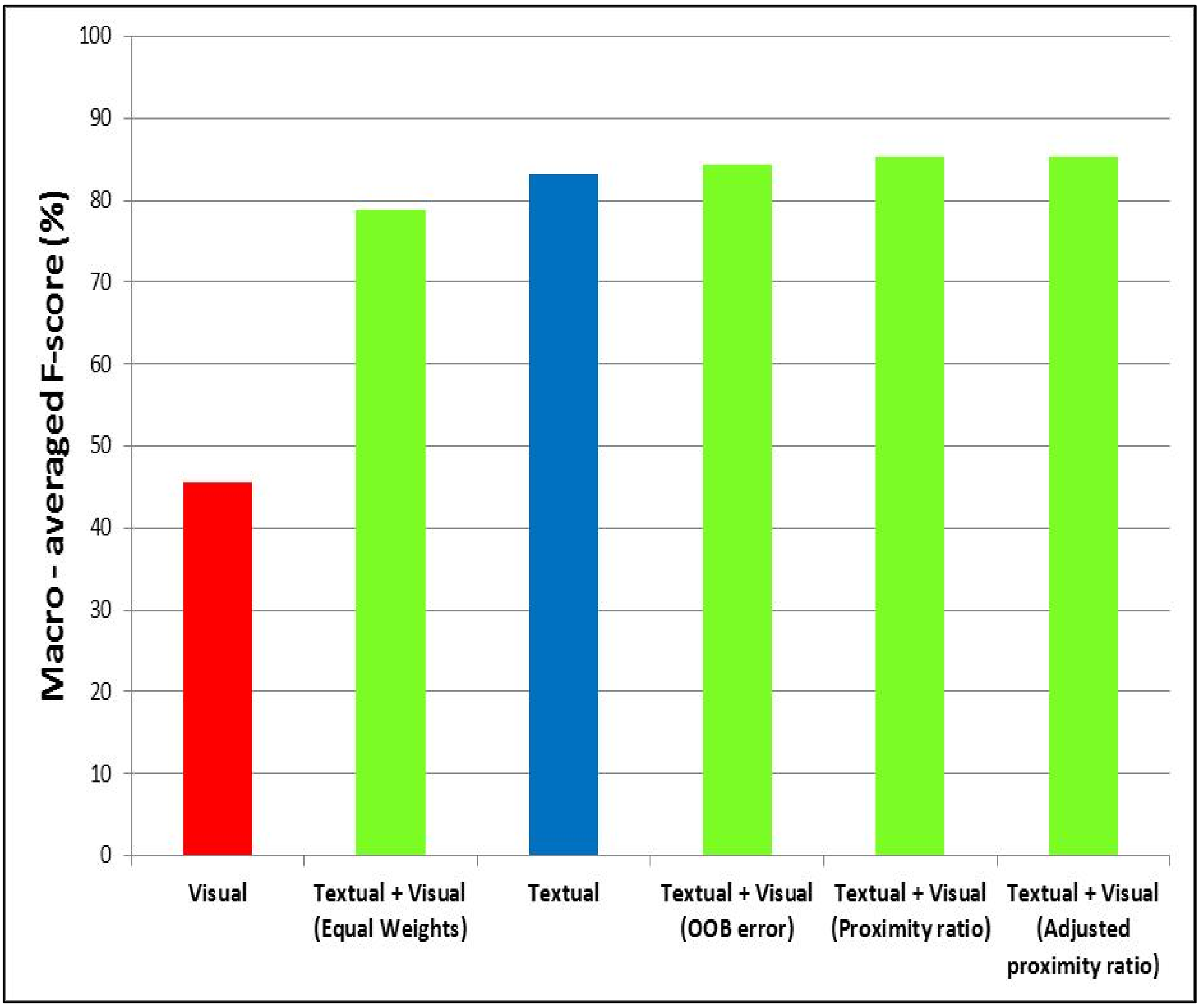


Figure 16: Macro-averaged F-score values for all RF models sorted in ascending order.

4.4.2 MULTISENSOR dataset

Set of topic categories - Dataset description

As already mentioned in Section 4.3, for the experiments within MULTISENSOR a set of topics for each use case was defined, with the help and guidance of user experts. The users that are interested in the topic classification are involved in the following two use cases: a) "home appliances" use case and b) journalistic use case which deals with energy-related News Items.

As far as the home appliances use case is concerned, the list of topics that are of interest is the following:

- **Technology**: Information on advancements on appliances or on modules used in appliances (e.g. http://www.ft.com/intl/cms/s/0/76edb496-a61a-11e3-8a2a-00144feab7de.html?siteedition=intl#axzz31nAFq531)
- **Sports**: Home appliance companies acting as sponsors for sports events or athletes (e.g. http://www.aargauerzeitung.ch/sport/ski/gebrochener-arm-geheilt-wendy-holdener-in-levi-am-start-127356261)

- **Economy&finance&business**: This topic would cover general corporate news, such as quarterly figures, appointment of managers, strategic topics and the like (e.g. http://www.nasdaq.com/article/whirlpool-plans-40-million-expansion-analyst-blog-cm336228#comment-log-in)
- **Arts&culture&entertainment**: (entertainment) Events or TV cooking shows might use specific appliances or be sponsored by appliance companies (e.g. http://beautifulkitchensblog.co.uk/2013/11/06/school-of-wok-top-wok-cooking-tips/)
- **Crime&law&justice**: (justice) This topic covers appliance companies that are involved in lawsuits, e.g. dealing with customer complaints or antitrust issues; maybe also regulation topics (e.g. http://www.dailyfinance.com/2014/02/24/class-action-lawsuit-get-boost-supreme-court/)
- **Disaster&accident**: (accident) Sometimes there is coverage of burning or exploding appliances (e.g. http://www.theguardian.com/money/2014/jan/11/kitchen-flames-domestic-appliances)
- **Lifestyle&leisure**: (lifestyle) This might be one of the most important topics: a lot of coverage is from interior design and women's magazines, promoting appliances as lifestyle and design elements. The targeted audience is the consumer himself, so the coverage has a huge importance as instrument for advertisement (e.g. http://www.bilanz.ch/bildergalerie/news-trends-herbst)
- **Nature&environment**: (environment) Sustainability is an important issue for appliance producers, as it is directly connected to their reputation (e.g. http://www.businesswire.co.uk/news/uk/20140326005584/en/Electrolux-unveils-climate-impact-target-2013-Sustainability)
- **Society**: (employment) Coverage of the companies as employers: this topic might overlap with economic and business coverage (e.g. http://www.bizjournals.com/albany/morning_call/2014/03/layoffs-at-general-electric-ny-factory-delayed.html)

The experiments conducted up to this point refer to the journalistic use case. This was decided since the data for the two use cases were similar. However, in order to have a complete understanding of the performance of our classification approach, further experimental studies on the "home appliances" use case dataset will be a target set for Deliverable 4.3. After receiving input from the user experts, the following set of topics was specified:

- **Economy&finance&business**: This topic would cover corporate and finance news related to energy sources (e.g. http://fuelfix.com/blog/2014/05/01/tesoros-crude-rail-plans-hit-delay-in-washington-state/)
- **Health**: This topic includes news about energy applications to the health sector (e.g. http://scienceblog.com/72122/acupuncture-helps-kids-manage-pain-nausea/)

- **Lifestyle&leisure**: (lifestyle) News about how energy-related innovations could be used for lifestyle and leisure purposes (e.g. http://www.washingtonpost.com/realestate/serving-as-his-own-general-contractor-an-arlington-man-transforms-his-home/2014/03/13/fe714e2c-7a3d-11e3-af7f-13bf0e9965f6_story.html)
- **Nature&environment**: (environment) This topic could cover news, such as the impact energy consumption has on the environment (e.g. http://www.bio-medicine.org/biology-news-1/John-P--Holdren-addresses-climate-change--stressing-need-for-international-cooperation-15601-1/)
- **Politics**: News about energy-related policies, initiatives and legislations (e.g. http://cleantechnica.com/2014/06/23/obamas-clean-power-plan-without-policy-mistakes-made-eu/)
- **Science&Technology**: (technology) This topic could cover news about scientific research and technological advances in the energy sector (e.g. http://www.bio-medicine.org/biology-news-1/The-JBEI-GT-Collection-3A-A-new-resource-for-advanced-biofuels-research-36420-1/)

The dataset[17] used in these experiments contains 2382 news articles retrieved from the MULTISENSOR News Repository for the aforementioned set of topics. The manual annotation of the news articles was necessary, since no description about the topic of each news article was provided in the News Repository. The numbers of news articles for each topic that are contained in the dataset are listed below:

- **Economy&finance&business**: 465 news articles
- **Health**: 187 news articles
- **Lifestyle&leisure**: 326 news articles
- **Nature&environment**: 447 news articles
- **Politics**: 277 news articles
- **Science&Technology**: 680 news articles

Feature extraction

For the feature extraction process, at first all the modalities created in WP2 (textual and visual features) and WP3 (sentiment and contextual information) were considered. However, due to the nature and content of the sentiment and contextual information, the features extracted for these two modalities proved non-representative and unsuitable for the classification task at hand. Therefore, we decided to represent each news article by the two modalities from WP2: a) the textual description and b) the images. Again, as in the case of the News sites dataset experiments, N-gram features were extracted (globally and not per category) from the textual description. Regarding the visual features, they were extracted from the biggest image of each article, as it was assumed to be the representative one.

[17] The dataset is publicly available at: http://mklab.iti.gr/files/ArticlesNewsSitesData_2382.7z

N-gram textual features

The same procedure with the one described in the News sites dataset experiments Section was applied for the N-gram textual features extraction. Again, four groups of N-grams were created (unigrams, bigrams, trigrams and four-grams). The number of features for each group is the following: a) 1000 unigrams; b) 200 bigrams; c) 40 trigrams; d) 8 four-grams (a reduction factor of 5 was applied for each group).

Visual features

The low-level visual feature that was used for capturing the characteristics of images is the RGB-SIFT (Van De Sande et al., 2010) visual descriptor, which is an extension of the SIFT. In general, SIFT descriptors belongs to the category of local descriptors that represent local salient points and thus capture the characteristics of the interest points (or keypoint) of images. When local descriptors are used, the first is the identification of interest points. Then around each keypoint a 16x16 neighbourhood is retrieved and it is divided into 16 sub-blocks of size 4x4. Moreover, for each sub-block, a 8 bin orientation histogram is created, so a total of 128 bin values are available. For each keypoint only the pixel intensity of it is considered while the colour information is dropped. On the other hand, RGB-SIFT is an extension of the SIFT descriptor that considers apart from the pixel intensity, the colour itself in the three channels Red, Green, Blue for each interest point. Thus it captures more information and is able to represent better the image compared to SIFT. However, when local descriptors are employed, and given that the whole procedure is arduous (in terms of time and CPU processing), a visual word assignment step is applied after the feature extraction step. Specifically, we apply K-Means clustering on these features vectors produced in order to acquire the visual vocabulary and finally VLAD encoding is realized for representing images (Jégou et al., 2010). Eventually, a descriptor is produced that gives an overall impression of the visual data. In this case, the dimensionality of the visual features set is 4000.

Experimental setup

Again, as in the News sites dataset experiments, the dataset was randomly split into training and test sets. Approximately 2/3 of the cases were kept for training purposes, whereas the rest (1/3) were used as test set, in order to estimate the classification scheme's performance.

The RF parameter setting that was selected is the following: The number of trees for the construction of each RF was set based on the OOB error estimate. After conducting several experiments with a gradually increasing number of trees, the OOB error estimate was stabilized after using 2000 trees and no longer improved. Thus, the number of trees was set to *T*=2000. For each node split during the growing of a tree, the number of the subset of variables used to determine the best split was set to $p = \sqrt{\mathrm{k}}$ (according to (Breiman, 2001)), where *k* is the total number of features of the dataset.

Finally, for the evaluation of the experiments, the precision, recall and F-score measures for each topic/category were computed, along with their corresponding macro-averaged values, as well as the accuracy on the entire test set (all categories included).

Results

Table 5 contains the test set results from the application of RF to each modality separately. We observe that the textual modality yields better performance than the visual one in all measures, both regarding each topic and the macro-averaged scores. This confirms the notion that the textual features are a more reliable source of information and more suitable for the classification task, compared to the visual features. Specifically:

- The textual RF achieves a macro-averaged F-score value of 78.2%, compared to 52.9% for the visual modality
- The accuracy for the textual modality's RF model is 77.6%, while the visual modality's RF model achieves only 54.9%
- The worst results for the textual RF model are attained for the topic "Nature-Environment" (recall 49% and F-score 59.8%). On the other hand, the best results for the visual RF model (in terms of F-score value) concern the topic "Lifestyle-Leisure" (60.2%)

Modality / Topics	Textual			Visual		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score
Economy-Business-Finance	67.9%	93.1%	78.5%	56.2%	62.9%	59.4%
Health	91.8%	80.4%	85.7%	95%	33.9%	50%
Lifestyle-Leisure	82.2%	95.1%	88.2%	54.3%	67.6%	60.2%
Nature-Environment	76.8%	49%	59.8%	56.2%	38.1%	45.4%
Politics	87.1%	69.3%	77.2%	65.2%	34.1%	44.7%
Science-Technology	78.8%	80.8%	79.8%	50%	67.9%	57.6%
Macro-Average	**80.8%**	**78%**	**78.2%**	**62.8%**	**50.8%**	**52.9%**
Accuracy	**77.6%**			**54.9%**		

Table 5: Test set results from RF application to each modality

In Table 6, the results from the late fusion of the two RF models, using the OOB error estimate weighting strategy are presented. Moreover, the weight values assigned to each modality's RF model for each class are depicted in the right side of the Table. It is obvious that this approach does not improve the overall performance significantly, as we have a 0.4% improvement in the accuracy value (78% compared to 77.6% for the textual RF model) and a 0.6% improvement in the macro-averaged F-score value (78.8% for the fused model and 78.2% for the textual modality).

Weighting method	**Textual + Visual** (Weighting based on OOB error per topic)			**Weight values assigned to each modality**	
Topics	**Prec.**	**Rec.**	**F-score**	**Textual**	**Visual**
Economy-Business-Finance	79.6%	90.6%	84.7%	0.64	0.36
Health	97.6%	71.4%	82.5%	0.75	0.25
Lifestyle-Leisure	91.3%	92.2%	91.7%	0.65	0.35
Nature-Environment	71.6%	50.3%	59.1%	0.6	0.4
Politics	87.5%	71.6%	78.7%	0.66	0.34
Science-Technology	69.4%	85.5%	76.6%	0.55	0.45
Macro-average	**82.8%**	**76.9%**	**78.8%**		
Accuracy	**78%**				

Table 6: Test set results after the late fusion of RF regarding the OOB error estimate weighting scheme

In Table 7, the results from the late fusion of the two RF models, using the proximity ratio weighting approach, together with the weights for each class and for each modality are contained. In this case, we notice that the performance of the fused RF model diminishes for all measures. More specifically, the accuracy value has dropped down to 73.7% and the macro-averaged F-score value down to 71.8% (a 7% loss compared to the OOB weighting strategy). This could be attributed to the fact that the visual modality gets higher weight values than the textual modality for the "Health" and "Nature-Environment" topics (due to higher averaged proximity ratio values, something that is not justified by the visual modality's OOB performance for these topics).

Weighting method / Topics	Textual + Visual (Weighting based on proximity ratio per topic)			Weight values assigned to each modality	
	Prec.	**Rec.**	**F-score**	**Textual**	**Visual**
Economy-Business-Finance	76.5%	88.1%	81.9%	0.61	0.39
Health	100%	37.5%	54.5%	0.44	0.56
Lifestyle-Leisure	90.9%	88.2%	89.5%	0.65	0.35
Nature-Environment	63.8%	47.7%	54.5%	0.43	0.57
Politics	84.7%	69.3%	76.2%	0.6	0.4
Science-Technology	65.7%	85%	74.1%	0.64	0.36
Macro-average	**80.3%**	**69.3%**	**71.8%**		
Accuracy	**73.7%**				

Table 7: Test set results after the late fusion of RF regarding the proximity ratio weighting scheme

If we opt to adjust the proximity ratio values (following the approach described in Section 4.2.1), we notice that there is a notable improvement in the fused results (Table 8). This time, the modality weights are assigned not only based on the proximity ratio values, but also based on the OOB accuracy performance for each class. This leads to an overall best performance among all weighting strategies (accuracy value 79.2%, macro-averaged precision 83.8% - recall 78.4% - F-score 80.1%).

Weighting method / Topics	Textual + Visual (Weighting based on adjusted proximity ratio per topic)			Weight values assigned to each modality	
	Prec.	**Rec.**	**F-score**	**Textual**	**Visual**
Economy-	79%	92.5%	85.2%	0.73	0.27

Business-Finance					
Health	100%	73.2%	84.5%	0.71	0.29
Lifestyle-Leisure	90.7%	95.1%	92.8%	0.77	0.23
Nature-Environment	72.4%	51%	59.8%	0.53	0.47
Politics	88.9%	72.7%	80%	0.75	0.25
Science-Technology	72%	85.9%	78.3%	0.67	0.33
Macro-average	**83.8%**	**78.4%**	**80.1%**		
Accuracy	**79.2%**				

Table 8: Test set results after the late fusion of RF regarding the adjusted proximity ratio weighting scheme

Finally, for comparison purposes, in Table 9 we include the late fusion results in the case of assigning equal weights for each modality. We can say that after following this approach, the performance of RF is greatly affected in a negative way, as there is a decrease in all performance measures, both for each topic and for their macro-averaged values.

The test set confusion matrices from all RF models can be found in Appendix A.

Weighting method / **Topics**	**Textual + Visual** (Equal weights per topic)			**Weight values assigned to each modality**	
	Prec.	**Rec.**	**F-score**	**Textual**	**Visual**
Economy-Business-Finance	74.4%	74.8%	74.6%	0.5	0.5
Health	100%	35.7%	52.6%	0.5	0.5
Lifestyle-Leisure	87.5%	82.4%	84.9%	0.5	0.5
Nature-	63.6%	49.7%	55.8%	0.5	0.5

Environment					
Politics	77.8%	63.6%	69.9%	0.5	0.5
Science-Technology	60.6%	84.2%	70.4%	0.5	0.5
Macro-average	**77.3%**	**65.1%**	**68%**		
Accuracy	**69.6%**				

Table 9: Test set results after the late fusion of RF (Equal weights considered)

5 CONCLUSIONS

In this Deliverable, we have presented the basic techniques for topic-based classification using a supervised learning approach, on top of a content representation framework. In this context, we have proposed a multimedia data representation framework that supports multimedia indexing and retrieval. Specifically, the model integrates in a unified manner the representation of multimedia and social features in online environments. Its flexibility and expressive power allow it to embrace the heterogeneity of multimedia content and its interconnections, thus making it unique in its ability to support a wide range of multimedia information processing, analysis, and access applications. Our aim is for SIMMO to be a reusable data model across such applications; to facilitate its adoption, we plan to extend its documentation, add utility methods (such as the implementation of standard indexing and retrieval operations), and identify and implement mappings to established data models (such as SIOC).

As far as the topic-based classification is concerned, it uses the data captured by SIMMO model for labelling the News Items with a predefined list of topic categories. The main findings from the experiments are the following: a) The textual modality is more reliable and suitable for the classification task than the visual modality; b) The late fusion approach and the selection of the proper weighting strategy improve the outputs from the separate RF models constructed from each modality.

Future work, which will be reported in D4.3, includes the full implementation of SIMMO, the design of a database for capturing the SIMMO objects and the implementation of a set of functions for inserting and retrieving records from it. It should be noted that different retrieval functions will be developed in order to cover the user requirements for data clustering, similarity search and other retrieval functionalities. As far as the topic-based classification techniques are concerned, the next step involves the combination of the proposed technique with rule-based approaches that will consider the relations among classes as well. Moreover, experiments by considering concept-based information (textual and visual) extracted in WP2 will be conducted. Finally, D4.3 will include the techniques for topic detection based on multimodal clustering exploiting the multimodal features of SIMMO (context, sentiment, spatiotemporal information).

6 REFERENCES

Arapakis, I., Cambazoglu, B. B., and Lalmas, M. 2014. “On the Feasibility of Predicting News Popularity at Cold Start”, In Proceedings of the 6th International Conference on Social Informatics (Socinfo 2014), Barcelona, 10-13 November 2014 (to appear).

Aung, W. T., and Hla, K. H. M. S. 2009. “Random forest classifier for multi-category classification of web pages”, In Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific, pp. 372-376.

Bandari, R., Asur, S., and Huberman, B. A. 2012. “The Pulse of News in Social Media: Forecasting Popularity”, In ICWSM.

Bay, H., Tuytelaars, T., and Gool, L. V. 2006. "SURF: Speeded Up Robust Features", Proceedings of the ninth European Conference on Computer Vision, May 2006.

Bojars, U., Breslin, J. G., Peristeras, V., Tummarello G., Decker, S. 2008. “Interlinking the social web with semantics”, IEEE Intelligent Systems, vol. 23(3), pp. 29-40.

Braga, I., Monard, M., and Matsubara, E. 2009. “Combining unigrams and bigrams in semi-supervised text classification”, In Proceedings of Progress in Artificial Intelligence, 14th Portuguese Conference on Artificial Intelligence (EPIA 2009), Aveiro, pp. 489-500.

Breiman, L. 2001. “Random Forests”, In Machine Learning, 45(1), pp. 5-32.

Brickley, D., Miller, L. 2012. “FOAF vocabulary specification 0.98”, Namespace Document, 9.

Burnett, I. S., de Walle, R. V., Hill, K., Bormans, J., Pereira, F. 2003. “MPEG-21: Goals and achievements”, IEEE MultiMedia, vol. 10(4), pp. 60-70.

Caetano, A., Guimaraes, N. 1998. “A model for content representation of multimedia information”, In Proceedings of the 1st workshop on the Challenge of Image Retrieval (CIR 1998), organised by the British Computer Society.

Chang, C. C., and Lin, C. J. 2001. “LIBSVM: a library for support vector machines”.

Chang, S.-F., Sikora, T., Purl, A. 2001. “Overview of the MPEG-7 standard”, IEEE Transactions on Circuits and Systems for Video Technology, vol. 11(6), pp. 688-695.

Chen, M. Y., and Hauptmann, A. 2004. “Multi-modal classification in digital news libraries”, Proceedings of the 2004 Joint ACM/IEEE Conference on, pp. 212-213.

Chen, N., Shatkay, H., and Blostein, D. 2006. “Exploring a new space of features for document classification: figure clustering”, In Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research, pp. 35, IBM Corp.

Daras, P., Axenopoulos, A., Darlagiannis, V., Tzovaras, D., Bourdon, X. L., Joyeux, L., Verroust-Blondet, A., Croce, V., Steiner, T., Massari, A., Camurri, A., Morin, S., Mezaour, A.-D., Sutton, L. F., Spiller, S. 2011. “Introducing a unified framework for content object description”, International Journal of Multimedia Intelligence and Security, vol. 2(3), pp. 351-375.

Erdélyi, M., Garzó, A., and Benczúr, A. A. 2011. “Web spam classification: a few features worth more”, In Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality, pp. 27-34, ACM.

Fox, C. 1989. “A stop list for general text”, ACM SIGIR Forum, 24 (1-2), ACM.

Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., and König, A. C. 2008. “BLEWS: Using Blogs to Provide Context for News Articles”, In ICWSM.

Giroux, P., Brunessaux, S., Brunessaux, S., Doucy, J., Dupont, G., Grilheres, B., Mombrun, Y., Saval, A., des Portes, P. d. 2008. “Weblab: An integration infrastructure to ease the development of multimedia processing applications”, In Proceedings of the International Conference on Software and System Engineering and their Applications (ICSSEA).

Glodek, M., Scherer, S., Schwenker, F., and Palm, G. 2011. “Conditioned Hidden Markov Model Fusion for Multimodal Classification”, In INTERSPEECH, pp. 2269-2272.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., and Rueckert, D. 2013. “Random forest-based similarity measures for multi-modal classification of Alzheimer's disease”, NeuroImage, 65, pp. 167-175.

HaCohen-Kerner, Y., Mughaz, D., Beck, H., and Yehudai, E. 2008. “Words as Classifiers of Documents According to their Historical Period and the Ethnic Origin of their Authors”, Cybernetics and Systems: An International Journal, 39(3), pp. 213-228.

Ho, A. K. N., Ragot, N., Ramel, J. Y., Eglin, V., and Sidere, N. 2013. “Document Classification in a Non-stationary Environment: A One-Class SVM Approach”, In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, (ICDAR), pp. 616-620.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. 2010. “Aggregating local descriptors into a compact image representation”, In IEEE Conference on Computer Vision & Pattern Recognition, pp. 3304-3311.

Klassen, M., and Paturi, N. 2010. “Web document classification by keywords using random forests”, In Networked Digital Technologies, pp. 256-261, Springer Berlin Heidelberg.

Lazebnik, S., Schmid, C., and Ponce, J. 2006. “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories”, In Proc. of IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 2169-2178.

Li, W., and Meng, Y. 2013. “Improving the performance of neural networks with random forest in detecting network intrusions”, In Advances in Neural Networks–ISNN 2013, pp. 622-629, Springer Berlin Heidelberg.

Lin, K. H. Y., Yang, C., and Chen, H. H. 2008. “Emotion classification of online news articles from the reader's perspective”, In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pp. 220-226, IEEE Computer Society.

Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., and Kompatsiaris, I. 2014. “News Articles Classification Using Random Forests and Weighted Multimodal Features”, In Multidisciplinary Information Retrieval, pp. 63-75, Springer International Publishing.

Lowe, D. G. 2004. “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision, 60(2), pp. 91-110.

Montagnuolo, M., and Messina, A. 2009. “Parallel neural networks for multimodal video genre classification”, Multimedia Tools and Applications, 41 (1), pp. 125-159.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. 2011. “Multimodal deep learning”, In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 689-696.

Papadopoulos, S., Kompatsiaris, Y. 2014. “Social multimedia crawling for mining and search”, IEEE Computer, vol. 47(5), pp. 84-87.

Qi, X., and Davison, B. D. 2009. “Web page classification: Features and algorithms”, ACM Computing Surveys (CSUR), 41 (2), 12.

Ramzan, N., van Zwol, R., Lee, J.-S., Clüver, K., Hua, X.-S. (eds.) 2013. “Social Media Retrieval”, Computer Communications and Networks, Springer.

Robnik-Šikonja, M. 2004. “Improving random forests”, In Machine Learning: ECML 2004, pp. 359-370, Springer Berlin Heidelberg.

Saarikoski, J., Laurikkala, J., Järvelin, K., and Juhola, M. 2011. “Self-Organising maps in document classification: A comparison with six machine learning methods”, In Adaptive and Natural Computing Algorithms, pp. 260-269, Springer Berlin Heidelberg.

Sebastiani, F. 2002. “Machine learning in automated text categorization”, ACM Computing Surveys, 34 (1), pp. 1-47.

Selamat, A., and Omatu, S. 2004. “Web page feature selection and classification using neural networks”, Information Sciences, 158, pp. 69-88.

Shi, L., Ma, X., Xi, L., Duan, Q., and Zhao, J. 2011. “Rough set and ensemble learning based semi-supervised algorithm for text classification”, Expert Systems with Applications, 38 (5), pp. 6300-6306.

Shin, C., Doermann, D., and Rosenfeld, A. 2001. “Classification of document pages using structure-based features”, International Journal on Document Analysis and Recognition, 3 (4), pp. 232-247.

Sikora, T. 2001. “The MPEG-7 visual standard for content description-an overview”, IEEE Transactions on Circuits and Systems for Video Technology, 11(6), pp. 696-702.

Srivastava, N., and Salakhutdinov, R. 2012. “Multimodal learning with deep boltzmann machines”, In Advances in neural information processing systems, pp. 2222-2230.

Swezey, R. M., Sano, H., Shiramatsu, S., Ozono, T., and Shintani, T. 2012. “Automatic detection of news articles of interest to regional communities”, IJCSNS, 12 (6), 100.

Ting, S. L., Ip, W. H., and Tsang, A. H. C. 2011. “Is Naive Bayes a good classifier for document classification?”, International Journal of Software Engineering and Its Applications, 5 (3), 37.

Tsikrika, T., Andreadou, K., Moumtzidou, A., Schinas, E., Papadopoulos, S., Vrochidis, S., Kompatsiaris, Y. 2015. "A Unified Model for Socially Interconnected Multimedia-Enriched Objects", 21st MultiMedia Modelling Conference (MMM2015), Sydney, Australia.

Van De Sande, K. E., Gevers, T., and Snoek, C. G. 2010. “Evaluating color descriptors for object and scene recognition”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9), pp. 1582-1596.

Xu, B., Ye, Y., and Nie, L. 2012. "An improved random forest classifier for image classification", In Information and Automation (ICIA), 2012 International Conference on, pp. 795-800, IEEE.

Xu, Z., Yan, F., Qin, J., and Zhu, H. 2011. "A Web Page Classification Algorithm Based on Link Information", In Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2011 Tenth International Symposium on, pp. 82-86.

Zeng, A., and Huang, Y. 2012. "A text classification algorithm based on rocchio and hierarchical clustering", In Advanced Intelligent Computing, pp. 432-439, Springer Berlin Heidelberg.

Zhou, Q., Hong, W., Luo, L., and Yang, F. 2010. "Gene selection using random forest and proximity differences criterion on DNA microarray data", Journal of Convergence Information Technology, 5(6), pp. 161-170.

A Appendix

A.1 Confusion matrices

A.1.1 News sites dataset

Predicted / Observed	Business-Finance	Lifestyle-Leisure	Science-Technology	Sports
Business-Finance	96	3	9	2
Lifestyle-Leisure	6	39	2	3
Science-Technology	15	2	53	5
Sports	3	1	3	105

Table A1: Test set confusion matrix (textual RF model).

Predicted / Observed	Business-Finance	Lifestyle-Leisure	Science-Technology	Sports
Business-Finance	63	1	18	28
Lifestyle-Leisure	12	6	10	22
Science-Technology	18	1	29	27
Sports	19	0	7	86

Table A2: Test set confusion matrix (visual RF model).

Predicted Observed	Business-Finance	Lifestyle-Leisure	Science-Technology	Sports
Business-Finance	102	1	5	2
Lifestyle-Leisure	7	37	3	3
Science-Technology	16	1	52	6
Sports	2	1	2	107

Table A3: Test set confusion matrix (RF late fusion – Weighting based on OOB error per topic).

Predicted Observed	Business-Finance	Lifestyle-Leisure	Science-Technology	Sports
Business-Finance	98	1	8	3
Lifestyle-Leisure	6	39	2	3
Science-Technology	12	1	57	5
Sports	3	1	3	105

Table A4: Test set confusion matrix (RF late fusion – Weighting based on proximity ratio per topic).

Predicted / Observed	Business-Finance	Lifestyle-Leisure	Science-Technology	Sports
Business-Finance	100	1	7	2
Lifestyle-Leisure	6	39	2	3
Science-Technology	13	1	56	5
Sports	3	1	3	105

Table A5: Test set confusion matrix (RF late fusion – Weighting based on adjusted proximity ratio per topic).

Predicted / Observed	Business-Finance	Lifestyle-Leisure	Science-Technology	Sports
Business-Finance	101	1	5	3
Lifestyle-Leisure	11	32	3	4
Science-Technology	18	1	49	7
Sports	12	1	2	97

Table A6: Test set confusion matrix (RF late fusion – Equal weights per topic).

A.1.2 MULTISENSOR dataset

Predicted **Observed**	Economy-Business-Finance	Health	Lifestyle-Leisure	Nature-Environment	Politics	Science-Technology
Economy-Business-Finance	148	0	3	3	1	4
Health	0	45	6	0	0	5
Lifestyle-Leisure	2	0	97	2	0	1
Nature-Environment	32	0	3	76	7	37
Politics	13	0	1	9	61	4
Science-Technology	23	4	8	9	1	189

Table A7: Test set confusion matrix (textual RF model).

Predicted **Observed**	Economy-Business-Finance	Health	Lifestyle-Leisure	Nature-Environment	Politics	Science-Technology
Economy-Business-Finance	100	0	6	12	6	35
Health	2	19	0	2	2	31
Lifestyle-Leisure	10	0	69	6	1	16
Nature-Environment	15	1	18	59	6	56

Politics	23	0	10	4	30	21
Science-Technology	28	0	24	22	1	159

Table A8: Test set confusion matrix (visual RF model).

Predicted Observed	Economy-Business-Finance	Health	Lifestyle-Leisure	Nature-Environment	Politics	Science-Technology
Economy-Business-Finance	144	0	1	6	3	5
Health	0	40	3	1	0	12
Lifestyle-Leisure	3	0	94	1	0	4
Nature-Environment	14	0	2	78	6	55
Politics	9	0	0	4	63	12
Science-Technology	11	1	3	19	0	200

Table A9: Test set confusion matrix (RF late fusion – Weighting based on OOB error per topic).

Predicted Observed	Economy-Business-Finance	Health	Lifestyle-Leisure	Nature-Environment	Politics	Science-Technology
Economy-Business-Finance	140	0	1	11	3	4
Health	1	21	2	2	2	28

Lifestyle-Leisure	4	0	90	4	0	4
Nature-Environment	16	0	3	74	6	56
Politics	10	0	0	5	61	12
Science-Technology	12	0	3	20	0	199

Table A10: Test set confusion matrix (RF late fusion – Weighting based on proximity ratio per topic)

Predicted Observed	Economy-Business-Finance	Health	Lifestyle-Leisure	Nature-Environment	Politics	Science-Technology
Economy-Business-Finance	147	0	1	6	2	3
Health	0	41	4	1	0	10
Lifestyle-Leisure	2	0	97	1	0	2
Nature-Environment	17	0	2	79	5	52
Politics	9	0	0	4	64	11
Science-Technology	11	0	3	18	1	201

Table A11: Test set confusion matrix (RF late fusion – Weighting based on adjusted proximity ratio per topic).

Predicted / **Observed**	Economy-Business-Finance	Health	Lifestyle-Leisure	Nature-Environment	Politics	Science-Technology
Economy-Business-Finance	119	0	1	12	6	21
Health	1	20	2	2	2	29
Lifestyle-Leisure	5	0	84	4	1	8
Nature-Environment	11	0	5	77	6	56
Politics	12	0	1	5	56	14
Science-Technology	12	0	3	21	1	197

Table A12: Test set confusion matrix (RF late fusion – Equal weights per topic).