**xLiMe**
**FP7-ICT-2013-10**
**ICT-2013.4.1**
**Contract no.: 611346**
www.xlime.eu

**Deliverable D5.4**

# Recommendation Prototype

| | |
|---|---|
| Editor: | Aditya Mogadala, KIT |
| Author(s): | Aditya Mogadala, KIT |
| Deliverable Nature: | Prototype (P) |
| Dissemination Level: | Public (PU) |
| Contractual Delivery Date: | M30 – 30 April 2016 |
| Actual Delivery Date: | M30 – 30 April 2016 |
| Suggested Readers: | Researchers and developers who are interested in providing content based recommendations in web shops based on social media content. |
| Version: | 1.0 |
| Keywords: | Content based recommendations; semantic product vectors; text annotation |

Disclaimer



| | |
|---|---|
| Full Project Title: | xLiMe– crossLingual crossMedia knowledge extraction |
| Short Project Title: | xLiMe |
| Number and<br>Title of Work package: | WP5<br>Cross-lingual, Cross-media Analytics and Reporting |
| Document Title: | D5.4 – Recommendation Prototype |
| Editor: | Aditya Mogadala, KIT |
| Work package Leader: | Blaz Novak, JSI |

# Executive Summary

The goal of this deliverable is to provide a detailed description of the research and development done to build a recommendation prototype. The aim of the prototype is to provide content based recommendations in web shops based on the mentions of products in social media by varied users. This prototype also leverages its solution to the challenges posed in a use case, where the customers of ECONDA i.e. web shops analyze social media content to understand the need of their customers.

Our approach to content based recommendations is built on the ideas of previous deliverables, mainly semantic disambiguation prototype (D4.2) and text annotation (D3.3.2). Here, we explore two different approaches and apply the best suitable approach for the use-case. In the first approach popular brands, products and categories of a web shop are identified in the social media such as Twitter messages using word distributed representations (i.e. embeddings) learned over a large corpus of social media content. While in the second approach, popular products, brands and categories are identified in the social media using information from a knowledge-base.

# Table of Contents



# List of Figures

# List of Tables



# Abbreviations

SPV Semantic Product Vectors

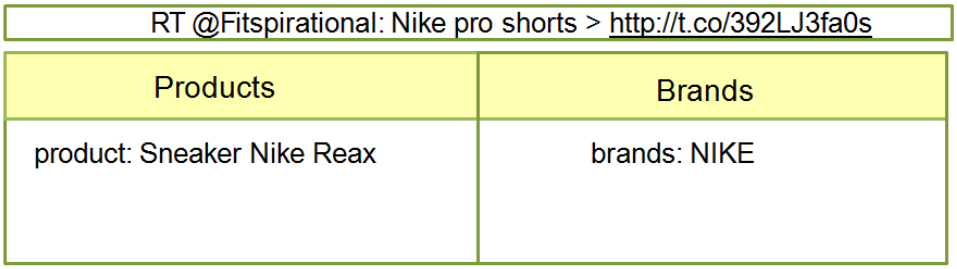
# Definitions

| | |
|---|---|
| Unstructured Corpora | Also considered as a document corpus without any structured metadata or any link to knowledge base. |
| Twitter Messages | Can be also called as microblogs tweets that are produced by Twitter (a social networking website). |

# 1 Introduction

Growth of social media content on the web has provided numerous opportunities. It has become an important medium to collect user opinions on various issues. Several applications that could collect user opinions have shown valuable insights in many domains. In our recommendation prototype, we also aim to leverage the user's interest about certain entities that could improve or diversify recommendations. In particular, we aim to improve web shop recommendations by identifying products/brands/categories in a web shop that are mentioned in the social media posts. This will help us to comprehend trending or popular products/brands/categories in which users are interested. Although, a caveat here is that we do not have an idea about user's sentiment about the product/brand. Nevertheless, as of now we limit ourselves to content based recommendations where we identify products/brands/categories in the social media that are same or similar to the products/brands/categories in the web shop catalogue to be further used as recommendation to the customers as the popular items. Figure 1 shows the overall idea with an example.

**Figure 1 Product and Brand Annotation for a Twitter Message (Example).**




To achieve our goal, we explore two different approaches inspired from the semantic disambiguation prototype [1] and text annotation [2]. Both of these approaches are unsupervised and are listed below.

1. The first approach is based on unstructured corpus. Vocabulary from social media corpus is extracted and represented with high-dimensional vectors. We name this approach as semantic product vectors (SPV).
2. The second approach is based on text annotations derived from a knowledge-base. We name this approach as text annotation approach.

Table 1 and Table 2 below provide the comparative analysis of the two approaches.

**Table 1 Pros of Text Annotations and Semantic Product Vectors**

| Pros | Text Annotations | Semantic Product Vectors (SPV) |
|---|---|---|
| 1 | Either general or domain specific knowledge-base can be used to map products to the world knowledge. | Any unstructured social media corpus can be used to learn high-dimensional vectors for products/brands/categories. (E.g. Twitter, Facebook posts etc...) |
| 2 | Can be effective to find similar products if the knowledge base specific features such as relational knowledge is leveraged. | Computationally less expensive to find similar products as compared to graph based methods. |
| 3 | Encode Domain knowledge. | Can encode product specific features that are inherent in the social media posts. |

**Table 2 Cons of Text Annotations and Semantic Product Vectors**

| Cons | Text Annotations | Semantic Product Vectors (SPV) |
|---|---|---|
| 1 | Cannot work with different social corpus that store different semantics. (E.g. Twitter, social media data etc…) | Cannot handle relations. |
| 2 | Computationally expensive due to the usage of different tools in a pipeline. | No world knowledge. Only capture semantics from the given corpus. |
| 4 | Problematic if the product is not present in the knowledge-base. | Different variations of the same product may have different semantic vectors. |

In the following sections, we briefly describe the data format and provide an in-depth description of the above mentioned approaches.

## 1.1 Data

Most of the data used for building the recommendation prototype is collected from external social media resources like Twitter [3], Vimeo [4], YouTube [5] etc., provided by VICO [6]. Data generated by VICO is in the format of xLiMe meta data model [7].

The xLiMe meta data model wraps the content generated from social media sites into Turtle format. We convert the data into the JSON format as it better suites our proposed approaches. A simple example of both of these formats collected from Vimeo is listed in the Table 3 and Table 4.

**Table 3 Vimeo messages in Turtle Format**

```
<http://vico-research.com/social/11cd99d9-99b7-3a73-b7a1-449d8a63b035>
        a                   sioc:MicroPost ;
        dcterms:created         "2016-04-13T13:34:04"^^xsd:dateTime ;
        dcterms:language        "de" ;
        dcterms:publisher       <https://vimeo.com/> ;
        dcterms:source          <https://vimeo.com/162669825> ;
        dcterms:spatial         [ rdfs:label  "" ] ;
        sioc:content            "Ein Paar spielt Tennis.Kinder laufen immer wieder über das Spielfeld, behindern
das Spiel aber nicht weiter." ;
        sioc:has_creator        <https://vimeo.com/user12835962> ;
        xlime:hasAnnotation     [ xlime:hasConfidence  "0.703"^^xsd:double ;
                                  xlime:hasEntity     dbpedia:Paar ;
                                  xlime:hasPosition    [ xlime:hasStartPosition  "4"^^xsd:long ;
                                                         xlime:hasStopPosition   "8"^^xsd:long
                                                       ]
                                ] ;
        xlime:hasAnnotation     [ xlime:hasConfidence  "0.798"^^xsd:double ;
                                  xlime:hasEntity     dbpedia:Game ;
                                  xlime:hasPosition    [ xlime:hasStartPosition  "84"^^xsd:long ;
                                                         xlime:hasStopPosition   "89"^^xsd:long
                                                       ]
                                ] ;
        xlime:hasAnnotation     [ xlime:hasConfidence  "1"^^xsd:double ;
                                  xlime:hasEntity     dbpedia:Tennis ;
```

```
                    xlime:hasPosition    [ xlime:hasStartPosition  "16"^^xsd:long ;
                                           xlime:hasStopPosition   "22"^^xsd:long
                                         ]
                  ] ;
        xlime:hasSourceType     "Video" ;
        xlime:keywordFilterName  "Ski Region" .
```

Article published is converted into JSON format and stored.

**Table 4 Vimeo messages in JSON Format**

```
{
   u'Lang': u'de',
   u'SourceURL':u'https://vimeo.com/user12835962',
   u'Publisher': u'Article',
   u'Text':u' Ein Paar spielt Tennis.Kinder laufen immer wieder über das Spielfeld, behindern das Spiel aber nicht weiter.',
  u'Date':u'2016-04-13 13:34:04'
}
```

## 1.2 Relation to Other Work Packages and KPI's

In this section, we present relation of this deliverable to other work packages in Table 5 and knowledge performance indicators (KPI's) in Table 6.

**Table 5 Core Functionality of this Deliverable**

| **Component (Core Functionality)** | **Receive input from WP** | **Provides input to WP** |
|---|---|---|
| The core contribution of this deliverable is to provide content based recommendations to web shops based on the popularity of the products/brands in social media. | D3.3.2 – Text Annotation<br>D4.2–Semantic Disambiguation | D7.4.2 – Fully Functional Prototype and Validation Report EXPLAIN. |

**Table 6 Relevant Key Performance Indicators**

| **Problem Definition** | **Objective Target (Evaluation Measure)** | **Measureable Progress** |
|---|---|---|
| Given social media information, identify the same or similar products/brands/categories in the catalogue of a web shop mentioned in the social media content. This will help to understand the popularity of | The target of the deliverable is to achieve good recall by identifying relevant products mentioned in the social media content. | Ability to build a sample prototype based on proposed approaches. |

| | | |
|---|---|---|
| items among users. | | |

## 1.3 Relation to Other Projects

In this section, we present the relation of this deliverable to other EU project deliverables in Table 7.

**Table 7 Relation to other EU project deliverables.**

| Component | Origin | Novel Contributions |
|---|---|---|
| Semantic annotation of text - XLisa | XLike | Application of the tool on social media content to understand its limitations over noisy text. |

# 2 Approaches

We divide our approaches into two different types based on the structured and unstructured information used to identify similar products in a web shop catalogue for content based recommendation. The first approach is based on distributional semantics which embed the semantics of words representing products/brands/categories in high-dimensional vectors. While, the second approach leverage text annotation provided with a knowledge-base.

## 2.1 Semantic Product Vectors (SPV)

Distributed representations of words (i.e. word embeddings) are used for several natural language processing tasks. Many approaches are also proposed to learn word embeddings [8, 9] with an extension to phrases and document embeddings [10, 11]. But, word embeddings were most of the times learned with carefully handcrafted text such as news and Wikipedia articles. In order to meet the requirements of short or noisy text generated in social media, some approaches [12] have designed variations of embeddings.

Here, we also use the variation of embeddings obtained by the standard word embedding approach adapted to the noisy text such as Twitter messages. Especially, we use word2vec model [8] adapted to learn vocabulary size of 3,039,345 with the dimensionality of 400 using 400 million Twitter messages proposed by Godin et al., [13] as a part of ACL W-NUT task [14].

Embeddings obtained for the words in noisy text is further used to map products/brands/category names of a web shop. For example, Table 8 shows the information about a particular product present in a web shop. It has information about many fields such as colour, material, name, brand etc.

**Table 8 Webshop Product Information (German)**


```
"http://xlime.econda.de/feed/00000505-932ea6a3-9fa4-4f5c-84ee-122a9ee7942d-1/products/1100057": {
    "Absatz": "cm",
    "Farbe": "schwarz",
    "Innenmaterial": "Synthetik, Canvas",
    "Laufsohle": "TR",
    "Obermaterial": "PU, Synthetik",
    "brand": "Graceland",
    "categories": [
      "http://xlime.econda.de/feed/00000505-932ea6a3-9fa4-4f5c-84ee-122a9ee7942d-
1/categories/productcategory#0000001574damen-schuhe-mokassins"
    ],
    "categoryNames": [
      "Mokassins"
    ],
    "description": "Mit sportivem Chic kommt der schwarze Mokassin von Graceland daher. Eine
wei\u00dfe 360 \u00b0-Schn\u00fcrung, helle Kontrastn\u00e4hte und die Wulstnaht auf dem Vorderblatt
sorgen f\u00fcr einen authentischen Look. Das Innenmaterial besteht aus Synthetik und gestreiftem
Canvas. Die Sohle und der 1,2-cm-Absatz kontrastieren in Beige-Braun mit der Grundfarbe des
Obermaterials. H\u00fcbscher Hingucker: der Stick in Rot, Blau und Wei\u00df an der Au\u00dfenseite.",
    "name": "Mokassin"
  }
```


We use important fields like **name** and **brand** present in English to assign them with the vectors obtained from the vocabulary of tweets. If more than one word is present, average of vectors is taken. This helps to map brand and product names in the web shop catalogue to the same space as of tweet words. Also,

another advantage of mapping brands/product names to vectors is that it is useful to obtain similar products with cosine similarity rather than identifying same products with string matching.

## 2.2 Text Annotation

Identification of entities in the social media can help to identify trending or popular products. In the text annotation approach, social media text is annotated with the product knowledge-base of a web shop. This helps to identify user's interest (either positive or negative) about particular products present in the web shop. This approach is similar to the text annotation approach [2]. For example a sample annotation can be visualized in Table 9 from the text annotation of D3.3.2. Here, DBpedia will be replaced with a web shop knowledge-base.

**Table 9 Text Annotation Example with Products (Example with DBpedia entites)**

**Social Media Post:** "Deals: FILA Men's Memory Solidarity Running Shoe (2 Colors) $24.99 (https://t.co/OZFhVB8O4U) https://t.co/raCQamy4mt" ;
**xlime:hasAnnotation** [ xlime:hasConfidence "0.476"^^xsd:double ;
xlime:hasEntity dbpedia:Memory ;
xlime:hasPosition [ xlime:hasStartPosition "18"^^xsd:long ;
xlime:hasStopPosition "24"^^xsd:long
] ] ;
**xlime:hasAnnotation** [ xlime:hasConfidence "0.85"^^xsd:double ;
xlime:hasEntity dbpedia:Shoe ;
xlime:hasPosition [ xlime:hasStartPosition "44"^^xsd:long ;
xlime:hasStopPosition "48"^^xsd:long
] ] ;
**xlime:hasAnnotation** [ xlime:hasConfidence "0.849"^^xsd:double ;
xlime:hasEntity dbpedia:Running ;
xlime:hasPosition [ xlime:hasStartPosition "36"^^xsd:long ;
xlime:hasStopPosition "43"^^xsd:long
] ] ;
**xlime:keywordFilterName** "Econda Shoes EN" .

## 2.3 Recommendation

The recommendation of products/brands/categories in a web shop is aligned to its popularity in the social media. Major challenge of identifying the popular products/brands/categories depends on their detection in social media. Sometimes syntactic matching may not be useful to find the closely related items, thus creating a need for sophisticated approaches.

In the following sections, we present the usage of our two proposed approaches SPV and text annotation to provide recommendations.

### 2.3.1 SPV Approach

Aim of the SPV approach is to identify the entities in the social media content that is closely related to the products/brands/categories in the database of a web shop. As the first step, instead of using embeddings for all the words in Twitter messages we use TweetNLP [15] to identify part-of-speech (POS) tags for the words in the tweets. This will help us to identify common or proper nouns that may represent entities or

possible products/brands/categories in a web shop catalogue. In the second step, nouns are assigned to the embeddings obtained from Godin et al [13].

Now, for the content base recommendations, words tagged by TweetNLP as 'N' for the Twitter message is used to identify closely related products/brands/categories in the database of a web shop using cosine similarity. A certain threshold is kept to eliminate false positives of similar items. This approach is much better than basic string comparison as it also helps to identify similar products by just limiting itself in identification of same products.

Now, popular products/brands/categories among users in social media are further used as recommendation in the web shop.

### 2.3.2 Text Annotation Approach

The recommendation step performed in the text annotation approach is similar to the earlier approach, but the way entities are identified in social media is different. Service developed in text annotation [2] is used for annotating products/brands/categories in the social media. This approach to identification of entities is different from semantic vectors as it depends on background knowledge.

# 3 Use-Cases

Here, we present a use-case where the Twitter messages generated from VICO is used to identify mentions of the products or brands present in the Deichmann web shop provided by ECONDA [16].

## 3.1 Data

The data used for preliminary evaluation is generated from VICO and ECONDA. VICO provided Twitter messages mostly belonging to certain brands or products, while ECONDA data is about products and categories in the Deichmann web shop. More information about the content is listed in Table 10 and Table 11.

**Table 10 Social Media Data**

| Language | Size | Twitter Messages |
|---|---|---|
| **English (en)** | 55MB | 787436 |
| **German (de)** | 1.7MB | 21476 |

**Table 11 Econda Diechmann Web Shop Data (Total)**

| **Products** | **Categories** | **Brands** |
|---|---|---|
| 290 | 235 | 63 |

Product and category information in the Deichmann web shop is present in both English and German. Here, we use only those products, brands and categories described in English. Table 12 present more information.

**Table 12 Econda Deichmann Web Shop Data (English)**

| **Products** | **Categories** | **Brands** |
|---|---|---|
| 99 | 97 | 60 |

## 3.2 Recommendations

Generating recommendations is associated with several challenges that require attention. In the following sections, we present challenges and the corresponding solutions proposed to overcome them.

### 3.2.1 Challenges

Challenges are segregated and are presented in Table 13.

**Table 13 Challenges**

| **Challenges** | **Description** |
|---|---|
| C1 | Deal with Noisy Social Media Text |

| | |
|---|---|
| C2 | Identification of entities in social media that may represent products or brands in a web shop. |
| C3 | Confidence or Threshold score to confirm an entity as product or brand |
| C4 | Evaluating recommendation quality. |

### 3.2.2 Approach

To address the challenges, we design a modular approach to solve them. Table 14 provides the solution to each of them separately.

**Table 14 Solution to Challenges**

| Challenges | Solution |
|---|---|
| C1: Noisy Text | Pre-processing |
| C2: Entities | TweetNLP to identify nouns that can be possible brands or products. |
| C3: Threshold Score | Empirical evidence is used to decide threshold score |
| C4: Evaluation | Manual annotated data set to evaluate the relevance of recommendations. |

### 3.2.3 Implementation

We leveraged SPV approach presented in Section 2.3.1 as opposed to the annotation approach to provide accurate and fast recommendations.

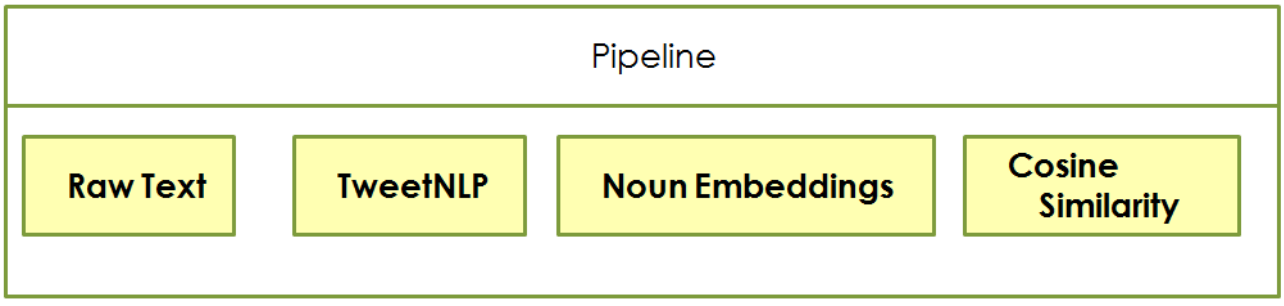
**Step-1:** Each Twitter message obtained from the data given in Table 8 is passed through the TweetNLP to obtain nouns. In most of the cases, nouns may represent entities representing products/brands/categories present in a web shop.

**Step-2:** Each noun is now searched in the vocabulary provided by Godin et al., for assigning high-dimensional vector.

**Step-3:** From Section 2.3.1, it is already known that products/brands/categories of a web shop were also mapped to the vocabulary of Godin et al. Now for the annotation of nouns obtained from step-2 with products/brands/categories of a web shop, cosine similarity between all products/brands/categories of a web shop is calculated to obtain top-3 similar ones. Finally, for the entire Twitter message intersection of all annotations is reported. If no intersection, top-2 from each one is reported.

The overall goal of the approach is shown in Figure 2.

**Figure 2 Implementation Pipeline**



### 3.2.4 Results – Qualitative Analysis

Nouns inside Twitter messages annotated with products, brands and categories of a web shop are pushed into Kafka [17] with the topic name "**KITSocialMediaWebshop**". Table 15 shows example annotations obtained on the data given in Table 10.

**Table 15 Annotations - Example**

{
"social media": "* 00:00 newells adidas"
"products":[slipper, mid cut priority mid, party clutch, sling pumps]
"brands":[victory,victory performance, frozen,monster high]
"categories":[get the look, online exklusiv, fernanda brandao, kids]
}

This Kafka message is a JSON object encapsulated with the following structure. Table 16 shows the structure information.

**Table 16 Key-Value Information**

| Key | Value |
|---|---|
| Social Media | Social Media Text |
| Products | Product annotations of Diechmann web shop. |
| Brands | Brand annotations of Diechmann web shop. |
| Categories | Category annotations of Diechmann web shop. |

Quantitative evaluation measures such as precision, recall and f1-score will be used to evaluate the approach in D7.4.2 with the manually annotated dataset from ECONDA using their Deichmann web shop.

# 4 Conclusion and Future Work

In this deliverable, we have explored two different approaches to perform content based recommendations. Both of these approaches identify entities in the social media content that might represent products in a web shop, thus helping to identify popular or trending products that can be used as recommendations to customers of a web shop.

In future, we will extend the approaches to incorporate more complex scenarios where the meta-data of the users or more information about products such as descriptions, colour etc., can be used to provide recommendations. Also, in D7.4.2 evaluation on a manually annotated dataset created by ECONDA will be presented using the presented approaches.

# References


[1] Semantic Disambiguation Prototype (D4.2)

[2] Text Annotation (D3.3.2)

[3] https://twitter.com/ (Twitter)

[4] http://vimeo.com/ (Vimeo)

[5] https://www.youtube.com/ (YouTube)

[6] http://www.vico-research.com/ (VICO)

[7] xLiMe Meta Data Model (D1.1)

[8] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

[9] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12 (2014): 1532-1543.

[10] Le, Quoc V., and Tomas Mikolov. "Distributed representations of sentences and documents." *arXiv preprint arXiv:1405.4053* (2014).

[11] Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. "Semi-supervised recursive autoencoders for predicting sentiment distributions." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 151-161. Association for Computational Linguistics, 2011.

[12] Severyn, Aliaksei, and Alessandro Moschitti. "Twitter sentiment analysis with deep convolutional neural networks." In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 959-962. ACM, 2015.

[13] Godin, Fréderic, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. "Multimedia Lab@ ACL W-NUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations." In *ACL 2015 Workshop on Noisy User-generated Text*, pp. 146-153. Association for Computational Linguistics, 2015.

[14] https://noisy-text.github.io/ (ACL 2015 Workshop on Noisy User generated Text)

[15] http://www.cs.cmu.edu/~ark/TweetNLP/

[16] http://www.econda.com/ (ECONDA)

[17] http://kafka.apache.org/ (Kafka)