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

Deliverable D2.3.1

Early Text from Social Media Prototype

Editor:	Luis Rei, JSI
Author(s):	Luis Rei, JSI; Blaz Novak, JSI; Dunja Mladenic, JSI
Deliverable Nature:	Prototype (P)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M12 – 31 October 2014
Actual Delivery Date:	M12 – 31 October 2014
Suggested Readers:	xLiMe Project partners
Version:	1.0
Keywords:	Text mining, natural language processing, named entity recognition

Disclaimer

Full Project Title:	xLiMe – crossLingual crossMedia knowledge extraction
Short Project Title:	xLiMe
Number and Title of Work Package:	WP2 Text Extraction from Multilingual Multimedia Natural Language
Document Title:	D2.3.1 - Early Text from Social Media Prototype
Editor:	Luis Rei, JSI
Work Package Leader:	Blaž Novak, JSI

Executive Summary

The deliverable presents an early prototype for processing text from social media streams. The main objectives of this deliverable are to produce a representation that allows further statistical processing including integration with information extracted from video and annotations from the extracted information. The developed prototype includes both a vector-space-model representation of the text and annotations, namely named entities on normalized social media text.

This early prototype has been optimized for social media and is focused on English language. The plan for the final prototype is to go beyond English handling other languages most probably German, Italian, Spanish and to investigate the coverage of a less spoken language such as Catalan or Portuguese.

It is important to point out that this deliverable fulfils the requirements laid out in xLiMe deliverable 1.4.1 "Requirements for Early Prototype".

Table of Contents

Abbreviations

CRF	Conditional Random Fields
IOB	Inside, Outside, Beginning
LDA	Latent Dirichlet Allocation
NER	Named Entity Recognition
NLP	Natural Language Processing
NP	Noun Phrase
RDF	Resource Description Framework
POS	Part-Of-Speech
OOV	Out-Of-Vocabulary
SVM	Support Vector Machines
URL	Unified Resource Locator

1 Introduction

Social media has made following developing news stories anywhere in the world both easy and popular. Tens, sometimes hundreds, of millions of social network users can view photos, videos and comments from people at the scenes of an event almost instantly. Publishers and readers alike post links to news articles about the events to the social networks, providing users with more information at the distance of a click. In many cases, such as for instance, recent protests in the North American town of Ferguson or the Asian metropolis of Hong Kong, traditional media now source their stories from social networks. In the realm of entertainment, TV shows, movies premiering in theatres, concerts, live shows and conferences are now accompanied by the live comments of their viewers. Brands have realized both the positive and negative potential of social networks with viral campaigns such as Cadbury's "Operation Goo" and infamous disasters such as the great Nestle-Greenpeace debacle of 2010.

While the need to real-time monitor and integrate social media streams into news, entertainment and companies' public relations and marketing efforts is widely acknowledged, several technical challenges complicate such integration. The first and the most obvious is volume: with 500M posts per day on the Twitter network alone, how can we process the social media stream efficiently in close to real-time? The second challenge is inherent to the social media text itself: a single short piece of text, which usually relies on external context and is often lacking the grammatical niceties we are accustomed to in other forms of text. How can we extract useful information from such text? Finally we are presented with the challenge of matching the social media stream contents to the contents of other continuous streams such as the TV video stream and the traditional news and blog streams. The process of matching information across different streams is made even more challenging by the need to also match content across languages.

In this deliverable (D2.3.1) we present an early prototype for a text representation, which allows further statistical and lightweight semantic processing as well as integration with other media. We create both a vector-space-model representation of the text and annotations, namely named entities on normalized social media text. This deliverable fulfils the requirements laid out in deliverable D1.4.1 Requirements For Early Prototype [1].

2 Social Media Text Processing

Text processing can be thought of as a pipeline, as shown in Figure 1 where a series of steps that occur in order to transform the input so that the desired output can be created or extracted. This pipeline contains 3 main parts:

1. Input data; independent streams of text data, documents, from different sources
2. Pre-processing; we apply a series of transformations to each document to prepare it for the output generation
3. Output; we extract the desired information from each pre-processed document in order to generate the output

In this section we describe each part in more detail.

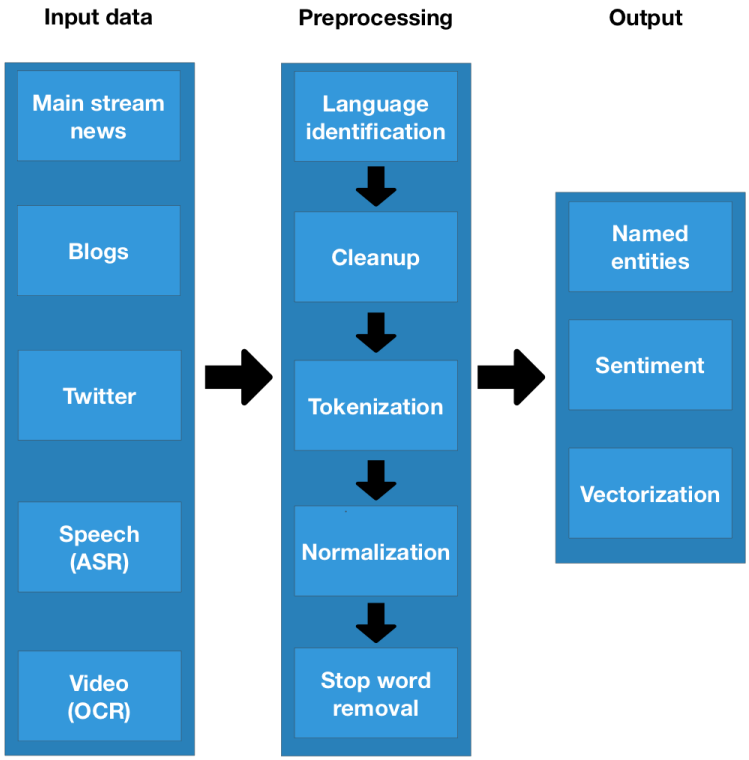

Figure 1: Text Processing Pipeline

2.1 Input Data

The main focus of this deliverable is to process social and generally noisy text. The data available are text documents obtained from the project partners:

- Main stream news & Blogs; from the JSI Newsfeed service (http://newsfeed.ijs.si/) [2] which collects articles from the RSS feeds of most worldwide online news sources and blogs
- Twitter; directly from the twitter public API as well as from VICO's monitoring platform [3]
- Speech; the output of Vecsys Automatic Speech Recognition system [4] used on Zattoo's video channels [5]
- Video; output of Optical Character Recognition used on Zattoo's video channels

While, as stated, the focus is on social media, the prototype described in this deliverable currently handles anything that is pushed as text. At the time of writing, this was just social media and main stream news. Nevertheless, this early prototype has been optimized for social media. It's also important to state that for this early prototype we have focused on English language only. The plan for the final prototype is to also handle German, Italian and Spanish. An investigation of the coverage of a less spoken language will be carried out during the second year.

2.2 Pre-Processing

2.2.1 Language Identification

Documents must be separated by language in order for subsequent steps to work as intended. All our data sources report language with an acceptable degree of accuracy: we know the language of Zattoo's channels, main stream news and blogs. Twitter uses machine learning to classify its data into languages, we take advantage of their classification. As reported previously, for this early prototype, we have focused on English and thus we currently ignore text that is identified as anything other than English. In the final prototype we will use the language identification to select different language-specific optimizations of the pipeline. If at any point it becomes necessary to perform language identification using machine learning, we are capable of doing it having tested components for this purpose.

2.2.2 Cleanup

Social media text contains special markup and elements that must be handled. These include mentions of other users, companies and topics, identified in twitter by preceding the token with the symbols '@', '$' and '#', respectively. URLs and smileys are also common in social media. All of these, as well as numbers, are identified in the text through the use of regular expressions and we can choose to either remove, replace or keep each of these special types of text depending on the needs of each step in the next part of the pipeline.

2.2.3 Tokenization

Tokenization consists of splitting a document into tokens, mostly individual words. This is often slightly harder on social media than news text due to common conventions not being followed, the presence of strings of punctuation (e.g. +====) and the presence of Unicode glyphs (e.g. U+2620, skull and crossbones). These can make standard tokenizers work poorly. We've implemented Tweetmotif's [6] tokenizer for the current Named entity recognition, which is the most sensitive task to tokenization issues while the other tasks use a simpler tokenization procedure based on splitting tokens on non-alphanumeric characters.

2.2.4 Normalization

For the vectorization task, all tokens are converted to lowercase letters, punctuation, special characters and accents are removed. Under this transformation, the German text is modified from 1) to 2):

Text 1:

1574 – Im niederländischen Unabhängigkeits-krieg gelingt es den Aufständischen, die Belagerung von Leiden durch die Spanier zu beenden

Text 2:

1574 im niederlandischen unabhangigkeits krieg gelingt es den aufstandischen die belagerung von leiden durch die spanier zu beenden

While this is a common procedure in text mining, it is particularly important in our case as social network text often omits accents. Alternative, language-specific adaptations of this transformation can also be

carried out. For example, the German Umlaute is normally replaced with the corresponding vocal followed by an "e". So in the example above *ä* would be replaced by *ae*.

2.2.5 Stop Word Removal

The stop words of a language are usually some or all of its most common words. These include function words such as the English words “the” and “and”. The removal of these words results in better computational performance and better results for most text mining tasks. Of course, the stop words that are removed are language dependent and there exist publicly available lists of stop-words for major languages. Since the current early prototype focus on English, we have only implemented this for the English language but we have lists ready for all other languages.

It is important to note that stop word removal is used for vectorization but not for NER since the NLP algorithms for this task work in a fundamentally different way to text mining algorithms. NER is a sentence level task reliant on the structure of the sentence, thus removing stop words is not advisable.

2.3 Named Entity Recognition

As the name suggests, Named entity recognition (NER) consists of identifying named entities, such as persons, organizations and locations, in a document. This is a subtask of information extraction, the improvement of which was the focus of the Message Understanding Conferences (MUC) sponsored by DARPA from the late 80s to the late 90s. The final results of the NER subtask in these conferences for English news text were close to human annotators [7] and subsequent improvements were obtained during the Conference on Computational Natural Language Learning (CoNLL) shared task in the early 2000s. The CoNLL shared task also added a new catchall entity type for entities that did not fit into the MUC types [8].

Many “off-the-shelf” NER tools exist, however these have weak performance on social media text. Social media text is informal, noisy, often ungrammatical, and importantly, lacks the context necessary for standard NLP tools to work well. Some recent attempts at creating specific pipelines for social media text have managed to obtain better results [9] but these still fall significantly below the performance achieved in news text.

In this early prototype we adopted the approach described in [9] which treats segmentation (is this an entity?) and classification (what is the type of this entity?) as separate tasks. Segmentation uses Conditional Random Fields (CRFs) with a wide range of features: orthographic, contextual, dictionary features, Brown clusters Part-of-Speech (POS) tags, Noun-phrase (NP) chunking and the output of a classifier that predicts whether capitalization is informative. Classifying entities in to types on twitter is particularly hard given that context is often as short as two or three words. In this subtask, distant supervision with LabelledLDA topic models is used together with dictionaries created from Freebase.

In

Table 1 we compare the results of this social text optimized approach with the news-trained classifier Enrycher [10]. The dataset used contains 2400 tweets containing a total 34000 tokens and 1127 annotated entities distributed among classes according to Table 2. The evaluation procedure consisted of a 4-fold cross validation for the component implemented in this deliverable: 75% of the data is used to train the model, 25% is used for evaluation in each fold. All data is tested and the standard precision, recall and F1 (harmonic mean of precision and recall) are calculated. Evaluation of Enrycher on this dataset did not require cross-validation as the system is not trained on this dataset.

In the following sections we provide additional details the current NER pipeline.

Table 1: Comparison of Named Entity Recognition results on a tweet corpus.

System	Precision	Recall	F1
Enrycher	0.49	0.22	0.28
xLiMe D2.3.1 [RITTER]	0.73	0.49	0.59

Table 2: Distribution of Entity Types in Tweet Dataset

Entity Class Type	Number of Entities
Person	436
Location	372
Organization	319
Total	1127

2.3.1 Part-of-Speech Tagging

We use traditional Penn TreeBank POS tags plus twitter specific tags for URLs, retweets, usernames and hashtags. Social media has a bigger vocabulary than main stream news. This large amount of "new" words result mainly from misspellings and abbreviations. For example, "yes" is commonly misspelled as "yess" while "tomorrow" can be abbreviated to "2morrow", "tmrow", "2mor" and several more ways. To account for OOV words and lexical variations, words have been hierarchically clustered resulting in Brown clusters from which the prefixes used are 4, 8 and 12 bits. Conditional Random Fields are used with the Brown clusters and a fairly standard set of features that include POS dictionaries, spelling and contextual features.

2.3.2 Shallow Parsing (NP-Chunking)

The shallow parsing sub-component also uses CRFs and the Brown cluster features described above together with the features originally described in [11].

2.3.3 Capitalization

Capitalization which is usually a very informative, easy to use, feature when recognizing named entities in news text, requires more work in social media as many documents have either no capitalization, are written in full caps, use capitalization just for emphasis or just have an inconsistent capitalization style. The approach we describe uses a capitalization classifier on the entire text of a document to decide whether or not the capitalization style used is informative or not (binary classification). Capitalization is considered uninformative if the document contains non-entity words which are capitalized, but do not begin a sentence, or if it contains any entities which are not capitalized. The classifier for this sub-task uses Support Vector Machines.

2.3.4 Named Entity Segmentation

Named Entity Segmentation consists of identifying segments (sequences of tokens) as entities in a sentence or, in our case, a tweet. A named entity can consist of one or more tokens e.g. "JFK" (1 token) "John Fitzgerald Kennedy" (3 tokens). This sub-task is modelled as a sequence-labelling (classification) task using the standard IOB encoding: each token is either inside a named entity (I), outside a named entity (O) or begins a named entity (B). Again, CRFs are used for inference. The features used include the POS tags obtained in 2.3.1 as well as the Brown clusters described there, the Chunking tags obtained in 2.3.2, capitalization features including the one described in 2.3.3 as well as orthographic, contextual and dictionary features. The dictionaries include a set type lists obtained from Freebase [12].

2.3.5 Named Entity Classification

Social media messages often contain very little context and rely heavily on the intended audience having that context. NLP systems usually do not have such context. Take the tweets “HYMYM is amazing” and “NPH is amazing”. Without context, both a human reader and an NLP system might be tricked into thinking that “HYMYM” and “NPH” are the same class of entities. With context, however, it is possible to know that “HYMYM” refers to a TV show (Miscellaneous class) and “NPH” refers to a person (Person class). This context can be obtained in different ways, by co-occurrence in other examples where more information is available and by leveraging dictionaries and entity lists. An additional way in which context can be enhanced is to create a topic model over each entity and constrain it over the entities’ possible types in Freebase. This is only possible for entities present in the Freebase database. In this scenario, the topic distribution for “Amazon” is constrained to the types Location and Organization. For entities not present on Freebase, the topic distribution must remain unconstrained.

2.4 Vectorization

To transform a social network text into a vector-space-model we use the Bag-of-Words representation originally described by Harris [13]. We also build n-grams specifically bigrams and trigrams.

Unlike with traditional media where vocabularies for a given language are in the tens of thousands of words and rarely change significantly, on social media they can easily reach millions of words and grow rapidly. New words can be important as they can describe particular topics or events being discussed such as the word “fappening” which described the recent celebrity hacking incident. As such, the traditional dictionary model of bag-of-words is ill suited for this problem. Another issue is that the traditional dictionary approach means that multiple processes handling media streams would need to keep their dictionaries synchronized.

We use the Hashing Trick [14] where words and n-grams are mapped to indices with a hashing function, which handles the previous mentioned issues and improves memory use. We use MurmurHash3 with 28 bits. The tweet in 1) is transformed into 2)

Tweet 1:

"Exactly two years ago we watched the jump of @BaumgartneFelix from @RedBullStratos and I met him this year. Crazy things. #respect"

Tweet 2:

{'9410507': 1, '14858927': 1, '4501016': 1, '8192699': 1, '9247738': 1, '3632780': 1, '2231977': 1, '11787678': 1, '587318': 1, '2897153': 1, '8775141': 1, '16448651': 1, '15280316': 1, '13561368': 1, '14034445': 1, '11374614': 1, '8166572': 1, '3826860': 1, '3810374': 1, '15201779': 1, '4679224': 1, '14537201': 1, '2805437': 1, '16324147': 1, '5751589': 1, '4602601': 1, '5772415': 1, '2303228': 1, '661528': 1, '13461415': 1, '11418980': 1, '9726511': 1, '15452273': 1}

3 Integration into the Project Stream

As detailed in deliverable D1.2 Prototype of Data Processing Infrastructure [15], the project uses Kafka with separate topics for each data provider and annotation provider. This deliverable is implemented as an annotation provider which pushes annotations to the “jsi-annotations” topic while listening to all topics that contain text data, most importantly, the VICO social media stream.

The RDF data model used was originally specified in deliverable 1.1 Prototype of the (Meta) Data Model [16], this deliverable both consumes and produces RDF data. The following is an example of a message originating with newsfeed and the generated annotation:

Newsfeed Message:

```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix freebase: <http://rdf.freebase.com/ns/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix kdo: <http://kdo.render-project.eu/kdo#> .
@prefix ma: <http://www.w3.org/ns/ma-ont#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix wikide: <http://de.wikipedia.org/wiki/> .
@prefix wikien: <http://en.wikipedia.org/wiki/> .
@prefix wikies: <http://es.wikipedia.org/wiki/> .
@prefix wikisl: <http://sl.wikipedia.org/wiki/> .
@prefix xlime: <http://xlime-project.org/vocab/> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://ijs.si/enryched/218922538> {
  <http://ijs.si/article/218922538> a kdo:NewsArticle ;
    dcterms:created "2014-10-16T15:40:07.349138"^^xsd:datetime ;
    dcterms:language "en" ;
    dcterms:publisher [ rdfs:label "Sify" ] ;
    dcterms:source <http://www.sify.com/news/telcos-trying-to-restore-mobile-connectivity-in-vizag-news-others-okqva6gbbbbej.html> ;
    dcterms:spatial [ gn:name "India",
            "New Delhi" ;
```

geo:lat 2.17866e+01 ;

geo:long 8.279476e+01] ;

dcterms:title "Telcos trying to restore mobile connectivity in Vizag" ;

sioc:content """Telcos trying to restore mobile connectivity in Vizag

New Delhi, Oct 16 (IANS) Mobile operators are working jointly to restore mobile connectivity in the affected areas of Andhra Pradesh's Vizag after the devastating cyclone Hudhud, the Cellular Operators' Association of India (COAI) said here Thursday.

The operators engaged in restoration work are Aircel, Bharti Airtel, Idea Cellular and Vodafone.

"After being hit severely by the cyclone which caused severe damage to telecom infrastructure, thereby impacting networks across all operators in a huge way, the telecom industry is working collectively to put its network back online," the COAI statement said."""" ;

sioc:topic "Business",

"Business_and_Economy",

"Computers",

"Europe",

"Mobile_Computing",

"Regional",

"Telecommunications",

"United_Kingdom" ;

xlime:hasAnnotation [rdfs:label "Andhra Pradesh" ;

xlime:hasConfidence 5.44e-01 ;

xlime:hasEntity wikide:Andhra_Pradesh,

wikien:Andhra_Pradesh,

wikies:Andhra_Pradesh],

[rdfs:label "Vodafone" ;

xlime:hasConfidence 3.39e-01 ;

xlime:hasEntity wikide:Vodafone_Group,

wikien:Vodafone,

wikies:Vodafone],

[rdfs:label "Bharti Airtel" ;

xlime:hasConfidence 1e+00 ;

xlime:hasEntity wikide:Bharti_Airtel,

wikien:Bharti_Airtel],

[rdfs:label "Mobile network operator" ;

xlime:hasConfidence 2.01e-01 ;

xlime:hasEntity wikide:Mobilfunkgesellschaft,

wikien:Mobile_network_operator,

wikies:Operador_de_telefonía_móvil],


```
[ rdfs:label "Aircel" ;
   xlime:hasConfidence 6.64e-01 ;
   xlime:hasEntity wikien:Aircel ],
[ rdfs:label "Cyclone" ;
   xlime:hasConfidence 7.9e-02 ;
   xlime:hasEntity wikide:Zyklon,
      wikien:Cyclone,
      <http://es.wikipedia.org/wiki/Ciclón_(fenómeno_natural)>,
      wikisl:Ciklon ],
[ rdfs:label "Infrastructure" ;
   xlime:hasConfidence 8e-02 ;
   xlime:hasEntity wikide:Infrastruktur,
      wikien:Infrastructure,
      wikies:Infraestructura_urbana ],
[ rdfs:label "Idea Cellular" ;
   xlime:hasConfidence 9.02e-01 ;
   xlime:hasEntity wikide:Idea_Cellular,
      wikien:Idea_Cellular ],
[ rdfs:label "Telecommunication" ;
   xlime:hasConfidence 2.22e-01 ;
   xlime:hasEntity wikide:Telekommunikation,
      wikien:Telecommunication,
      wikies:Telecomunicación,
      wikisl:Telekom ],
[ rdfs:label "Vodafone" ;
   xlime:hasEntity freebase:m.02d6ph ],
[ rdfs:label "Visakhapatnam" ;
   xlime:hasConfidence 4.88e-01 ;
   xlime:hasEntity wikide:Visakhapatnam,
      wikien:Visakhapatnam,
      wikies:Visakhapatnam ],
[ rdfs:label "New Delhi" ;
   xlime:hasConfidence 5.75e-01 ;
   xlime:hasEntity wikide:Neu-Delhi,
      wikien:New_Delhi,
      wikies:Nueva_Delhi,
      wikisl:New_Delhi ],
```

```
        [ rdfs:label "India" ;
          xlime:hasConfidence 8.9e-01 ;
          xlime:hasEntity wikide:Indien,
            wikien:India,
            wikies:India,
            wikisl:Indija ] .
}

{
  <http://ijs.si/enryched/218922538> prov:wasGeneratedBy [ a prov:Activity ;
          dcterms:title "JSI Newsfeed" ;
          prov:endedAtTime "2014-10-16T15:41:00+00:00"^^xsd:datetime ;
          prov:startedAtTime "2014-10-16T17:40:56.785981+02:00"^^xsd:datetime ;
          prov:wasAttributedTo <http://ijs.si> ] .
}
```

Resulting annotations:

```
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix kdo: <http://kdo.render-project.eu/kdo#> .
@prefix ma: <http://www.w3.org/ns/ma-ont#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .
@prefix xlime: <http://xlime-project.org/vocab/> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

{
  <http://ijs.si/enryched/218922538> prov:wasGeneratedBy [ a prov:Activity ;
          dcterms:title "JSI Annotator" ;
          prov:endedAtTime "2014-10-16T15:41:21+00:00"^^xsd:datetime ;
          prov:startedAtTime "2014-10-16T17:41:17.733679+02:00"^^xsd:datetime ;
          prov:wasAttributedTo <http://ijs.si> ] .
```


}

<http://ijs.si/enryched/218922538> {
<http://ijs.si/article/218922538> kdo:hasSentiment [kdo:hasScore 7.54308e-01 ;
kdo:sentiment kdo:negativePolarity] ;
xlime:hasAnnotation [rdfs:label "Bharti Airtel" ;
xlime:hasPosition [xlime:hasStartPosition 369 ;
xlime:hasStopPosition 381]],
[rdfs:label "Association of India" ;
xlime:hasPosition [xlime:hasStartPosition 263 ;
xlime:hasStopPosition 282]],
[rdfs:label "Andhra Pradesh" ;
xlime:hasPosition [xlime:hasStartPosition 175 ;
xlime:hasStopPosition 188]],
[rdfs:label "COAI" ;
xlime:hasPosition [xlime:hasStartPosition 286 ;
xlime:hasStopPosition 289]],
[rdfs:label "Vizag New Delhi" ;
xlime:hasPosition [xlime:hasStartPosition 48 ;
xlime:hasStopPosition 62]],
[rdfs:label "Mobile" ;
xlime:hasPosition [xlime:hasStartPosition 82 ;
xlime:hasStopPosition 87]],
[rdfs:label "Idea Cellular and Vodafone" ;
xlime:hasPosition [xlime:hasStartPosition 385 ;
xlime:hasStopPosition 410]],
[rdfs:label "Vizag" ;
xlime:hasPosition [xlime:hasStartPosition 48 ;
xlime:hasStopPosition 52]],
[rdfs:label "COAI" ;
xlime:hasPosition [xlime:hasStartPosition 286 ;
xlime:hasStopPosition 289]] ;
xlime:hasNgrams [xlime:ngrams "{\"4230366\": 1, \"3073330\": 1, \"9429082\": 1, \"5285735\": 1, \"8938141\": 1, \"8324092\": 1, \"8627914\": 1, \"1699299\": 1, \"1680895\": 1, \"13102980\": 1, \"11005329\": 1, \"16613991\": 1, \"8417232\": 1, \"11566639\": 1, \"14783984\": 1, \"3232951\": 1, \"8713076\": 1, \"5860815\": 1, \"15825855\": 1, \"12166233\": 1, \"5893113\": 1, \"1777966\": 1, \"8484593\": 1, \"13175945\": 1, \"5730892\": 1, \"10273461\": 1, \"6115507\": 1, \"10266851\": 1, \"13057237\": 1, \"8395089\": 1, \"11168919\": 1, \"4976883\": 1, \"12262351\": 1, \"15358264\": 1, \"6689733\": 1, \"14564309\": 1, \"9200617\": 1, \"6464893\": 1, \"12731498\": 1, \"15584568\": 1,

\"15599110\": 1, \"12168197\": 1, \"3535052\": 1, \"10433976\": 1, \"8484469\": 1, \"11731554\": 1, \"1980586\": 1, \"11703047\": 1, \"8253990\": 1, \"11705411\": 1, \"15434493\": 1, \"8774610\": 1, \"5731431\": 1, \"9513854\": 1, \"2216858\": 1, \"9636398\": 1, \"5966736\": 1, \"16280668\": 1, \"2148181\": 1, \"923795\": 1, \"9006202\": 1, \"14906445\": 1, \"11531532\": 1, \"2847060\": 1, \"5133109\": 1, \"362124\": 1, \"22108\": 1, \"15466503\": 1, \"3725796\": 1, \"3323824\": 1, \"12849008\": 1, \"8700931\": 1, \"7367450\": 1, \"11812602\": 1, \"7645986\": 1, \"14721507\": 1, \"4716011\": 1, \"16761633\": 1, \"7703709\": 1, \"256076\": 1, \"7521324\": 1, \"826091\": 1, \"3870169\": 1, \"8773775\": 1, \"4995490\": 1, \"7502300\": 1, \"477269\": 1, \"1739217\": 1, \"13241727\": 1, \"12842785\": 1, \"13833219\": 1, \"3766547\": 1, \"11922661\": 1, \"16148343\": 1, \"15339144\": 1, \"4993317\": 1, \"8153136\": 1, \"10696835\": 1, \"1550277\": 1, \"5348943\": 1, \"5361090\": 1, \"9740330\": 1, \"13612285\": 1, \"12218456\": 1, \"6176339\": 1, \"409826\": 1, \"10254244\": 1, \"15511174\": 1, \"2467374\": 1, \"4820492\": 1, \"4802708\": 1, \"3843819\": 1, \"2549291\": 1, \"1072328\": 1, \"6065384\": 1, \"832670\": 1, \"4736848\": 1, \"3218866\": 1, \"8268752\": 1, \"3567826\": 1, \"12616900\": 1, \"2778503\": 1, \"845243\": 1, \"10420141\": 1, \"8678350\": 1, \"5935760\": 1, \"9411913\": 1, \"3513289\": 1, \"16044042\": 1, \"14449209\": 1, \"4141439\": 1, \"6282403\": 1, \"14852089\": 1, \"14399215\": 1, \"2615058\": 1, \"16650866\": 1, \"14892461\": 1, \"4713217\": 1, \"16600202\": 1, \"7574996\": 1, \"11238811\": 1, \"213965\": 1, \"11361809\": 1, \"11289419\": 1, \"12428271\": 1, \"87798\": 1, \"1587579\": 1, \"13209554\": 1, \"5874994\": 1, \"2058339\": 1, \"11312016\": 1, \"5172231\": 1, \"879179\": 1, \"13410032\": 1, \"15424001\": 1, \"2365136\": 1, \"1150949\": 1, \"14181018\": 1, \"1637538\": 1, \"5182508\": 1, \"499213\": 1, \"6594223\": 1, \"12852627\": 1, \"10198271\": 1, \"8861287\": 1, \"7281736\": 1, \"5807187\": 1, \"12283875\": 1, \"4425511\": 1, \"7013032\": 1, \"14524433\": 1, \"11315611\": 1, \"13273619\": 1, \"4851118\": 1, \"13039979\": 1}"] .

}

4 Future Work

In this early prototype the developed NER approach is only functional for English, the final prototype will cover all languages supported in the project. Additionally, we will investigate the coverage of a less spoken language, possibly Catalan or Portuguese. We would also like to follow on some research ideas for improving quality of the current approach implemented in the prototype. One such idea is to add an additional representation for text, namely Paragraph Vectors [17] that showed promising results in retrieval and sentiment analysis. We also would like to provide a better NER performance in social text. A promising line of research is the neural network based deep learning approach described in [18] and [19] which do not rely heavily on human feature engineering and could potentially avoid difficulties in obtaining the same performance across different languages.

5 Conclusion

This deliverable describes the work done for the xLiMe early social media text annotation prototype. It details the pre-processing applied to text, the named entity recognition approach and the vectorization. The approaches described were tailored or adapted specifically for social media.

It is important to point out that this prototype meets all the required criteria previously defined in the project. We provide results of the initial evaluation of named entity recognition on social media which meet the goals of the project. While results are sufficient, it's our hope that they can be further improved.

Named Entity Recognition needs to be expanded to support other languages and evaluation on them must be conducted for the final prototype of text from social media.

References

[1] XLime deliverable "D1.4.1" – Requirements For Early Prototype

[2] Trampus, Mitja and Novak, Blaz: The Internals Of An Aggregated Web News Feed. Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).

[3] http://vico-research.com/

[4] http://www.vecsys-technologies.fr/

[5] http://zattoo.com/

[6] O'Connor, Brendan, Michel Krieger, and David Ahn. "TweetMotif: Exploratory Search and Topic Summarization for Twitter." ICWSM, 2010.

[7] Elaine Marsh, Dennis Perzanowski, "MUC-7 Evaluation of IE Technology: Overview of Results", 29 April 1998

[8] Tjong Kim Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003.

[9] Ritter, Alan, Sam Clark, and Oren Etzioni. "Named entity recognition in tweets: an experimental study." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

[10] Štajner, Tadej: From unstructured to linked data: entity extraction and disambiguation by collective similarity maximization. In Identity and Reference in web-based Knowledge Representation (IR-KR): Proceedings of the IJCAI-09 workshop, 29-34.

[11] Sha, Fei, and Fernando Pereira. "Shallow parsing with conditional random fields." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

[12] http://www.freebase.com

[13] Harris, Zellig (1954). "DISTRIBUTIONAL STRUCTURE". WORD 10 (2/3): 146–62

[14] Weinberger, Kilian, et al. "Feature hashing for large scale multitask learning." Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009.

[15] XLime deliverable "D1.2" - Prototype of Data Processing Infrastructure

[16] XLime deliverable "D1.1" - Prototype of the (Meta) Data Model

[17] Lee, Quoc and Mikolov, Tomas "Distributed. "Representations of Sentences and Documents" ICML 2014

[18] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. "Natural language processing (almost) from scratch". The Journal of Machine Learning Research, 12, 2493-2537.

[19] Dos Santos, Cicero, Zadrozny, Bianca. "Learning Character-level Representations for Part-of-Speech Tagging", The Journal of Machine Learning Research, 32, 1818-1826, 24

