1. INTRODUCTION/BACKGROUND

The Global Security and Crisis Management Unit of the IPSC supports the Union’s policies to strengthen the EU’s resilience to crises and disasters as well the EU’s aim to promote stability and peace through its research in crisis management technologies and in information mining and analysis. The Unit’s EMM (Europe Media Monitor) project (formerly OPTIMA - Open Source Text Information Mining and Analysis action), develops innovative solutions for retrieving and extracting information from the internet, and especially from online news and social media, serving many Commission Services, EU agencies and some EU Member State authorities.

2. JRC PROJECTS WORKED ON

This annual report concerns the work carried out within the EMM project for the development and maintenance of automatic multilingual text analysis tools. In particular, research and development work has been performed on the EMM Open Source Intelligence Suite (EMM OSINT Suite). It is a desktop software application which helps to find, acquire and analyse data from the Internet and local sources. It provides automatic means to empower users to gather intelligence from open available sources by removing the need to search manually through vast data sets. The specific JRC projects that I have been working on are:

- 456, OSINT for HOME and Law Enforcement. The project enhances and further develops the OSINT Suite software for use by law enforcement agencies and other parties such as the International Criminal Court. Its activities focus mainly on providing customisation features in order to adapt the OSINT suite to the user domain. Other developments will increase the number of analysis components by including more EMM derived modules. The project will increase OSINT capacity building by further dissemination of the OSINT suite in MS.

- 965, EMM SERVICES 2013. The project enhances and supports the DG-COMM Media Monitoring platform on request of EC DG Communication. Its activities focus mainly on researching of multi-lingual information text mining and analysis tools and systems that exploit traditional online information sources as well as emerging information sources and social media.
3. **Progress of the Project, Timing and Results**

Regarding the OSINT Suite project, the main task consisted in the development of new functional modules and features to the desktop tool:

- **Local Search module.** This involved the development and testing of the core module and the user interface. This module allows performing local searches within the documents previously downloaded or imported into OSINT. It provides a rich query language through which the user can perform wildcard searches or boolean queries. Help documentation about how to use this new feature was provided in the support web site (https://wiki.emn4.eu):
  - Tutorial: “Using Local Search” ([link](https://wiki.emn4.eu))

![Local Search module](image)

*Figure 1: New Local Search module running on the OSINT Suite*

- **Export documents to XML.** This feature allows exporting the documents previously downloaded or imported into OSINT to a unique XML file or several XML files in plain text.

- **Category Matching module.** This involved the development and testing of the core module and the user interface. This module allows classifying or grouping documents according to a field of interest (a category or alert). I developed a particular Domain-Specific Language (DSL) to allow users to define a category or alert in an easy way. Help documentation about how to use this new feature was provided in the support web site:
  - Tutorial: “Using the Category Browser view” ([link](https://wiki.emn4.eu))
  - Task: “Creating a Category Definition File” ([link](https://wiki.emn4.eu))
  - Task: “Defining categories or alerts” ([link](https://wiki.emn4.eu))
All these developments along with other performance improvements and bug fixes have resulted in the release of several versions of the OSINT Suite during this year (from 2.2.0 to 2.2.5). In addition, I have been working on adding and improving the general online documentation in the support website. I contributed to the OSINT Workshop held in September 2013 by providing new on-site tutorials and improving the previous online documentation with new screenshots.
Regarding the research work related to the multi-lingual information text mining and social media project, I co-participated in several conferences by carrying out different scientific works. Specifically, I processed the training and test corpora of the Workshop on Sentiment Analysis at SEPLN (TASS 2013). This involved the programming of scripts to clean and extract the lemmas of tweets written in Spanish language. We participated in the task called “Sentiment Analysis at global level” which consisted in classifying tweets into 3 or 5 classes (from positive to negative). We participated with a system designed for English and adapted to Spanish, by employing in-house built dictionaries and machine-translated data for training. Additionally, we tested the manner in which four separate classifiers could be used in cascade to determine the sentiment. Although this was our first participation, the results obtained were promising as shown in Table 1 and 2 for all the experiments carried out. Precision (P), recall (R) and F1-measure were used to evaluate the classification results obtained by the systems.

<table>
<thead>
<tr>
<th>Experiment Id.</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRC-tuek-train-base-DICT</td>
<td>0.612</td>
<td>0.612</td>
<td>0.612</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-SVM</td>
<td>0.608</td>
<td>0.603</td>
<td>0.608</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-DICT</td>
<td>0.607</td>
<td>0.607</td>
<td>0.607</td>
</tr>
<tr>
<td>JRC-tuek-train-lemma-DICT</td>
<td>0.599</td>
<td>0.599</td>
<td>0.599</td>
</tr>
<tr>
<td>JRC-tuek-train-lemma-SVM</td>
<td>0.599</td>
<td>0.599</td>
<td>0.599</td>
</tr>
<tr>
<td>JRC-tuek-train-base-SVM</td>
<td>0.597</td>
<td>0.597</td>
<td>0.597</td>
</tr>
<tr>
<td>JRC-scamvula-train-lemma-SVM</td>
<td>0.590</td>
<td>0.590</td>
<td>0.590</td>
</tr>
<tr>
<td>JRC-scamvula-train-base-SVM</td>
<td>0.585</td>
<td>0.585</td>
<td>0.585</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-4CLS</td>
<td>0.582</td>
<td>0.582</td>
<td>0.582</td>
</tr>
</tbody>
</table>

Table 1: JRC 3-way official results in TASS 2013

<table>
<thead>
<tr>
<th>Experiment Id.</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>JRC-tuek-train-base-DICT</td>
<td>0.519</td>
<td>0.519</td>
<td>0.519</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-SVM</td>
<td>0.515</td>
<td>0.515</td>
<td>0.515</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-DICT</td>
<td>0.507</td>
<td>0.507</td>
<td>0.507</td>
</tr>
<tr>
<td>JRC-tuek-train-lemma-DICT</td>
<td>0.505</td>
<td>0.505</td>
<td>0.505</td>
</tr>
<tr>
<td>JRC-tuek-train-lemma-SVM</td>
<td>0.501</td>
<td>0.501</td>
<td>0.501</td>
</tr>
<tr>
<td>JRC-tuek-train-lemma-DICT</td>
<td>0.497</td>
<td>0.497</td>
<td>0.497</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-4CLS</td>
<td>0.481</td>
<td>0.481</td>
<td>0.481</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-4CLS</td>
<td>0.477</td>
<td>0.477</td>
<td>0.477</td>
</tr>
<tr>
<td>JRC-tuek-train-lemmaStop-4CLS</td>
<td>0.477</td>
<td>0.477</td>
<td>0.477</td>
</tr>
</tbody>
</table>

Table 2: JRC 5-way official results in TASS 2013

4. Objectives of the Project, Main Results and Conclusions

Only for the final report, concerns the full duration of the project

5. Publications


- **Experiments using varying sizes and machine translated data for sentiment analysis in Twitter**, XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN), Madrid, Spain, 2013.