D1.4 SoMeDi Outlook

WP1 –Vision, architecture and data integration – T1.4 SoMeDi Vision and Content Evaluation

Delivery Date:	*Project Number*:	*Responsible partner*:
M37 - 31/12/2019	ITEA3 Call2 15011	HI iberia

DOCUMENT CONTRIBUTORS

Name	Company	Email
Elena Muelas Cano	HI Iberia Ingeniería y Proyectos	emuelas@hi-iberia.es
Raúl Santos de la Cámara	HI Iberia Ingeniería y Proyectos	rsantos@hi-iberia.es
Adrian Pasat	BEIA Consult	adrian.pasat@beia.ro
Raffaele Perini	TAIGER	raffaele.perini@taiger.com
J. Fernando Sánchez Rada	UPM	jfernando@dit.upm.es
Ersin Ergün	Turkcell	ersin.ergun@turkcell.com.tr
Olcay Gürsel Baltaoğlu	EVAM	gursel@evam.com
Ahmet Sever	Turkgen	ahmet.sever@turkgen.com.tr
M. Özgür Gungor	Semantik	ozgur.gungor@turkgen.com.tr

DOCUMENT HISTORY

Version	Date	Author	Description
0.1	27/06/2019	HIB	First ToC distribution requesting contributions presented at Bucharest plenary meeting.
0.6	01/12/2019	HIB	Detailed ToC incorporating changes after final review.
0.7	19/12/2019	BEIA	Contributions for the HR Use Case
0.8	08/01/2020	HIB	Contributions throughout the document, incorporation of text to the annexes.

0.9	09/01/2020	HIB, BEIA, UPM	Added Annexes A, B and C.
1.0	03/02/2020	Turkcell, EVAM, Semantik, Turkgen	Finished sections 4 and 7.

TABLE OF CONTENTS

1. INTRODUCTION

This deliverable is the final work undertaken in SoMeDi. It is a self-reflective overall conclusion to the project, a summary of the achievements and the retrospective analysis of the advancements and end results that have been achieved in the three years that the project has been running. Since SoMeDi hasn't been running in an isolation chamber, the context (technological but also social and business) that has been active during this period is also important and as such it will be briefly examined in its impact on SoMeDi. Finally, in the Annex of this document a summary of the technical approach chosen is provided to illustrate the technology that we have built to achieve the project objectives.

After this brief introduction in Chapter 1, the deliverable is organized as follows:

- In Chapter 2 we revisit the SoMeDi Manifesto that we delivered in one of our first deliverables, D1.1, released almost three years ago. The manifesto was the explicit statement of the intended results of the project so we can track how did we achieve this during the course of the execution.
- In Chapter 3 we provide a brief analysis of important tecnhnological, social and business stories that have occurred during the execution of the project (2016 to 2020) and we analyse how did we fare during the project to integrate them into our activities.
- In Chapter 4 we present the three main innovations that we believe are the most long lasting successes in the project: the usage of Romanian NLP to detect sentiment in text, which was improved over the state of the art, the usage of computer vision techniques to improve the results of social media monitoring for marketing and the usage of real-time social media analytics to directly influence business actions.
- In Chapters 5 through 7 we present a summary of the use case of choice for the project: social media analysis for marketing purposes (5), social media and sentiment analysis for recruitment purposes (6) and finally social media analysis for generating live business rules for next best business action calculation in the Telco industry (7). We present the final status and the prospective new lines brought forward by the participation in SoMeDi.
- Finally in Chapter 8 we provide a summary of the document and some overall conclusions that provide an end to the activities of the project.
- Since the above discussion is more general reader-friendly than dedicated to the nitty gritty of the technology, the main text is followed by annexes that provide shareable accounts of the technology produced in the project in topics such as architecture, NLP tools and computer vision approaches.

2. SoMeDi MANIFESTO – POST PROJECT ANALYSIS AND UPDATE

The project started in November 2016 with the usual recollection of general objectives and aims that is usual during the preparation stage of proposals. In order to better outline the approach in the project from the beginning, the first deliverable produced, at project month 3 was *D1.1 SoMeDi Vision*. This document contained a more detailed account of the current state of the art, planned innovations and preliminary approaches to the use cases. This was conveniently packaged as the **SoMeDi Manifesto**, which was a declaration of interests for the project that was done before any technical groundwork was started. In this section we revisit this piece and conclude how did we fare compared to this initial approach.

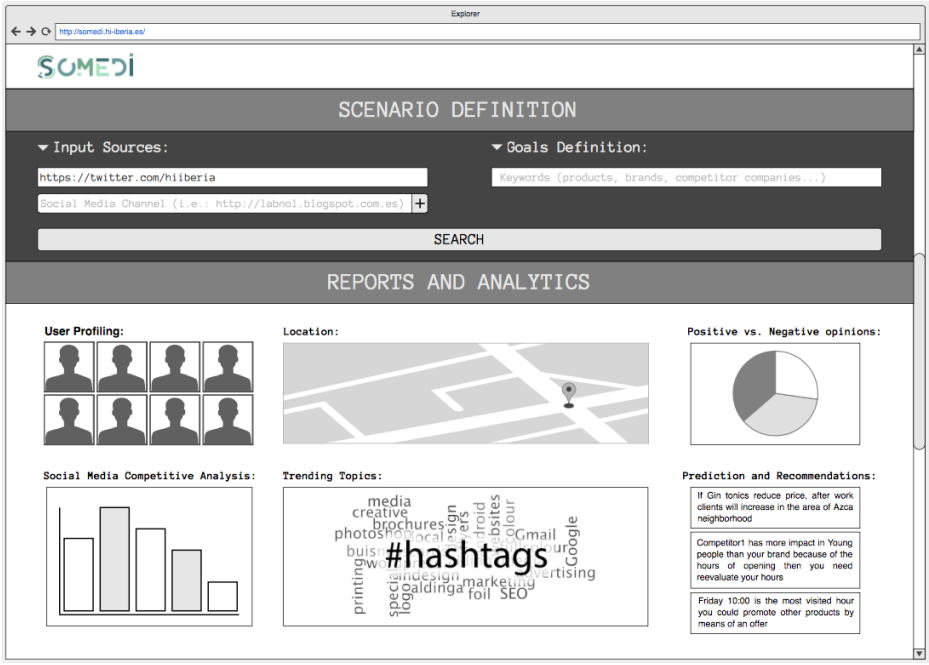


FIGURE 1 - SOMEDI MARKETING MOCK-UP DID

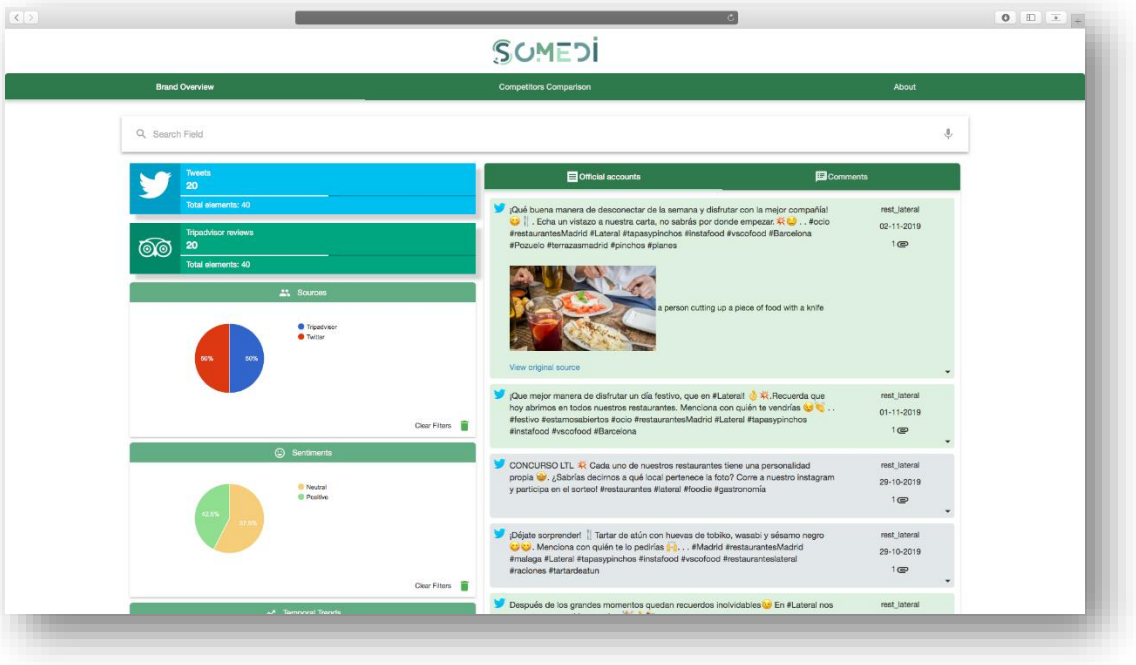


FIGURE 2 - SOMEDI MARKETING FINAL DID

We will do so by providing verbatim text from the original manifesto and then our commentary in light of the work in the project following these excerpts.

DID tool

> *The main output of SoMeDi project will be the DiD tool; […] SoMeDi DID tool main user interface is composed by a personalized dashboard that will allow the user to access to different functionalities:*

We see how the overall goal for SoMeDi was the DID (Digital Interaction Data) tool. This has remained a constant throughout the project, with use cases Marketing and Recruitment providing their own, personalized DID (visualization) tools as depicted in sections 5 and 6. The Next Best Action use case did not as it provided its results via a pre-existing commercial approach, the system in place at Turkcell.

> *Scenario definition: The scenario definition will allow to the user to determine the input sources to be analyzed as well as determine the parameters that want to be consider for the analysis:*

Input source section where based on a specific scenario you could add the sources from which obtain the data from social media like twitter, specific blogs, etc. The tool will collect and analyze the users' data as an input from different social media channels and fusion them.

This was fully achieved in the project. In Marketing, for example, we allowed users to seamlessly monitor results in Twitter, Facebook and TripAdvisor. Recruitment used its own means of collecting this data (connecting to recruitment platforms) and Next Best Action connected to Twitter and to custom bot networks to analyse texts generated by the users. Note that some sources mentioned (e.g., blogs) were not that relevant and as such not targeted by the end-product.

Target goals definition: Scenarios could be personalized by defining, for example, specific target goals like analyze specific products/brand ("I'd like to know the impact of the promotion including the Spanish omelet"); or search for other companies with similar profiles and compare its influence/presence in the social media channels (brand competitors analysis).

This targets were extremely important in the Next Best Action and Recruitment applications, as they were required for the operators to track the benefits of the system. In the Marketing use case the customers asked for a more *free-form* analysis, with the system just presenting aggregated data and the operator making their own jusdgements.

Output: Two main output formats are envisaged for the DID tool: reports and direct visualization in the dashboard through scenarios comparison and performance parameters. Reports generated will include:

Analysis of positive and negative opinions: to identify possible reputation damage or confirm a successful promotion. How the sentiments change on time for a specific promotion?

Completely achieved in all use cases and for all of the work languages in the project (English, Spanish, Romanian and Turkish) and with a breakthrough technology approach discussed in section 4.1 of this document.

Finding hot/trending topics: Which topics or themes are the main focuses of discussions? What are authors on the social Web talking about in terms of a brand or its product attributes? How do the topics of conversation differ from what the client would like authors to talk about?

Fully achieved in Marketing, with even custom categories added by the end users (e.g., restaurant products, service of waiters…).

User Profiling: demographic (age, gender, location, etc) study based on tweets and community relationships of the users.

Only high level was available and it was not emphasized due to concerns with the growing awareness of limitations imposed by GDPR.

Influential users: Identify potential influencers to promote some products or services.

Partially realized in the most relevant use case (Marketing) in which the most active users in the customer's network are pinpointed.

Social media competitive analysis: analysis of the publicly available social media data of a business and its competitors to gain perspective on their performance, identify weaknesses, find new opportunities and adjust their social media strategy.

Addressed in Marketing in which not only the main end user (LATERAL restaurant chain) but two of its local competitors (La Sureña, 100 montaditos and Morao Tapas) are tracked to extract information that can support the competition analysis,

Predictions and support automatic decision making processes based on the analysis performed. Recommendations based on analysis "Spanish omelet is highly demand on Fridays-> a promotion could be added to attract more clients"

This is the very basis of the Next Best Action use Case, in which a Telecom operators chooses their next 'move' in the dialogue with their customers based on the analysis of the data provided by SoMeDi.

Visualization of the main performance parameters and comparison of different scenarios. The visualization platform will be based on a backend developed with big data components to ensure search and retrieval facilities of social media assets and a frontend based on W3C Web Components standards, in order to enable its customization and foster the reusability of components.

The visualization components in Marketing are provided via the usage of UPMs Sefarad 3.0, which is an open source technology based on standards (W3C WebComponents).

SoMeDI will deploy a unique and common platform where the different use cases and applications envisaged in the project could be developed on top.

SoMeDi finally used a standard extensible architecture (documented in Deliverable D1.3) that enabled particular deployment of different analysis tools (modules of Digital Interaction Intelligence or DIIs) that required particular deployments. This was based around a flexible organization of the data workflows by using the very powerful open source orchestrator Luigi[1] and a number of industry standard technologies such as Elasticsearch[2]. The general architecture can then be fine tuned by using different configurations and tools to build dedicated complex applications like Marketing or Next Best Action.

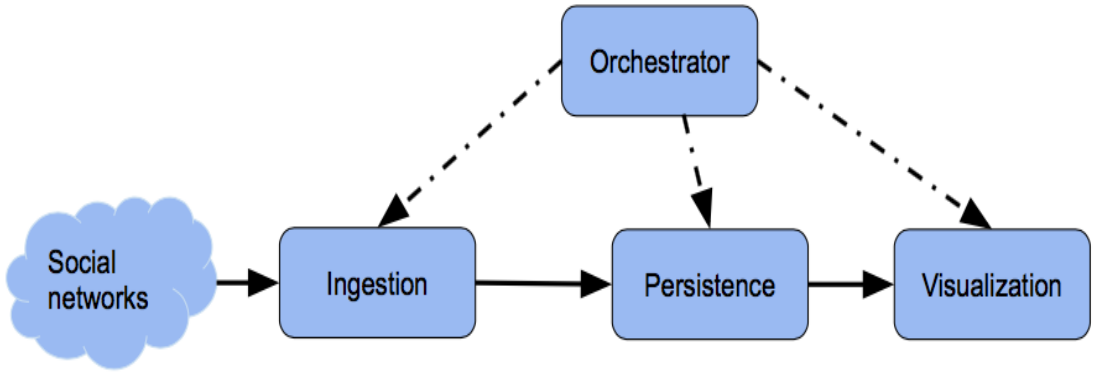


FIGURE 3 - HIGH LEVEL APPROACH OF THE DID ARCHITECTURE IN SOMEDI

> *The common architecture proposed for SoMeDi is structured in four layers: networking, presentation, controller and data access layer.*
>
> *On networking layer, we will use REST.*

This was the standard for all of the modules in SoMeDi, which were constructed as micro-services using a RESTful approach.

> *The view layer will be implemented by JavaScript, specifically with AngularisJS[…] The NoSQL database proposed, MongoDB, use also JSON documents in order to store records and offers facilities with the use of REST.*

[1] https://github.com/spotify/luigi - Spotify Luigi workflow orchestrator.

[2] https://www.elastic.co/ - Elasticsearch data repositories and indexing engine.

Here some changes were applied to the original plan to keep up with the technology. The view was mostly constructed using Sefarad 3.0 which provides a more high level model based on W3C standard WebComponents for a more integrated look and feel. For data storage we chose Elasticsearch which has steadily came to dominate the market in the data administration and warehouse market.

> *About controller layer, the server will be developed using Wildfly, an application server which implements JavaEE specification, characterized by its flexibility because it runs on multiple platforms. […] Looking at a micro services approach, technologies like Apache Web Server and PHP programming language can be used in order to build some of the services. Apache HTTP Server is the most popular web server on the internet since 1996, bringing a very long history of reliability and performance. […]*

Here more divergence was necessary to keep up with the times. Finally we used deployments based on Docker[3], which has became the industry go-to tool for simplifying the release and deployment to production of large web applications. As for servers, we still use Apache for some aspects but mostly have migrated to the rapidly rising open source development nginx[4].

[3] https://www.docker.com/ - Docker virtualization solution for deployment.

[4] https://nginx.org/ - nginx server.

3. STATE OF THE ART EVOLUTION DURING SOMEDI

The original idea for SoMeDi started to brew in September 2015 during the ITEA PO Days in Brussels. This means that at the end of the activities of the project, almost five years have elapsed, which is a considerable amount of time in the world of technology. In this chapter we pick some of the most relevant changes that have occurred in that time frame and briefly discuss how they have impacted the context in which SoMeDi has operated.

3.1. Changes in the landscape of social media

Since the main source of information to process in SoMeDi comes from the Social Media channels publicly available on the Internet, tracking the evolution of these over the time of the execution of the project has been a very important activity.

We see in statistics how the global usage of social media has continued on the rise during the execution of the project:

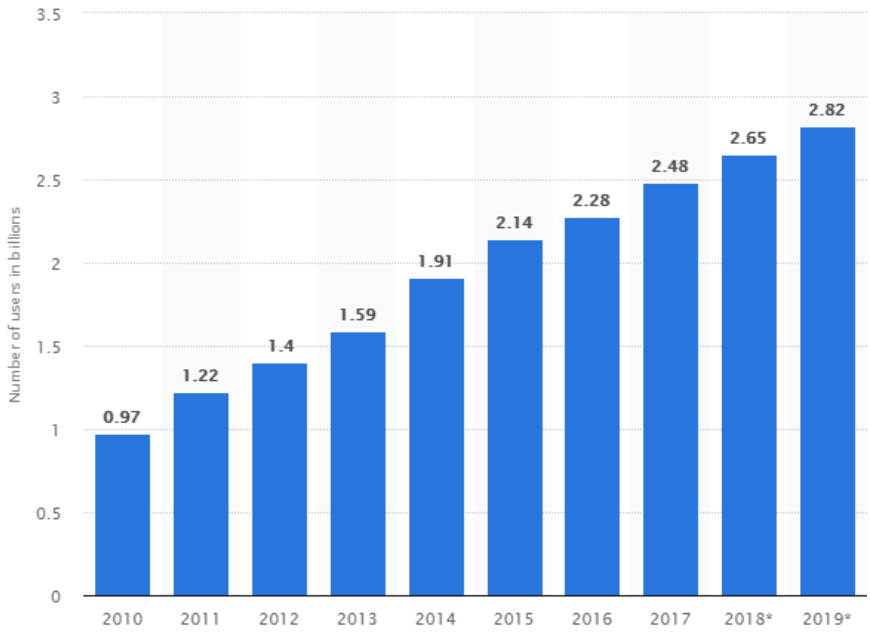


FIGURE 4 – GLOBAL SOCIAL MEDIA USERS 2010-2019

We see how in the period that we can consider SoMeDi's lifespan (2015 to 2019) the figure has grown from around 2 billion users to around 3 billion. This is strong growth but some

considerations apply. The rate of growth itself is slowing down which shows that eventually it will plateau, which is to be expected considering the finite number of humans on the planet.

What is more interesting is the evolution of platforms and capabilities offered. According to data by statista.com[5], the most used social media technologies as of October 2019 are as follows:

Facebook:	2.414.000.000 users (+6%)
YouTube:	2.000.000.000 users (+0%)
WhatsApp:	1.600.000.000 users (+0%)
Facebook Messenger:	1.300.000.000 users (+0%)
WeChat:	1.133.000.000 users (+2%)
Instagram:	1.000.000.000 users (+0%)

(figures of active [at least one usage per month] accounts, % variation from 2019 H1)

These results were periodically monitored by the project. Penetration of social media usage among users in the EU is slowly rising and now hovers slightly above 60% of the global internet usage[6]. Over 48% of these users have daily connections to social media so the interest for the target users of the Project and associated future products is still very high.

Breaking down the number of users per social media site, and removing the sites that are focused almost exclusively on person to person messaging (e.g., Whatsapp, Facebook Messenger) or those whose reach is focused regionally (e.g., QZone is almost exclusively used in China), we can see how the older social media sites (e.g., Twitter, Youtube, to some extent Facebook) remain stable or increasing, while networks that focus on multimedia sharing (e.g., Instagram) show strong increases and hence show the way to go for SoMeDi. This was confirmed by our potential customers in the project (as documented during 2018H2 in deliverable D4.3 and further confirmed in D4.3 v2 released at the end of the prioject) that try to focus for many of their offerings on Instagram which has over half its users (over 500M)

[5] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[6] https://www.statista.com/statistics/221377/internet-users-using-social-networks-since-2010-by-country/

checking the site daily and whose engagement for any given piece of content is +24% compared to the same contents in Facebook[7].

From the perspective of SoMeDi two key considerations emerged duing the course of the project:

- Twitter remains the gold standard for data access and text processing for data intensive systems such as SoMeDi. Thus, it is in our interest as a project to keep support for it, even with stagnant to declining user bases.
- The use of Instagram/Facebook and other Stories (ephemeral content that is automatically deleted 24 hours after its publication) was the highlight of the decade and the lifespan of SoMeDi, rising dramatically as well, with over 70% of younger (Gen Z) users consuming it daily and an increased likelihood of engagement – some users don't want to add permanent content to their storylines based on a promotion but may be willing to do so on a temporary basis. Tapping onto this data stream is extremely promising due to it being quite untapped other than by the social media operators themselves, but is not an easy task as it requires APIs that the operators might not provide and also requires a fluidity and immediateness in reaction that is difficulty to achieve by mass data processing systems such as SoMeDi. Baby steps were undertaken by the incorporation of computer vision on to the technology platform but this was a minimal effort due to lack of planning at the proposal stage.

3.2. Massive fake news surging.

FIGURE 5 – GOOGLE TRENDS EVOLUTION OF INTEREST IN "FAKE NEWS" 2015-2020

As is extremely evident in the Figure, the concept of "Fake News", which was a no-concept in 2015 at the start of SoMeDi, suddenly started growing in popularity, especially coinciding with the Donald Trump election campaign during summer and fall 2016. He was the real responsible of making the term popular, but beyond his usage of the term to categorize

[7] https://adespresso.com/blog/instagram-statistics - AdEspresso: Instagram statistics.

media which is critical on him, a greated undercurrent associated with the term emerged in the period associated to this: the deliberate usage of news elements which are purposefully lacking context or veracity to try and swing the opinion of the public, particularly in Social Media. This was powered by small but increasingly powerful companies of data analysts, the most infamous of what was Cambridge Analytica.

Cambridge Analytica was a British company which offered legal, grey area and illegal data mining and analysis to gather large quantity of data from users of social media, particularly Facebook. In March 2018, multiple media and news outlets broke news of the mismanagement of data by CA and its usage in recent major electoral processes throughout the world (e.g., the aforementioned US 2016 Election).

Today in the United States we have somewhere close to four or five thousand data points on every individual ... So we model the personality of every adult across the United States, some 230 million people.

— Alexander Nix, chief executive of Cambridge Analytica, October 2016 [8].

Cambridge Analytica sold wholesale the aggregated data of millions of users to advertising companies that were involved in the generation of direct advertising in their Facebook accounts as well as more unethical operators that controlled large networks of bots (computer generated users) that generated biased content to try and persuade the users of a number of things, ofter to support candidates which were ultimately paying for the advertisement campaigns.

All of this was done without the explicit knowledge of the users of the social media, whose data were used without informing of the final objective. In an even more worrying turn, it was revealed that there might be connections with parties and countries external to the elections (such as the Russian Interference in the 2016 elections in the USA, which is still under investigation).

Although the technology basis for all of this was long established (on the surface it is a powered-up version of similar algorithms used for decades in micro targeted advertising such as AdWords), it was the scale of it and the overall aim (influencing on national elections) which were the most shocking to the public. They definitely raised concerns on the usage of

[8] Cheshire, Tom (21 October 2016). "Behind the scenes at Donald Trump's UK digital war room". Sky News.

Social Media as a purely benevolent or neutral operator for your data[9] [10] and changes in how companies are expected to notify you of their intentions.

Cambridge Analytica quickly became the target of major criticism and that was the end of their business, notwithstanding the still ongoing legal battle for the actions in the elections, which are still far from finished. For SoMeDi, the repercusions of this have been limited, because the majority of our work doesn't concern business with activities as dramatic as shifting their users' political opinions (e..g., Marketing deals with their usage of restaurants while Next Best Action is all about their purchases of data plans). However, the changes in the public perception are very profound and a new attitude of defensiveness against this mass user profiling is in place. This should be taken with extreme care in a subsequent industrialization of the results, and the tool to certify this would be the GDPR, as we will analyse in the next subsection.

3.3. Advent of GDPR.

Even before Cambridge Analytica raised to the general public the problems, the European Union was working on a solution, or at least a legal framework to defined what is our data and what can we do to track its usage and consent or not to its processing and manipulation.

The GDPR[11] or General Data Protection Regulation was devised as a general norm for coutries of the EU to unify their prior Data Protection laws and, as a bonus, to update them to a new standard: for example in Spain, the law for data protection was the *Ley Orgánica de Protección de Datos 15/1999*[12] which as the name implies was 19 years old by the time GDPR was put into effect and as such, didn't regulate much of the data protection issues that are necessary on a modern social media site such as Facebook.

Thus, at the start of the 2010s it was decided that a new piece of EU-wide legislation was required and by the 25th of January 2012 the first approach was stared to discuss in the EU institutions, finally coming to fruition on April 2016 by consensus in the European Parliament

---

[9] https://optus.agency/articles/after-cambridge-analytica-who-can-you-trust/ - "After Facebook/Cambridge Analytica – Who do you trust with your data?", Optus Digital.

[10] https://www.theatlantic.com/technology/archive/2018/06/did-cambridge-analytica-actually-change-facebook-users-behavior/562154/ - People Are Changing the Way They Use Social Media, The Atlantic, October 2018.

[11] https://eur-lex.europa.eu/eli/reg/2016/679/oj - General Data Protection Regulation full text.

[12] https://www.boe.es/buscar/doc.php?id=BOE-A-1999-23750 Organic Data Protection Law from 1999, in use in Spain until the enforcement of GDPR in May 2018.

and the Council of Europe. A transitional phase started in May 2016 for countries that wished to have a smooth change from local law, until on May 25th 2018 the law became the top level legislation on the matters covered for the entire EU (later on the EEA) rendering obsolete prior local law.

The impact for products such as SoMeDi is substantial and the full analysis would require a very long in-depth analysis that falls beyond what is possible in this document. Unfortunately this wasn't built into the work plan of the project at the proposal stage (which occurred in winter 2016 before the adoption of the final text), so even when SoMeDi has tried to be compliant with the spirit of GDPR this has been done only partially because no resources were allotted to this. But the most relevant areas where this would impact the system would be:

- **Roles and responsible figures:** systems such as SoMeDi are required to signal persons that can be hold responsible as Data Protection Officer for the whole operation. This post was assumed by Mr. Raúl Santos as Project Coordinator for the duration of the project.
- **Rights of the data subject**: SoMeDi would be required to be transparent upon (a) which data is collecting from users (b) how this can be consulted and accessed, (c) rectified and (d) marked by the user as out of bounds for particular usages such as marketing. This is usually enforced in websites by means of not only terms of use display available to the user but also with controls detailing elements such as the web cookies selection and usage.
- **Processing of the data**: critically for SoMeDi, this requires the system to record any sensitive information in a *pseudoanonymized* manner that can't be traced back to individual persons. Also, all processing activities should be recorded and traceable, and for this SoMeDi provides extensive logs of the analysis undertaken in the DII tools. Finally, and very importantly any data processor such as SoMeDi should have means in place to notify without delay any breaches in the confidentiality of any of the personal data that it might be recording.

3.4. Data breaches in major companies.

Connecting with the last point in the previous subsection regarding data breaches, this concept has hit the headlines with increasing regularity since the beginning of the project. Since the start of the proposal some of the most serious have been as follows:

2015	British operator TalkTalk	> 4 million customers
	Ashley Madison	> 37 million customers
	Office of Personnel Management of the U.S. → 22.1 million customers	> 22.1 million customers

2016	FBI staff files → more than 20k sensitive profiles	20k security sensitive profiles
	"Panama Papers" scandal with many political officers affected (Iceland PM, Spanish former government officials)	Thousands of financially sensitive profiles.
	Yahoo	Initially Yahoo reported >500 million accounts in 2014, later reported to be its full customer base (>3 billion accounts)
2017	Equifax	>145,500,000 consumer records from the US, UK and Canada, the largest known data breach in history at the time.
	United States-South Korea classified military documents.	>235 gigabytes of military documents including wartme contingency and strategy plans.
2018	Meldown and Spectre vulnerabilities exposed	>1.7 billion affected devices, potential data breaches could occur.
	Facebook and Cambridge Analytica data scandal.	Detailed profiles of more than 230 milllion persons in US, UK and Canada sold illegally.
	MyFitnessPal	>150 million user accounts.
	SingHealth →	>1.5 million health records in Singapore.
2019	Democratic Senatorial Campaign Committee	6.2 million email accounts stolen.
	Ecuador national citizen records	>17 millilon records stolen.

From these breaches, which are just the visible tip of the iceberg, one can have a good impression of the vested interests that companies have in protecting their data – almost as high as the hackers in accessing said data. As we move to an ever more data-centric not only economy, but also health systems, financial sector and political systems, such large breaches of information security are a means of acquiring power by controlling the information. And

so, actors that collect large bodies of information such as SoMeDi have a responsibility on ensuring that data is as impregnable as technically possible.

In that regard, and also helping positioning SoMeDi from a GDPR standpoint, the toolbox and architectural elements in SoMeDi have been designed as to encourage self-hosting solutions. This not only obviates the problems arisiung by using cloud infrastructure provided by the major non-EU players such as Google, Amazon and Microsoft, but also imply that, ultimatelty, the securization of the data falls fully in the hands of the provider of SoMeDi service. This is a double-edged sword: as much as it makes it more self-contained it makes for a system less vulnerable to commonplace exploits of the infrastructure of these very same large cloud providers.

3.5. Generalization of AI, blockchain.

When SoMeDi started its activities in 2015, artificial intelligence was slowly on the rise in the academia. But it was around 2015 when it came to a tipping point in which the usually complex technology was supplemented with libraries in the most commonplace computer programming languages such as Python or ones that were quite high level such as Lua. This opened the floodgates to the massive uptake of AI in companies all over the world.

When the SoMeDi proposal was written, the topic was still growing slowly. As such, it is mentioned around 10 times in the whole FPP and wasn't the core technology for any of the participants. But in just some years, it has quickly become the byword for technology R&D with every major field applying concepts of AI their business. Now all of the companies participating in SoMeDi list AI as one of its main technological axes. And subsequently, some problems that at the outset of SoMeDi were planned to be tackled using classical approaches, such as rule-based sentiment analytics, were quickly migrated to more modern alternatives based on AI. Some results of this are further explained in secion 4.1 of this document and on deliverables D3.2/D3.3. Deep Learning algorithms for computer vision (using networks such as DenseCap for making text descriptions of images and frames in video) were also used although they were not planned at the beginning of the project. This was a feature that proved quite innovative for Marketing analytics.

The other rising axis of new technology during the course of the project was Blockchain and similar cryptographic and traceability ensuring mechanisms using distributed encrypted ledgers. This was almost unheard of by the start of the project outside of very limited academic domains but quickly spiked during 2017 and 2018 to be one of technology (and finance) hotspots. SoMeDi did not build anything based on this technology despite some expertise available in the consortium (by BEIA Consult) but it is surely one of the tools for the near future for these systems: one can easily see how fostering the securization of transactions will come into play in applications for our use cases in Marketing, Recruiting and Next Best Action.

4. INNOVATIONS IN SOMEDI

4.1. Romanian language sentiment analysis (TAIGER/BEIA)

Opinions are central to almost all human activities because they are key influencers of our behaviors. With the explosive growth of social media (e.g. reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making. Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products services, organizations, individuals, issues, evens, topics, and their attributes

In the last year of the project, BEIA and TAIGER collaborated intensively, and have trained a Romanian sentiment analyzer.

The Sentiment Analysis tool deployed within the Recruitment Use Case was based on the Google Translation API, thus the Romanian content provided in the Recruitment platform was firstly translated from Romanian in to English, and the achieved translation was analyzed through the Sentiment Analysis engines. So, we approached the idea of advancing a solution which would not require the step of translation from Romanian to English. In order to train a text classifier, a labelled dataset is required. However, there were not publicly available Romanian sentiment datasets, subset of IMDB sentiment dataset (English) was first machine translated using Google Translate, then manually corrected by humans. It was a time consuming process requiring about 200 human hours for in total of 720 labeled examples.

Since it took great amount of time to label datasets, the number of datasets were limited. It is shown that the number of training examples required are greatly reduced by using pretrained language models, either by feature extraction[13] or by fine tuning[14] a trained model rather than training a model from scratch. However, still large enough data sets are required[15]. Therefore, when we have used only Romanian dataset to the Romanian language model, the trained model couldn't generalize very well to the new dataset that was not from

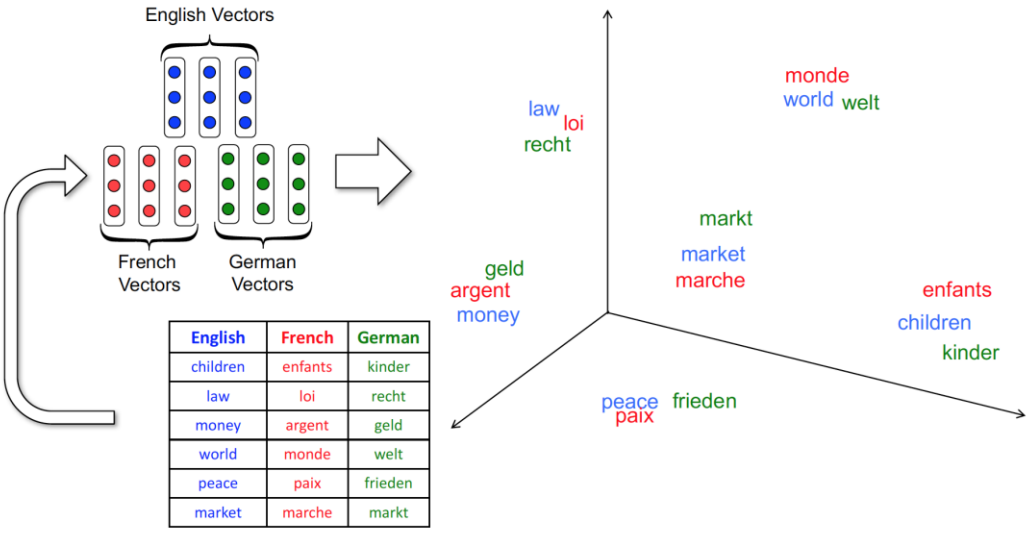
[13] Akbik, A., Blythe, D. & Vollgraf, R., 2018. Contextual String Embeddings for Sequence Labeling. 27th International Conference on Computational Linguistics

[14] Devlin, J. , Chang, M-W., Lee, K. , Toutanova, K., BERT: Pre-training of deep-bidirectional transformers for language understanding

[15] Howard, J., Ruder, S., Universal Language model fine-tuning for text classification

IMDB. Which could have been due to IMDB dataset having a specific profile such that the text was usually more than 5 sentences for each instance.

Therefore we have used a multilanguage embeddings approach which embed multiple languages into a single embedding. Crosslingual or multilingual word embeddings enable us to compare the meaning of words across languages, and enable model transfer between languages[16]. With multilanguage embeddings, there is a single encoder that can handle multiple languages, so that semantically similar sentences in different languages are close in the resulting embedding space[17]. (See Figure below).



English	French	German
children	enfants	kinder
law	loi	recht
money	argent	geld
world	monde	welt
peace	paix	frieden
market	marche	markt

FIGURE 6. MULTI LANGUAGE EMBEDDING (TAKEN FROM HERE, WHICH WAS GENERATED BASED ON AMMAR ET AL. (2016) AND SMITH ET AL. (2017)

We have trained multilingual sentiment analyzer, on top a Multilingual embedding (Multilingual Cased Bert embedding) which is a multi-layer bidirectional Transformer based encoder, that has shown to give state of the art results on multiple NLP tasks.

[16] Ruder, S., Vulic, I., Sogaard, A., A survey of cross-lingual word embeddings models

[17] Artetxe, M., Schwenk, H., Massively multilingual sentence embeddings for zero-short cross-lingual transfer and beyond

Training set used were from 3 languages, namely English, Romanian and Spanish. For English and Spanish we have used twitter sentiment analyzer data[18] [19]. The size of the dataset used for training was 989203 (10026 (ESP) + 978617 (ENG) + 560 (RO)).

In the table shown below we present a set of example sentences from the three languages and responses from the trained model.

Example sentences	In English	Response from the model	Language
This movie is great. I would like to watch again.	This movie is great. I would like to watch again.	P (0.92)	EN
No me gusta este restaurante. No voy a volver.	I don't like this restaurant. I won't return.	N (0.88)	ES
Sunt foarte fericit pentru tine. Lumea este încă un loc minunat.	I am very happy for you. The world is still a great place.	P (1.00)	RO
I would say it is a so so.	I would say it is a so so.	P (0.54)	EN

As for the future work, zero-shot learning ability of the model could be tested[20]. That is, whether in 101 languages where the task is not explicitly trained on, sentiment classification is possible. Due to the time constraint, we haven't tested this feature extensively.

[18] http://www.sepln.org/workshops/tass/tass_data/download.php?auth=NtQapsDsq45eTvZeZry

[19] https://www.cs.york.ac.uk/semeval-2013/

[20] Artetxe, M., Schwenk, H.Massively multilingual sentence embeddings for zero-short cross-lingual transfer and beyond

The achieved Romanian sentiment analyzer was deployed as a web service (using Flask) and its endpoint was integrated in the Sentiment Analysis Microservice. Thus, we achieved an unified tool which allows sentiment analysis in multiple languages by running multiple engines – Google, Azure, Standord NLP. The SA Microservice is implemented as a node.js application and can be consumed as a JSON REST API.

4.2. Social media analytics including deep learning image recognition (HIB)

At the start of the project, the planned analytic tools for the data captured in the social media were mainly focused on two aspects of data: the textual dimension (that is, text that can be retrieved from the media and analysed using text processing and/or Natural Text Processing Toolkits) and the interaction processing (e.g., analysing sequences of actions such as clicking on particular elements of a UI).

However, when we built our first version of the Marketing use case dashboard by the time of the project review for its first year, it was apparent that in our domain of activity as dictated by our end-user (that is, marketing for restaurants and restaurant chains) there was an enormous quantity of data not only in the text, but also in images and videos shared from the premises by the customers. This also coincided with the major trend detected in social media landscape (see section 3.1 of this document) in which the most rapidly growing networks were Instagram and others related to multimedia sharing.

Thus, it was decided in accordance with the end-users to devote resources to the integration of a functional protoype image description engine that took images from the timeline of the social media accounts and processed them so they could be converted to text that could be further processed by the rest of the SoMeDi NLP DII modules.

We proposed that for SoMeDi we would usa the following resources which we will now examine in detail:

- Deep Learning Models
- Deep Learning Frameworks: Torch
- Objects Datasets: Imagenet[21], VisualGenome[22], MS-COCO[23]

[21] "ImageNet - Wikipedia." https://en.wikipedia.org/wiki/ImageNet. Accessed 2 Jan. 2020.

[22] "VisualGenome." http://visualgenome.org/. Accessed 2 Jan. 2020.

[23] "MS-Coco." http://mscoco.org/. Accessed 2 Jan. 2020.

In the final system we use two DL meta-architectures working in parallel. These DL meta-architectures, ResNet and DenseCap, are specialized in different tasks.

ResNet[24] models is trained in 1000 labels (objects) using ImageNet dataset. ResNet models get very accurate results in object recognition tasks. These models have very good relation accuracy vs. speed. **DenseCap**[25] model task is to describe images in natural language. DenseCap identifies isolated objects and group of objects like one entity. DenseCap is trained/validated on the Visual Genome[26] dataset, which comprises 94,000 images and 4,100,000 region-grounded captions. The architecture is composed of a CNN[27], a dense localization layer and an RNN[28].

In the use case for Marketing the overall architecture is like follows:

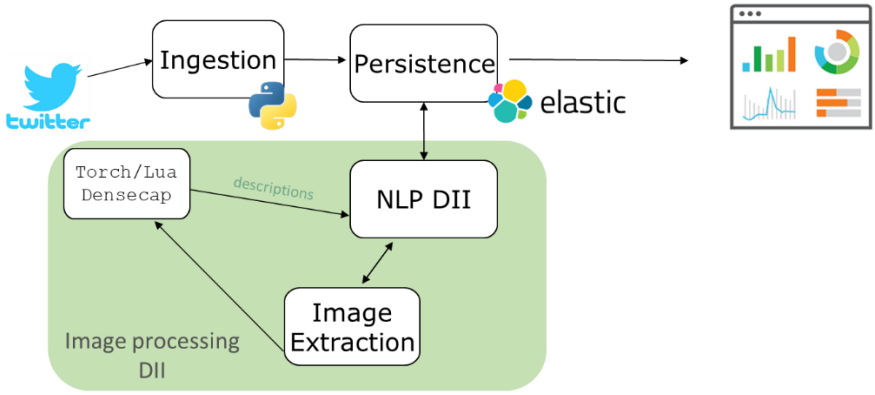


FIGURE 7. MARKETING ARCHITECTURE WITH COMPUTER VISION MODULES

[24] "[1512.03385] Deep Residual Learning for Image ... - at www.arxiv.org.." 10 Dec. 2015, https://arxiv.org/abs/1512.03385. Accessed 2 Jun. 2017.

[25] "DenseCap: Fully Convolutional Localization Networks for Dense ...." 24 Nov. 2015, https://arxiv.org/abs/1511.07571. Accessed 2 Jun. 2017.

[26] "VisualGenome." http://visualgenome.org/. Accessed 2 Jun. 2017.

[27] "Chapter 9: Convolutional Networks - Commonlounge." https://www.commonlounge.com/discussion/3040dd68b6454695a63cbfbb78d2e557/post. Accessed 2 Jun. 2017.

[28] "Deep Learning Book." http://www.deeplearningbook.org/. Accessed 2 Jun. 2017.

The internal architecture for the extraction and metadata DII module embedded in the green box is as follows:

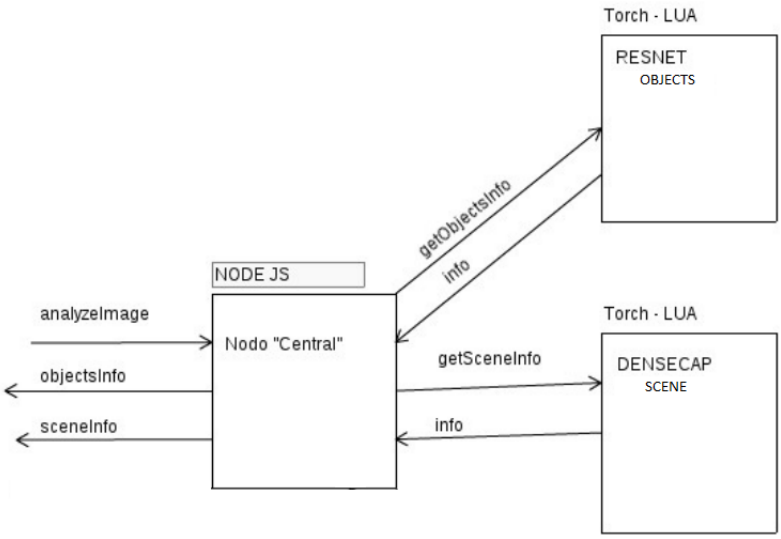


FIGURE 8. PROPOSED ARCHITECTURE FOR THE IMAGE METADATA EXTRACTION

As we can see, internally it is comprised of two distinct analysis modules. One is based on the RESNET network and is used for recognition of objects of interest and the other is based on the DENSECAP network and is used to provide scene description of images. As it can be seen in the Figure, we have divided the analysis of the images into 4 parts:

The **"central" node**, which receives the messages (images scraped from the Social Media) to analyze and is responsible for distributing according to the type of analysis requested by the source.

The **analysis nodes** (according to the information to be obtained from each frame): RESNET [objects in the image], DENSECAP [global description of the image and the different parts of it]. The "central" node is written in Node.js, and only performs message forwarding functions of both the part of middleware towards deep learning, as well as the part of deep learning towards middleware. The RESNET part uses the Torch library to perform all the analysis and is written in LUA. For each message it receives, it generates a list of objects that appear in the image, ordering this list according to the preponderance of the same in the scene. We always show the 5 first, but this value can be changed to show some more, although in this case it would increase the number of mistakes and errors in the results. The DENSECAP network, like RESNET, uses Torch and is written in LUA. With it we get descriptions of the scenes. This network divides the image into as many pieces as indicated (we usually work with 50) and analyzes the image in a generic way and each one of those (50) sub-images to give a description.

With the help of these components, our system runs quite efficiently in our test servers with ample headroom to accommodate connection to intense streams coming from social media for the Use Cases of the project.

These results are only a starting point. In conversations with the end-user (LATERAL) it was suggested that more advanced functionality could be used (e.g., recognition of number of people and particular restaurant in group photos, identification of particular dishes in photos) but these had to be left out of the activities in SoMeDi given the limited resources in the project. They are now potential functionalities for future derived products.

4.3. Integration of Next Best Action algorithms with chatbots and reductions in time for management

Next Best Action algorithm that we used aims to offer products to customers by using their sentiments about products and their packages. In order to do this, we used Somedi chatbot and also customer comments on the facebook page. With these comments we are able to understand the customers' satisfaction about their and other products/packages. After the customers are commented in chatbot/facebook page, this will be sent to the our flow with request, the sentiment value about this product is easily saved to our Business Moment Cache.

This cache is saving these sentiment values by actor id. So when the customer comes for predict a new product/package from chatbot with the same way, our flow is responding a suitable product list by using Business Moment cache. With this cache we are able to respond to customer in a short time. While deciding product list for customer, we are also checking the customers' current product/package and its price.

With this product for this scenario the company needs only two flows to give best offers to their customers. One of them is for training of customers sentiment values, another one is for prediction of customers offers. These flows can easily be draw in Evam tool. And they can also add validations to eliminate customers by using their features (age,segment,payment type ...).

For the train flow, the user should also create business moment cache for products. For the prediction flow to get offers for customers, GetSortedEligibleProduct action is need to be called and then the prediction system will be ready.

5. USE CASE 1: SOCIAL MEDIA FOR MARKETING PURPOSES

5.1. Summary of final version of the use case demonstration

This use case focused on social media for marketing purposes will be bases on three axes:

[1] Competitor Analysis based on Social Media,
[2] Brand monitoring (analyse the reputation of the brand) and
[3] Event detection with sentiment detection.

The main goal of this use case is to develop marketing strategies based on the insight collected and continuously analysing the impact of marketing campaigns, testing these strategies within the context of accelerating innovations, focusing especially on the needs of startup and SME companies. Principal partners involved in this use case are HI-Iberia and Taiger (with considerable help from subcontracted party Universidad Politécnica de Madrid).

In this use case, SoMeDi platform will be tested together with Lateral restaurants. Lateral is currently a customer of HI-Iberia as they are providing Management Software for the restaurant within other company department (POS, Warehouse management, CRM). In consequence HI-Iberia will be in close contact with them in order to present SoMeDi solution for its evaluation but also to enable the expansion of their current technology thanks to SoMeDi platform.

From the tools available in the project, the Marketing use case makes use of the following:

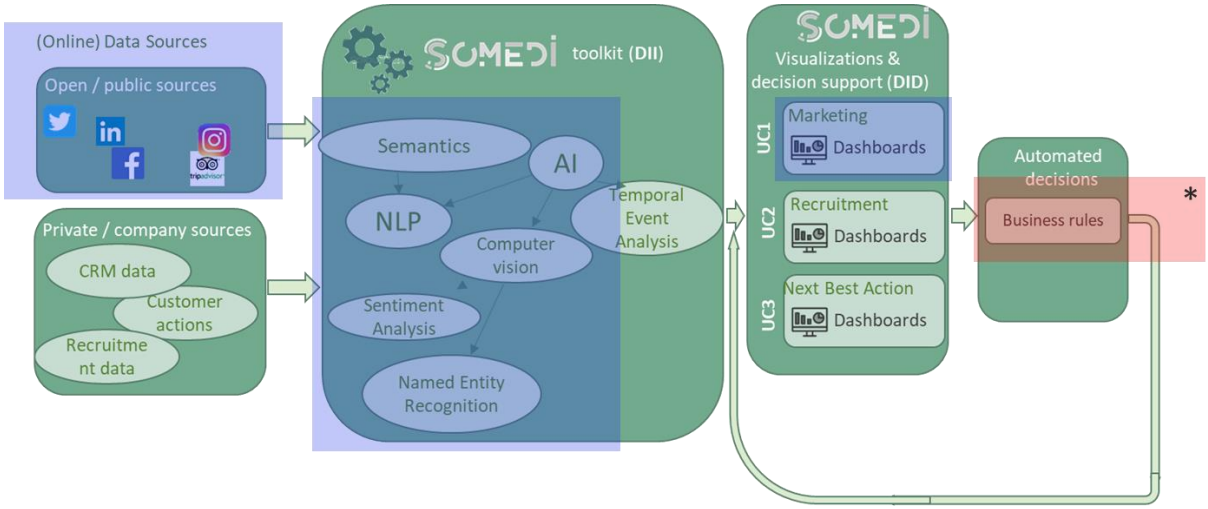


FIGURE 9. SOMEDI COMPONENTS USED IN USE CASE MARKETING

The final prototype enables users of the product (owners of restaurant chains and their marketing officials to monitor their social network activity. This is depicted in the following screenshots of the final prototype in action:

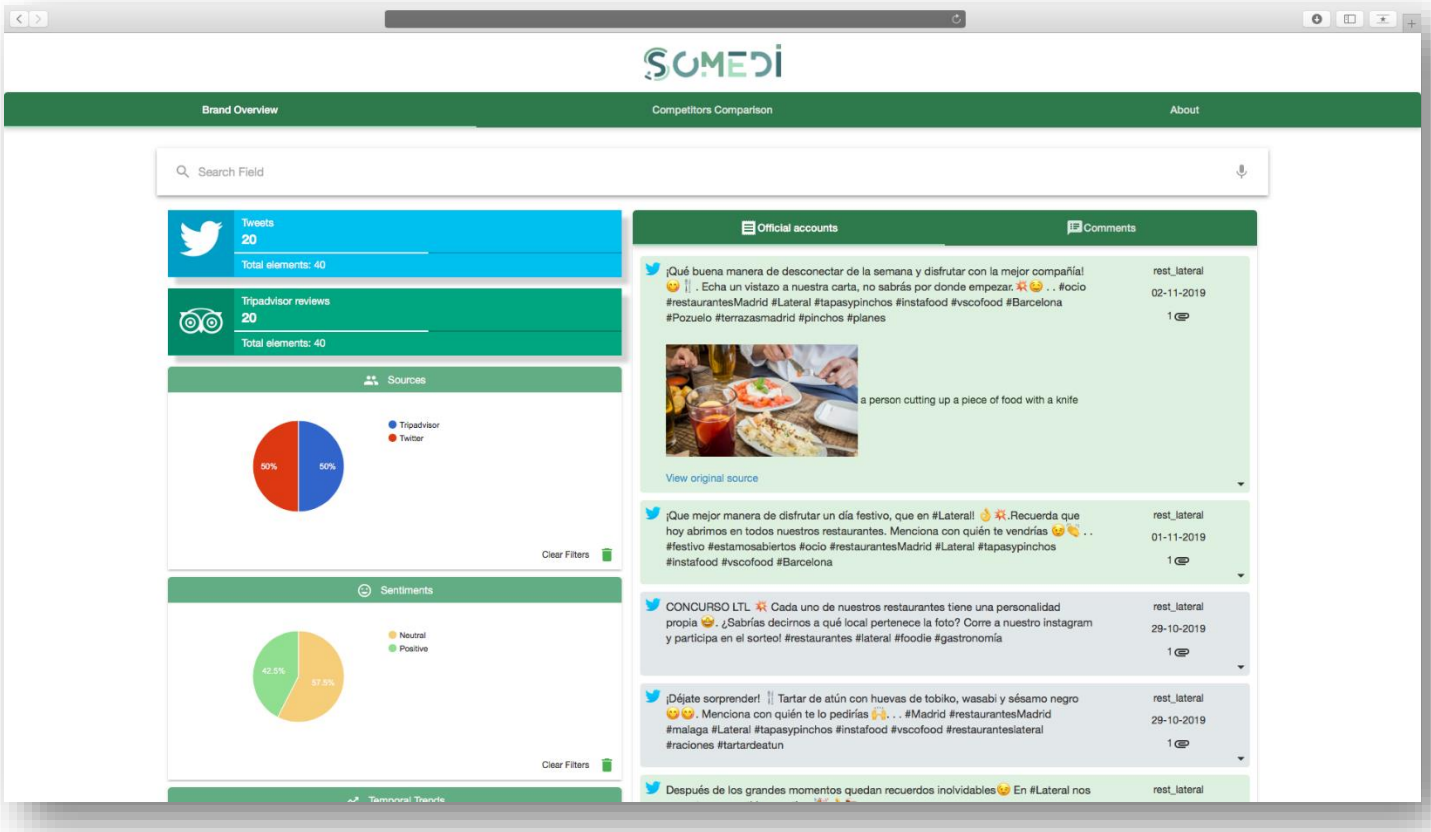


FIGURE 10. SCREENSHOT OF USE CASE MARKETING

We can see how the system provides monitoring of Twitter (a generalistic social media network) and TripAdvisor (which is more focused on the restaurant and hospitality business) The messages exchanged by users that are connected to the restaurant accounts of their competitors are then analysed for their sentiment, their addressed topics and the most active users in the extended social network.

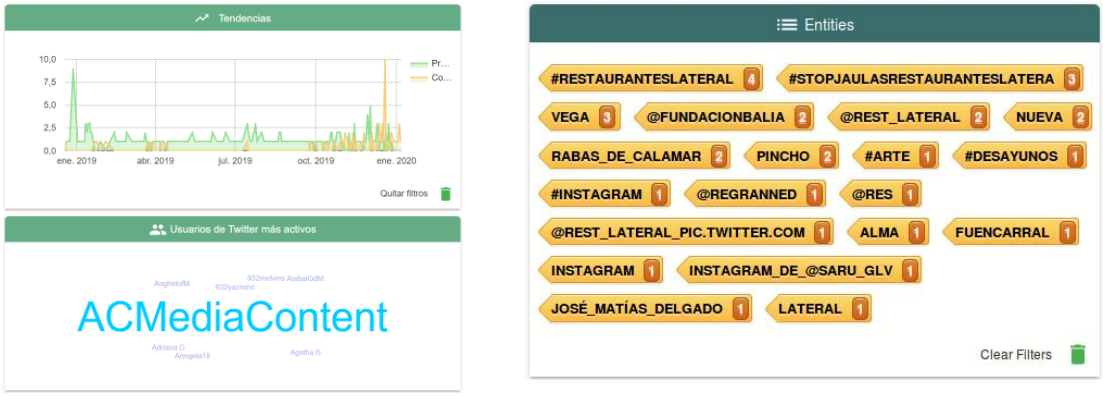


FIGURE 11. DETAILED ANALYTICS FROM USE CASE MARKETING: TRENDS AND MOST ACTIVE USER (LEFT) AND MOST ACTIVE TOPICS (RIGHT)

The system was co-designed hand in hand with a prospective end-user (the restaurant chain LATERAL with over 20 restaurants in operation in Spain) and the validation and evaluation was done also with them. This is an important alliance fro SoMeDi since they are not official participants in the consortium but have participated on their own expense.

5.2. Changes in related SotA since the delivery of D3.1 v2.

From the first delivery of the technology presented midterm during the project, the most important new aspects are as follows:

- Usage of TripAdvisor instead of Facebook for more domain-specific metrics for the restaurant business.
- Introduction of the computer vision capabilities. This and the associated SotA is presented in section 4.2 of this document.
- Usage of novel AI-based sentiment analyser, as described in section 4.1 of this document.
- Introduction of competition metrics. This makes easy for operators of the system to monitor and track the activity of other companies whose business may have relations with the operators' (restaurants in this case). This was done in SoMeDi by applying the same processing algorithms (sentiment, trend detection, etc.) as on the 'main' target.

5.3. Future lines and exploiation of the results

On January 2017, more than 200 million posts were tagged as *#food* and 23 million as *#drinks*, and food and beverage photos are easily some of the most popular types of content on Social Networks. It's very likely that your customers are posting with or without your

interaction. Besides, considering that 88% of people are influenced by reviews and online comments, having a social media strategy for your company is important.[29]

In consequence, it's clear that businesses (e-commerce) need to have a strong social media presence in order to drive more sales. However, listening in on the discussions surrounding your brand and products will help you better understand your buyer's behaviours and how they use social media to search for and share their experience with your product or services. This will allow the enterprise to focus its social media strategy where it matters most.

As solution SoMeDi, intends to support enterprises to improve their presence in social media by providing feedback extracted from social media data from clients. The overarching goal for the results of marketing use case in the future beyond SoMeDi would be to improve the marketing performance for restaurants that are using one of HI Iberia's products already (the Point of Sale system HIPPOS[30]). The marketing approach would be to package the SoMeDi functionality as an add-on to the base POS system, thereby providing a new functionality that differentiates it from the competition. The company will improve its marketing proposition by receiving personalized recommendation based on social impact information helping the business to modify their marketing campaign to align it with client needs, understand their brand impact and performance and in a whole help them to increase their impact and incomes.

The base approach used in SoMeDi (generic management of social media traffic and processing of messages) could be extended based on three main axes that could be exploited in the future:

1) Competitor Analysis based on Social Media.
2) Brand monitoring (analyse the reputation of the brand).
3) Event detection with sentiment detection.

The solution needs to support information from social media related to companies (restaurants and e-commerce) and provide the analysis details to them in an understandable and useful format.

[29] https://sproutsocial.com/insights/bars-restaurants-social-media-guide/ - Tips for social media management for restaurant domains.

[30] https://hippos.io/ - HIPPOS Point of Sale for restaurants website.

Particularly for the Marketing use case, the added value provided by SoMeDi is the provision of a tool able to monitor the activity of the enterprise profile in the social media and integrating very advanced features such as analysis of images with computer vision. Starting from the analysis of the evolution of mentions; the evolution of public sentiment; the impact of different topics; co-mentions; evolution of campaigns and promotions posts, the tools will recommend the enterprise with some practises to enhance the marketing performance of companies.

6. USE CASE 2: SOCIAL MEDIA FOR RECRUITING PURPOSES

6.1. Summary of final version of the use case demonstration

The release of the Internship Recruitment Demonstrator implied that the above-described NLP applications to be tested on as many candidates as possible, and so Digital Interaction Data will be created.

The strategy was to structure the DID as Metadata (DataBase) and then process it using Data Mining type Clustering and Text Analytics methods to find the following information / patterns:

- Company User Metrics (a set of visual instruments available for the company users' designed to improve the assessment of the internship campaign). This visual instruments will display the following statistics – the candidates age, level of education, field of study, work experience; also, after the candidates provide their text input regarding the company's fields of activity we will present certain tendencies – which field of activity is most attractive, etc.

- Student User Metrics – several reporting tools which will present the status of the user internship applications;

- Internship campaigns Metrics – these reporting tools will present the candidates feedback after the internship programme.

So, BEIA has launched its internship recruiting campaign for the next year and encouraged the students from the main Universities in Bucharest to register and apply to the company internship program. The figure presented below presents the answers collected (and analyzed through the SA microservice) from over 25 candidates.

The use of this visual and reporting instruments allows the staff of an HR department to reduce the time allocated for candidates screening, and also the time required to interview a large pool of candidates personally. By using the SoMeDi Internship Recruitment platform, its more accessible to select the candidates based on their field of study, specialization, and preferences regarding the company departments.

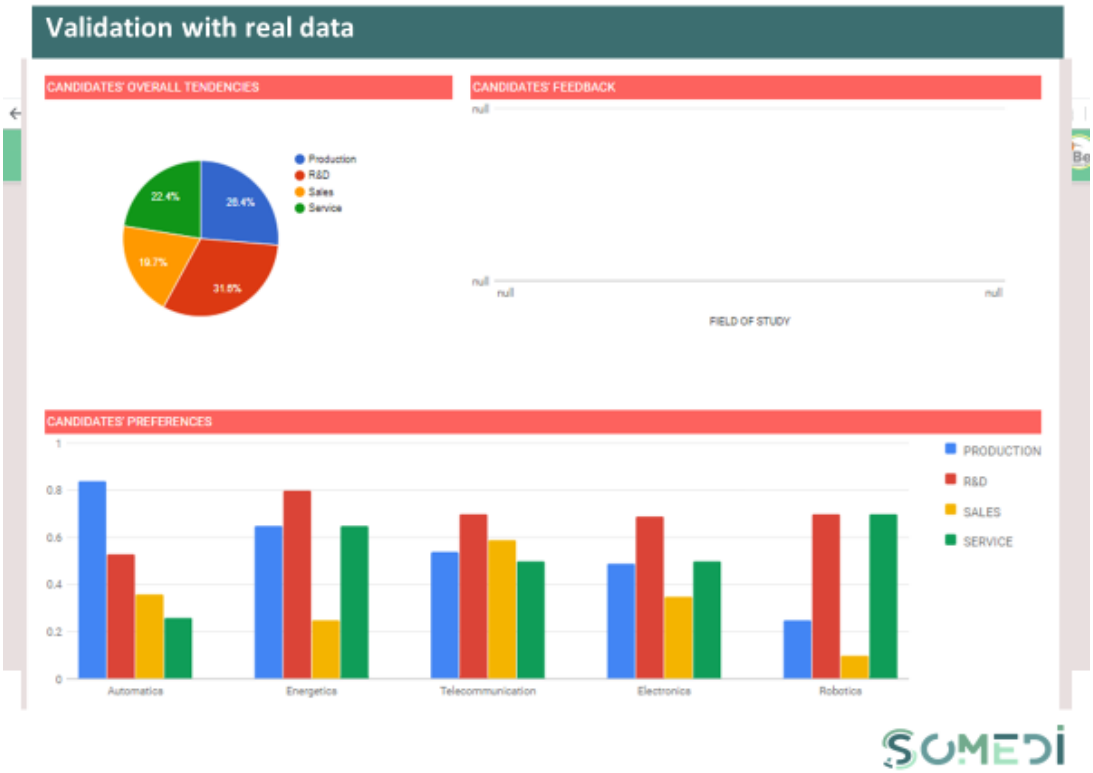


FIGURE 12. VISUAL INSTRUMENTS BASED ON THE SENTIMENT ANALYSIS MICROSERVICE

6.2. Changes in related SotA since the delivery of D3.1 v2.

Trends in NLP architectures

The first trends we can look at are based on the deep learning neural network architectures, which have been at the core of NLP research in recent years. We observed there is still significant doubt about what architectures can deliver the best results.

The SoMeDi consortium has tested, deployed, and integrated various architectures for NLP tasks in multiple business cases. Moreover, BEIA has performed a comparison of the main Sentiment Analysis engines available on the market, either paid or open-source. In conclusion, the trends for the NLP architectures in the near future regards the following aspects:

- Previous word embedding approaches are still important
- Recurrent Neural Networks (RNNs) are no longer an NLP standard architecture
- The Transformer will become the dominant NLP deep learning architecture

- Pre-trained models will develop more general linguistic skills
- Transfer learning will play more of a role
- Fine-tuning models will get easier
- BERT will transform the NLP application landscape
- Chatbots will benefit most from this phase on NLP innovation
- Zero-shot learning will become more effective.

6.3. Future lines and exploiation of the results

Both Beia and Siveco, have already started to use the platform for their yearly internship campaigns. Moreover, the exploitation perspective is encouraging due to the numerous collaboration with important universities countrywide – the University of Polytechnics in Bucharest, The Bucharest University of Economic Studies, etc.

Also, BEIA plans to exploit the SA microservice in specific markets, targeting three main domains – telecom, marketing, and banking. The exploitation strategy, in this case, is based on the easy integration of the SA microservice with third-party platforms, thus it can be easily linked to a call center channel or an online commerce platform.

7. USE CASE 3: NEXT BEST ACTION

The purpose of this section is to describe how to use the SoMeDi platform in terms of the functionalities provided to the end-users, as well as the roles defined by them.
We have two side in this project, one of them is getting information from customers by using social media (Facebook). The other part is giving opportunity to customers to ask best offer prediction by using ChatBot. These are our use cases:

Social Media - Data Bundle/Voice Bundle Opinion Declaration

1. User shares a text related with data/voice bundles from Facebook.
2. Then user's gsm information and facebook account id will be matched.
3. Sentiment value will be extracted from text by using NLP algorithm.This will find the text has positive or negative meaning.
4. Sentiment value is sent to Evam Engine for this user to train our flow somediTrain.

ChatBot best offer prediction for Data Bundle

- User enters the chatbot.
- When the user enters the chatBot, it asks gsmNo to the user. After entering gsmNo, user can ask for data bundle packages.The prediction event will be sent to predict flow and the engine will send response back to the ChatBot.

There are 4 cases for offer suggestions.

1. If there is no sentiment value information for this user's data bundle package, the lowest priced package which is more than the price of user's data package will be recommended.
2. If the sentiment value is equal to positive, the highest priced data package will be recommended.
3. If the sentiment value is equal to negative, the highest priced package which is lower than the price of user's data package will be recommended.
4. If the sentiment value of voice package is also negative, the promotion package will be recommended.

ChatBot best offer prediction for Voice Bundle

- User enters the chatbot.
- When the user enters the chatBot, it asks gsmNo to the user. After entering gsmNo, user can ask for data bundle packages.The prediction event will be sent to predict flow and the engine will send response back to the ChatBot.

There are 4 cases for offer suggestions.

1. If there is no sentiment value information for this user's voice bundle package, the lowest priced package which is more than the price of user's voice package will be recommended.
2. If the sentiment value is equal to positive, the highest priced voice package will be recommended.
3. If the sentiment value is equal to negative, the highest priced package which is lower than the price of user's voice package will be recommended.
4. If the sentiment value of data package is also negative, the promotion package will be recommended.

7.1. Summary of final version of the use case demonstration

The main purpose of this use case, to increase the marketing performance of telecommunication companies. The focus is on using customers social media sharing, comments and like, we will offer best suitable packages to customers by using SoMeDi's Chat Bot Tool.

Using SoMeDi's Chat Bot Tool, the platform will generate:

- Offering the most suitable data and voice package based on Social Media .
- Increase package sales
- Considering customer requests
- Targeted social activity
- Brand reputation increase
- Evaluating the positive and negative comments of customers on social media
- Maintaining customer satisfaction.

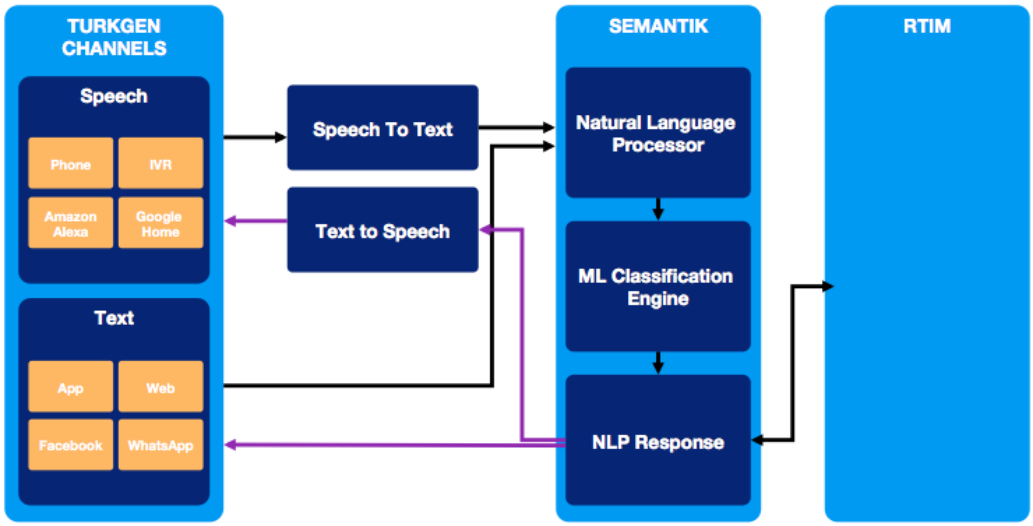
Main KPIs identified to assess the impact of SoMeDi in this use case are:

- Increase customers' access the impact
- Online sales increase
- To increase social activity with clients'

Semantik company focused on the commercialization of Chatbot in Telecom domain by smart product advice management on digital services / data on the social media specific for telecom sector.

Turkgen will work on development of an product as a SaaS platform for the analysis of social media data.

A combination of those two solutions are merged under the SOMEDI architecture.



1.1. Changes in related SotA since the delivery of D3.1 v2.

From Chatbot Perspective, the attention an chatbots are still increasind and its still in top pririorty in parallel with AI solutions. Text mining techniques are still being investigating in different academic and private institutions but the success rates increase modestly. Turkgen and Semantik continuously follow-ups the technological innovation in this domain and ready to test and implement new approaches as well as the algorithms.

In the digital world, with the increase of analytical solutions and customer data, customer satisfaction has become more important in companies. We aimed to increase customer satisfaction with personalized estimates of the most suitable packages and products for our customers. Today, with 3,5 billion internet users, following customer requests and needs from customer data, increasing customer satisfaction with analytical models has become one of our most important goals.

7.3 Future lines and exploiation of the results

We have gone for sales activities especially during the last year of the project. The list of the companies we have gone for selling interviews around the world are the following:

USA : T-Mobile (Telco)

Egypt : Citibank, Standart Chartered

Slovenia : Zagrebacka Bank

Nigeria : Group M (Telco), SmileComs (Telco)

Jordan : Umniah (Telco)

United Arab Emirates : Emirates NBD, Careem

Turkey : TEB, İş Bank, Akbank, YKB, Ziraat Bank, Garanti Bank

Within the scope of commercialization, Turkgen has carried out multiple trials with potential customers and academia in order to create business models and steps for early commercialization of the project. Several scenarios were developed together with these stakeholders based on classification algorithms. Turkgen wil continue to exploit of the SOMEDI experience as much as posible to position it as a producto in the Market.

Semantik works on prediction models and tries to capture the intention of the subscribers or users intention. Afterwards, Semantik proposes the next best action for a process or for a product / services. This must have high accuracy to benefit most the end users and the companies. Semantik implemented its solution to Turkgen's Chatbot solution and tried to measure the performance of the predication engine for the quarries that are related to banking operations.

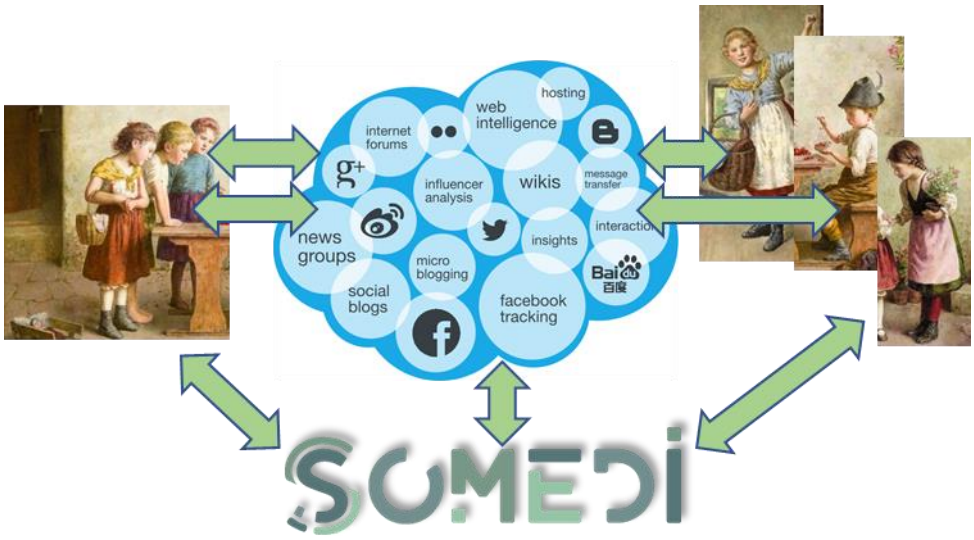
(Fintek)The measured success rate was above the expectations and now, it will be implemented as a demo to other two national banks as to understand and direct the bank account owners, so instead of calling the call-center, their requests will be operated in the chatbot system..

8. CONCLUSIONS

In this document we have provided a summary of the results of SoMeDi as a whole, studied the socio-economical context in which the project has evolved during its lifetime and provided hints of the possible directions that the field might be taking in the coming years. Every project is part of its era and SoMeDi will be remembered perhaps as one of the first approaches to produce a domain-specific set of tools and approaches with which companies in Europe can try to comprehend and analyse the massive data that is generated as part of the interaction with their customers.

We have provided a set of tools for information analytics that enabled end-users in three use cases (Marketing, Recruitment and Next Best Action) to improve their business and the quality of information they receive from their customers.

We expect that some results of SoMeDi will live on and that the relationship between companies and their customers gets even richer, more satisfying and more profitable for all sides with time and we have provided some tools to at least measure and improve these relationships.



ANNEX A - METHODOLOGY AND PRACTISE: FROM USER GENERATED & SOCIAL MEDIA DATA INTO DIGITAL INTERACTION INTELLIGENCE – ARCHITECTURAL SUMMARY

In this Annex we present some findings of SoMeDi in the field of the system architecture. Some of these were covered in past deliverables of the project, most notably D1.2, D4.2 and D1.3. Some information and technological descriptions that were deemed confidential for the partners in SoMeDi has been left out.

1 ARCHITECTURE OVERVIEW

The abstract architecture of SoMeDi is composed of two modules: DID and DII. Both modules are explained in the following sections.

1.1 DID overview

The DID architecture is based on an orchestrator that executes data pipelines. Figure 3 provides a high-level specification of the DID architecture, that consists of four main components.

These components are described below:

1. ***Ingestion***. Input interface of the platform. This software connects with sources of information (social media like Twitter and Facebook, news, websites, ...) to collect data into the platform.
2. ***Persistence***. A persistence level will be deployed in order to store the data collected and processed by the platform. In addition to these data, documents like dictionaries or taxonomies will be included in the storage system of the platform.
3. ***Visualization***. A visualization module will be used to obtain graphical representations of the data.

4. *Orchestrator*. Software that manages the execution of various platform modules in a pipeline. It monitors the travel of the data along the platform, coordinating the execution of the modules specified by the user.

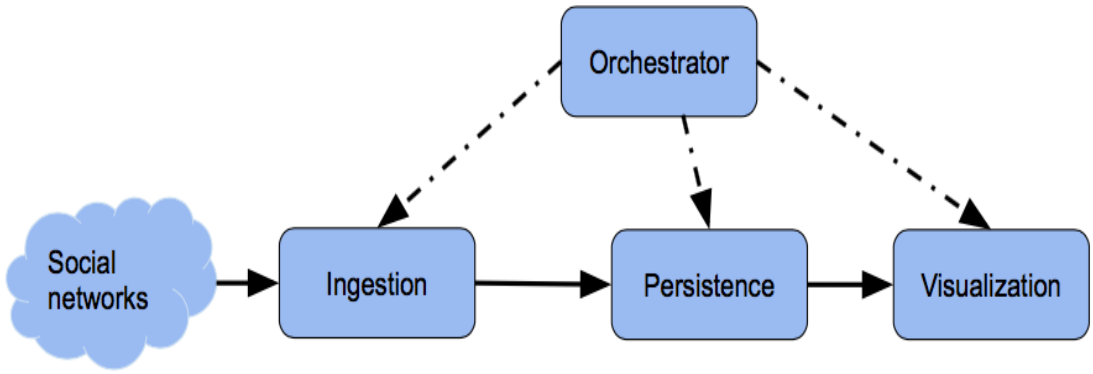


FIGURE 13 - HIGH LEVEL APPROACH OF THE DID ARCHITECTURE IN SOMEDI

1.2 DII overview

DII is the research domain that encompasses the application of advanced data mining techniques coming from advances in machine learning and artificial intelligence to extract information and value hidden in the digital information that is generated by humans in digital interactions with and true computer systems.

While each DII application project will have its unique domain specific modeling and parameterization, we can derive, capture and distill common guidance, heuristics, best practices and technology approaches that cross domain and application boundaries. The toolkit will integrate existing Artificial Intelligence and Machine learning libraries and services, thus leveraging the massive ongoing industry and academic investments taking place in this area. We will shape and build onto those foundations higher level instruments that can be leveraged by a wider community of service and product developers specifically focused on extracting meaning and value from human machine interactions.

With regards to DII architecture, it consists of a set of modules which provide a well-defined REST API that aims to return a semantic output expressed in JSON-LD. An overview of this architecture is shown in Figure 2.

Text intelligence is composed of tools and techniques for dealing with textual information, which can be the content of Social Media interactions, but also the textual parts of application data produced by the users. A multilingual-capable framework will be provided

that generalizes the capture of data and identifies a common set of resources (localized versions of WordNet, SentiWordNet and high level ontologies such as SUMO to homogenize the different resources).

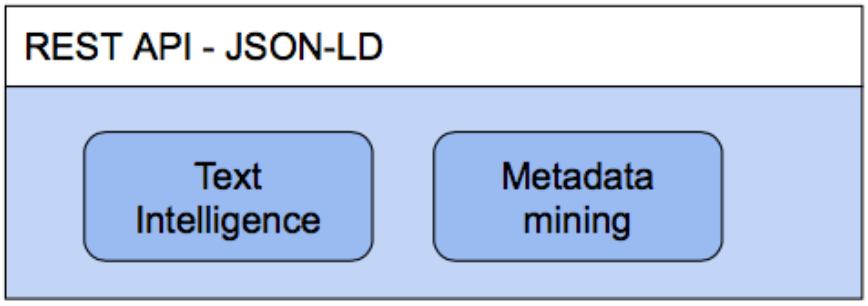


FIGURE 14 - HIGH LEVEL APPROACH TO THE DII ARCHITECTURE IN SOMEDI

Using this overall framework, several NLP toolchains will be proposed for the different working languages to enable deep syntactic, semantic and sentiment analyses. Throughout the design, Big Data processing strategies will be proposed (e.g., Apache Spark [1]) to enable the processing of large quantities of data in a scalable manner.

Textual data represents a significant source of information in digital interactions with and across applications; significant value can be mined from other types of data well. A first source is alternative content such as voice interactions and other media types. As an example of these we can think of the importance of voice calling data in call center applications of customer engagements. A second major source is the metadata accompanying the interactions, such as the patterns in timings, participants and the correlation between the interactions and external events. As an example of these consider mining recurring task patterns in a task management system. Projected onto vertical domain models of typical processes in a business segment, these can unearth business sector discovery, process discovery and participant roles unspecified and implicit in the conversation content itself.

The modules will be provided as Docker [2] modules, in order to guarantee the easy installation and connection between them. In addition, they can be executed in a distributed fashion, as well as in Big Data platforms such as Spark.

2 DID ARCHITECTURE

Once an overview of the architecture has been given, this section describes the whole architecture of the DID module, as well as its integration with DII. The integration between DID and DII is based on different pipelines depending on use cases. For example, Figure 3 and Figure 4 show the integration for pre-analysis pipeline and post-analysis pipeline respectively.

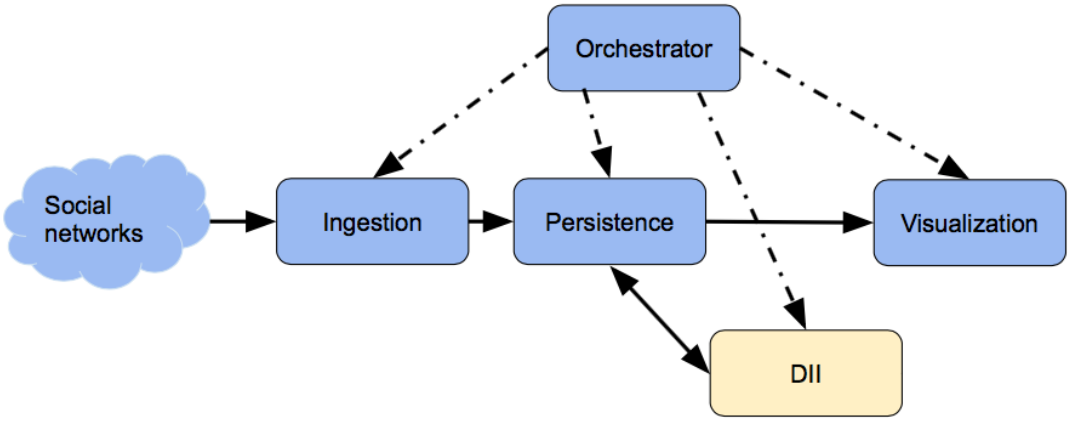


FIGURE 15: POST-ANALYSIS PIPELINE

As shown in the above diagrams, in the pre-analysis pipeline the DII services which perform the data processing are located after the ingestion process. In this way, after the data ingestion, data will be processed and then stored. In the post-analysis pipeline, DII services are located after the persistence process, so the data stored will be processed and then stored again. The different process of the pipeline will be exhaustively described below.

2.1 Ingestion

These modules are responsible for collecting data from the internet (i.e., DW website and social media). It acts as an interface of the platform to interact with the rest of the world. The collected data is stored in the platform for subsequent analysis. Specifically, ingestion modules must be capable of storing the collected data in the HDFS of the platform.

Some of the ingestion modules [4] in the SoMeDi platform are described in Module &Tools section (Twitter Crawler, Facebook Crawler and Youtube) :

For the process of ingestion, a data processing pipeline must be build and capable of ingests data from multitude of sources simultaneously, transform if needed and sends it to the stash. (Logstash tool, described Module &Tools section)

2.2 Persistence

For persistence, special emphasis has been put into using distributable databases, which are the only ones capable of managing large amounts of data. For our platform, three main scenarios have been devised depending on the type of the data to be stored.

- First, the crawler (input) data, which is almost raw data that has to be later processed. As this data is not to be queried, but there will be a great volume of writing and reading, the efficiency is paramount. For this kind of persistence, HDFS has been selected due to its simplicity, efficiency, distribution architecture and ease of integration with Apache Spark (technology suggested to implement the platform orchestrator when it is operated as a big data solution).
- Second, the persistence for the processed data, or data results from NLP analysis. This persistence is to be queried and will be the base for modules dedicated to data representation. This persistence should be searchable, scalable, distributable and usable for visualization. (Elasticsearch is an appropriate tool of choice, described Module &Tools section)

2.3 Analytics and visualization

In order to be able to appreciate and optimize the results based on the use of the advanced linguistic technologies it is fundamental to be able to numerically represent every aspect of the data during the transformation phase as well as measuring the reaction with the consumers.

Some of the visualization tools are described in Module &Tools section and more of these are described and implemented in deliverables D2.3, D3.2 and D3.3 of the project.

2.4 Pipeline orchestrator

The pipeline orchestrator will be the element responsible for getting the configuration of the SoMeDi platform from the user and running the software according to the specified requirements. It must take into account the following issues:

- Pipeline. The order of use of the modules in the platform.
- Timing. The moment every module involved in the pipeline is launched.

The configuration of the pipeline should provide information on:

- The source of the data.
- The language or languages to be analysed.
- The processing modules: NLP and social network modules.
- Storage system (i.e., where the results are going to be stored).

As a result, the configuration file should provide a description of the instance of the platform, which would look as follows: “Collect data in English from a certain source (e.g. Twitter, Facebook), process them for certain analysis (extraction of sentiment and/or emotion and/or topic and/or entity and/or knowledge graph insights), store the results and plot certain variables with the visualization tool”.

Some of the orchestration modules options for the SoMeDi platform are described in Module &Tools section (Luigi).

3 DII ARCHITECTURE

The DII processing modules refer to the set of functionalities in the SoMeDi platform that provide the capabilities for analysis that will be applied in a parallel way on the set of items to be processed. These modules are provided by different partners of the consortium and, as a result, they will be developed on distinct platforms and technologies. Therefore, the integration of these functionalities into the SoMeDi platform represent the main challenge to be addressed. In the following points, we describe the strategies adopted for the integration of modules in the SoMeDi platform.

To design the integration of Natural Language Processing (NLP) functionalities in the SoMeDi big data platform, it is required to define a standard format for the data processed by them. Hence, this format will define both the input and the output of NLP methods.

In general, NLP functions are expected to accept a piece of text, which will be analyzed for the computation of a result related to a given context or problem. Those addressed in SoMeDi are the following ones:

- ***Topic extraction***: the topic areas that the text corresponds to.

- ***Entity/concept extraction***: the entities identified in the text.

- ***Sentiment extraction***: the sentiments expressed in the text.

- ***Emotion recognition***: the emotions expressed in the text.

As commented before, the input will tend to be common for all the NLP modules. However, it may differ from a module to another. On the other hand, the output will clearly depend on the addressed NLP problem. Hence, a NLP function for concept extraction may yield a set of words indicating the entities found, while a function for emotion extraction will provide a set of tags reflecting the emotions perceived from several possible choices.

To address this issue, we propose the definition of a generic input/output. We propose using a common JSON format to accommodate the particular behaviours of every NLP function. Hence, a given NLP module would follow the template:

```
myNLPfunction(String input) → String output
```

where the “input” string refers to the JSON fed into the method, whereas the “output” string corresponds to the JSON returned by the NLP function.

Finally, the chosen format for data integration is JSON-LD (JSON for Linked Data) which is an implementation of NIF for JSON.

It is worth noting that the definition of the input/output format aims to standardize the exchange of information between processing modules in the platform. However, as every module is developed in a different technology, a suitable strategy must be designed for their integration in the platform. The strategies for the integration of NLP modules in the SoMeDi big data platform depend on the implementation of each module.

With regards to the specification for text intelligence, the main purpose is the use of an approach based on microservices which share a common interface. For this reason, they will be mainly based on W3C NIF, Marl and Onyx, that will be described below.

- **W3C NIF** is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. NIF consists of specifications, ontologies and software (overview), which are combined under the version identifier "NIF 2.0", but are versioned individually.
- **Marl** is a standardised vocabulary designed to annotate and describe subjective opinions expressed on the web or in particular Information Systems.
- **Onyx** is an ontology designed to describe the emotions expressed by user-generated content on the web or in particular Information Systems, that aims to complement the Marl Ontology by providing a simple means to describe emotion analysis processes and results using semantic technologies.

```
{
  "@context": "http://mixedemotions-project.eu/ns/context.jsonld",
  "@id": "me:Result1",
  "@type": "results",
  "analysis": [
    {
      "@id": "me:SAnalysis1",
      "@type": "marl:SentimentAnalysis",
      "marl:maxPolarityValue": 1,
      "marl:minPolarityValue": 0
    }
  ],
  "entries": [
    {
      "@id": "http://micro.blog/status1",
      "@type": [
        "nif:RFC5147String",
        "nif:Context"
      ],
      "nif:isString": "Dear Microsoft, put your Windows Phone on you
      "entities": [
      ],
      "suggestions": [
      ],
      "sentiments": [
        {
          "@id": "http://micro.blog/status1#char=80,97",
          "nif:beginIndex": 80,
          "nif:endIndex": 97,
          "nif:anchorOf": "You'll be awesome.",
          "marl:hasPolarity": "marl:Positive",
          "marl:polarityValue": 0.9,
          "prov:wasGeneratedBy": "me:SAnalysis1"
        }
      ],
      "emotionSets": [
      ]
    }
  ]
}
```


FIGURE 16: SENTIMENT ANALYSIS OF AN INPUT

```
{
  "@context": "http://mixedemotions-project.eu/ns/context.jsonld",
  "@id": "me:Result1",
  "@type": "results",
  "analysis": [
    {
      "@id": "me:EmotionAnalysis1",
      "@type": "onyx:EmotionAnalysis"
    }
  ],
  "entries": [
    {
      "@id": "http://micro.blog/status1",
      "@type": [
        "nif:RFC5147String",
        "nif:Context"
      ],
      "nif:isString": "Dear Microsoft, put your Windows Phone on you
      "entities": [
      ],
      "suggestions": [
      ],
      "sentiments": [
      ],
      "emotions": [
        {
          "@id": "http://micro.blog/status1#char=0,109",
          "nif:anchorOf": "Dear Microsoft, put your Windows Phone on
          "prov:wasGeneratedBy": "me:EmotionAnalysis1",
          "onyx:hasEmotion": [
            {
              "onyx:hasEmotionCategory": "wna:liking"
            },
            {
              "onyx:hasEmotionCategory": "wna:excitement"
            }
          ]
        }
      ]
    }
  ]
}
```


FIGURE 17: EMOTION ANALYSIS OF AN INPUT

ANNEX B - METHODOLOGY AND PRACTISE: FROM USER GENERATED & SOCIAL MEDIA DATA INTO DIGITAL INTERACTION INTELLIGENCE – ANALYTICS SUMMARY

In this Annex we present some findings of SoMeDi in the field of Data Anaysis for Natural Language and Computer Vision. Some of these were covered in past deliverables of the project, most notably D3.2 and D3.3. Some information and technological descriptions that were deemed confidential for the partners in SoMeDi has been left out.

1. DATA EXTRACTION

Data extraction is the act or process of retrieving data out of (usually unstructured or poorly structured) data sources for further data processing or data storage (data migration). The import into the intermediate extracting system is thus usually followed by data transformation and possibly the addition of metadata prior to exporting to another stage in the data workflow.

In the following sections are presented the data extraction methods applied for each of SOMEDI use cases.

1.1. Data Extraction Techniques [Marketing Use Case]

The data used in the use case is gathered from Twitter, Facebook and TripAdvisor accounts of various restaurant chains, including LATERAL (our main end-user) as well as their competitors (100 Montaditos, La Sureña and Morao tapas) In order to acquire content from the internet the components of the Marketing use case are selected from a selection of different approaches developed during the project:

- Direct access to the social media using their offered APIs (for example, for Twitter[31] and Facebook[32]). These APIs offer standardized access to the data published in the accounts of these social media sites. There is a large selection of tools in the state

[31] https://developer.twitter.com/en/docs/api-reference-index - Twitter API reference.

[32] https://developers.facebook.com/docs/graph-api/ - Facebook Graph API reference.

of the art to perform these actions. In SoMeDi we have reused Logstash[33] that offers very streamlined access to the Twitter and Facebook data directly into the Elasticsearch backend that we're using for storage of data. Some modifications and configurations have been necessary to align the two approaches.

- For data that doesn't offer an API or requires a more complicated pre-processing, we're using a dedicated Scraper called Scrapy[34]. This enables the creation using Python of powerful rule-based scrapers that collect data from not fully structured sources (such as the HTML in web pages) and returns formatted elements in standard mark-up formats such as XML or JSON. We have also used gsicrawler[35] by UPM, providing equivalent functionality but with a less widespread usage.

With these software tools, we're able to get the results into the processing stage of the marketing use case. This is then done as batch processing with a low frequency (from twice daily to daily typically) as the processing does not require a high urgency. This will be further presented in section 7 of the document.

1.2. Data Extraction Techniques [Recruiting Use Case]

Data collection from the end users was realized by using several Web Forms who write data to a SQL database on a SQL server.

The information gathered through the HTML forms is later handed to a downstream data extraction process. The HTML form invokes a Common Gateway interaction (CGI) request to the Web server.

[33] https://www.elastic.co/es/products/logstash - Logstash Twitter API pipeline for Elasticsearch

[34] https://scrapy.org/ - Scrapy python scraper.

[35] https://gsicrawler.readthedocs.io/en/latest/ - GSI crawler.

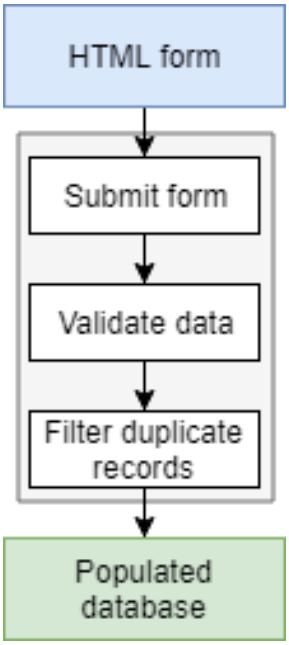


FIGURE 18. DATA COLLECTION FROM END-USERS IN UC2

The two methods to submit a form for CGI processing are presented below:

- First method, using the HTTP POST verb, the HTML forms can be submitted with (name, value) pairs encoded in the body of the request.
- Second method, using the HTTP GET verb, forms can be submitted by supplying (name, value) pairs in the URL.

The SoMeDi platform includes the option to secure communication and submissions of HTML forms through website by using Secure Hyper Text Transfer Protocol or https.

Submissions through the secured web form are stored in a way that only authorized and authenticated users can view the results.

1.3. Data Models

Data Warehouse

For the Somedi recruiting use case InnoDB engine is used, which supports foreign key and transaction. The default character set for this table is UTF8, which supports all languages for internationalization. The Data store repository with complete view of the business data is as follows:

- Aggregated data from multiple sources
- Active users, applicants, companies
- Programs

DATA MODELS

A Database model defines the logical design and structure of a database and defines how data will be stored, accessed and updated in a database management system. While the Relational Model is the most widely used database model, there are other models too:

- Hierarchical Model
- Network Model
- Entity-relationship Model
- Relational Model

The Somedi Recruitment platform is built on October CMS. October CMS provides a simple Active Record implementation for working with the database environment, based on Eloquent by Laravel. Each database table has a corresponding "Model" which is used to interact with that table. Models allow to query for data in all tables, as well as insert new records into the table.

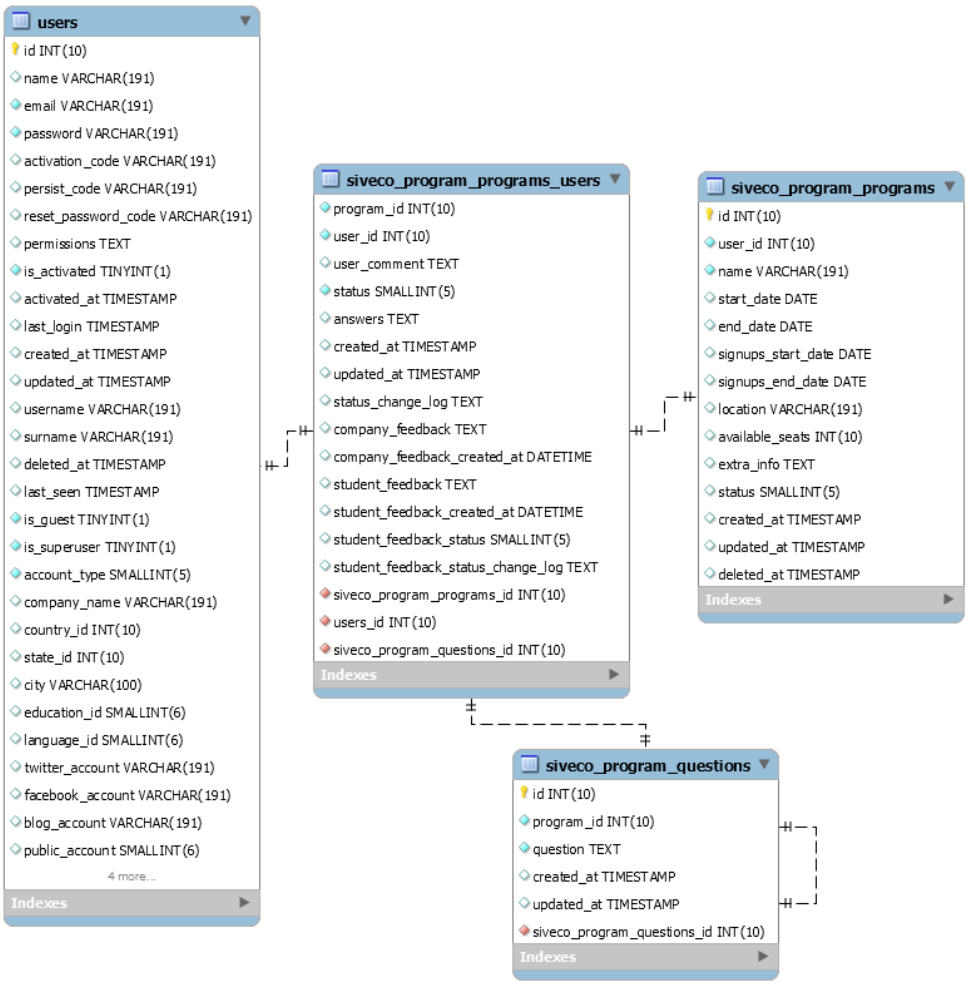


FIGURE 19. DATA BASE MODEL FOR UC2

2. DII TEXT INTELLIGENCE TOOKIT FOR MARKETING USE CASE

2.1. Architecture

We have used Python3 with conda[36] and used PyTorch[37] for deep learning framework.

Recently there have been big advances in Natural Language Processing with deep learning, and lots of open sourced tools are available: to name a few those from allenNLP[38] (ELMO). Zalando Research (Flair[39]), FastAI[40] and Google (BERT[41]). We have used Flair, due to its simplicity in its use, and its light architecture as well as due to the fact that it showed state of the art results for a couple of NLP tasks (as of 2018 November, NER English , NER German, Chunking and PoS tagging).

Flair [42] implements contextualized character-level word embeddings which combine the best attributes of the previously existed embeddings: the ability to (1) pre-train on large unlabeled corpora, (2) capture word meaning in context and therefore produce different embeddings for polysemous words depending on their usage, and (3) model words and context fundamentally as sequences of characters, to both better handle rare and misspelled words as well as model subword structures such as prefixes and endings.

Character level contextual embeddings are based on neural language modeling (LM) that have allowed language to be modeled as distributions over sequences of characters instead

[36] https://conda.io/docs/

[37] https://pytorch.org/

[38] https://github.com/allenai/allennlp

[39] https://github.com/zalandoresearch/flair

[40] https://github.com/fastai

[41] https://github.com/google-research/bert

[42] Akbik, A., Blythe, D. & Vollgraf, R., 2018. Contextual String Embeddings for Sequence Labeling. 27th International Conference on Computational Linguistics.

of words [43] [44] [45]. Recent work has shown that by learning to predict the next character on the basis of previous characters, such models learn internal representations that capture syntactic and semantic properties: even though trained without an explicit notion of word and sentence boundaries, they have been shown to generate grammatically correct text, including words, subclauses, quotes and sentences [46]. More recently, Radford and colleagues [47] showed that individual neurons in a large LSTM-LM can be attributed to specific semantic functions, such as predicting sentiment, without explicitly trained on a sentiment label set.

2.2. Architecture for training NER (Name Entity Recognition) and sentiment classifier

In SoMeDi, we have implemented English and Spanish Named Entity Recognizer as well as Sentiment Analyzer. For this we have trained Spanish character based context aware language model on the GPU machine on AWS (p2.xlarge, P2 Tesla K-series K-80) with 61 GB of memory and 12 GB of GPU memory for 2 weeks (each model).

Trained character based language model was used for Spanish NER and Spanish Sentiment analyser, as part of embedding. For Spanish NER, and Spanish and English sentiment analyser, we have trained each of them about one day.

NER (Named Entity Recognition)

Named Entity Recognition is a task where entities such as Person, Organization, Location are extracted from the unlabeled sentences.

[43] Sutskever, I., Vinyals, O. & Le, Q., 2014. Sequence to sequence learning with neural netowrks. Advances in neural information processing systems, pp. 3104-3112.

[44] Graves, A., 2013. Generating sequences with recurrent neural networks. arXiv:1308.0850.

[45] Kim, Y., Jernite, Y., Sontag, D. & Rush, A. M., 2015. Character-aware neural language models. arXiv:1508.06615.

[46] Karpathy, A., Johnson, J. & Fei-Fei, L., 2015. Visualizing and understanding recurrent networks. arXiv:1506.02078.

[47] Radford, A., Jozefowicz, R. & Sutskever, I., 2017. Learning to generate reviews and discovering sentiment. arXiv:1704.01444.

We have used CONLL2002 dataset[48] to train Spanish NER. We have acquired F1 value of 85.92 for validation set, and 87.58 for the test set, which is higher than the previous state of the art of Spanish NER (85.77 by [49]).

For English NER, we have used existing implementation of Flair. This model is current state of the art model with F1 value of 93.09.

SENTIMENT ANALYSIS

Opinions are central to almost all human activities because they are key influencers of our behaviors. With the explosive growth of social media (e.g. reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for decision making. Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products services, organizations, individuals, issues, evens, topics, and their attributes.[50]

We have trained English sentiment analyzer with a dataset[51] which contains 1,578,627 classified tweets. We have achieved F1 value of .816 with this dataset. For Spanish sentiment analyzer, we have combined datasets from TASS2012[52] and TASS2018[53], which in total contained 10,026 classified tweets. We have reached F1 value .4763 with this dataset.

[48] https://www.clips.uantwerpen.be/conll2002/ner/

[49] Yang, Z., Salakhutdinov, R. & Cohen, W. W., 2017. Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. arXiv:1703.06345.

[50] Liu, B., 2012. Sentiment analysis and opinion mining. s.l.:Morgan & Claypool publishers.

[51] http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/

[52] http://www.sepln.org/workshops/tass/2012/about.php

[53] http://www.sepln.org/workshops/tass/2018/

```
# tag type for prediction

tag_type = 'ner'

# making tag dictionary from the corpus

tag_dictionary = corpus.make_tag_dictionary(tag_type=tag_type)

print(tag_dictionary.idx2item)

# initialize embeddings

embedding_types: List[TokenEmbeddings] = [

    WordEmbeddings('es-glove'),

    CharLMEmbeddings('./resources/taggers/language_model_es_forward_long/best-lm.pt'),

    CharLMEmbeddings('./resources/taggers/language_model_es_backward_long/best-lm.pt'),

]

embeddings: StackedEmbeddings = StackedEmbeddings(embeddings=embedding_types)

# initialize sequence tagger

from flair.models import SequenceTagger

tagger: SequenceTagger = SequenceTagger(hidden_size=128,

                                        embeddings=embeddings,

                                        tag_dictionary=tag_dictionary,

                                        tag_type=tag_type,

                                        use_crf=True)

# initialize trainer

from flair.trainers.sequence_tagger_trainer import SequenceTaggerTrainer


trainer: SequenceTaggerTrainer = SequenceTaggerTrainer(tagger, corpus, test_mode=False)
```

Part of Code for training Spanish NER.

```
# making a list of word embeddings
word_embeddings = [WordEmbeddings('en-twitter-glove'),
                   CharLMEmbeddings('mix-forward'),
                   CharLMEmbeddings('mix-backward')]


# initialize document embedding by passing list of word embeddings
document_embeddings: DocumentLSTMEmbeddings = DocumentLSTMEmbeddings(word_embeddings,

hidden_states=512,

reproject_words=True,

reproject_words_dimension=256,)


# create text classifier
classifier = TextClassifier(document_embeddings, label_dictionary=label_dict,
multi_label=False)


#initialize text classifier trainer
```

Part of code for training English sentiment classifier.

3. DII TEXT INTELLIGENCE TOOKIT FOR RECRUITING USE CASE

3.1. Recruitment Scenario Description

This use case applies Sentiment Analysis (SA) techniques to improve the recruitment processes aiming to increase the efficiency of internship campaigns by ensuring a better match between the candidates' professional skills and the hiring company fields of activity.

Therefore, by accessing the SoMeDi platform, the candidates will complete their profile data and further - browse, select, and apply to specific internship programmes. The application process implies that the internship candidates will complete several forms providing feedback regarding the companies fields of activity.

The NLP tool advanced in this project analyzes each field completed, containing the text written by the candidate. The application delivers a score (that approximates how interested the candidate is to work in each area) namely a value between 0 and 1. If the sentiment analysis score is close to 0 means that there is no interest, while a value close to 1 signifies that the candidate is interested.

As mentioned earlier, the sentence analysis is performed by using NLP Text Analytics, namely Sentiment Analysis. The current version of the DII tool is delivered in this phase of the project by deploying the following methods (services) for sentiment analysis:

a) the first method uses services from Microsoft Azure - Cognitive Services;
b) the second method is built with open source Stanford CoreNLP.

We developed a sentiment analysis application using an open source code written in C # to present an appropriate GUI for the internship application. The sentiment analysis application has versions in English and Romanian, and for analyzing Romanian language content (text input) was used a Romanian-English translation service from MS Azure Translator Text.

In the following sections we will present the methods used for developing the sentiment analysis applications, describe each of the SA solutions (Azure and Stanford CoreNLP), and also release the instructions on how to use/test the above-mentioned applications.

3.2. Methods used for Sentiment Analysis

Sentiment Analysis is part of the Text Analytics.

Understanding and analyzing unstructured text is an increasingly popular field and includes a broad spectrum of problems such as sentiment analysis, key phrase extraction, topic modeling/extraction, aspect extraction and more. A simple approach is to keep a lexicon of words or phrases that assess negative or positive sentiment to a sentence (e.g., the words “bad”, “hate”, “not good” would belong to the lexicon of negative words, while “good”, “great”, “like” would belong to the lexicon of positive words). But this means such lexicons must be manually curated, and even then, they are not always accurate.

Methods based on Machine Learning

A more robust approach is to train models that detect sentiment. Here is how the training process works – a large dataset of text records is created that was already labeled with sentiment for each record. The first step is to tokenize the input text into individual words,

then apply stemming[54] (stemming is the process of reducing inflected -or sometimes derived- words to their base or root form). Next, it is necessary to construct features from these words; these features are used to train a classifier. Upon completion of the training process, the classifier can be used to predict the sentiment of any new piece of text. It is essential to construct meaningful features for the classifier, and the list of features includes several from state-of-the-art research:

- N-grams[55] denote all occurrences of *n* consecutive words in the input text. The precise value of *n* may vary across scenarios, but it's common to pick *n=2* or *n=3;*
- Part-of-speech tagging [56] [57] is the process of assigning a part-of-speech to each word in the input text;
- Word embeddings[58] [59] are a recent development in natural language processing, where words or phrases that are syntactically similar are mapped closer together. Neural networks are a popular choice for constructing such a mapping. For sentiment analysis, neural networks that encode the associated sentiment information are used as well. The layers of the neural network are then used as features for the classifier.

3.3. Description of the Microsoft Azure Cognitive Services – Text Analytics Project

Text Analytics uses a machine learning classification algorithm to generate a sentiment score between 0 and 1. Scores closer to 1 indicate positive sentiment, while scores closer to 0 indicate negative sentiment.

The model is pretrained with an extensive body of text with sentiment associations. Currently, it is not possible to provide your own training data. No labeled or training data is needed to use the service.

The model uses a combination of techniques during text analysis, including text processing, part-of-speech analysis, word placement, and word associations.

Sentiment analysis is performed on the entire document, as opposed to extracting sentiment for a particular entity in the text. In practice, there is a tendency for scoring accuracy to improve when documents contain one or two sentences rather than a large block of text. During an objectivity assessment phase, the model determines whether a document as a

[54] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[55] https://web.stanford.edu/~jurafsky/slp3/4.pdf

[56] https://web.stanford.edu/~jurafsky/slp3/10.pdf

[57] https://www.nltk.org/book/ch05.html

[58] http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/

[59] https://medium.com/@josecamachocollados/on-the-contribution-of-neural-networks-and-word-embeddings-in-natural-language-processing-c8bb1b85c61c

whole is objective or contains sentiment. A document that is mostly objective does not progress to the sentiment detection phrase, resulting in a 0.50 score, with no further processing. For documents continuing in the pipeline, the next phase generates a score above or below 0.50, depending on the degree of sentiment detected in the document.

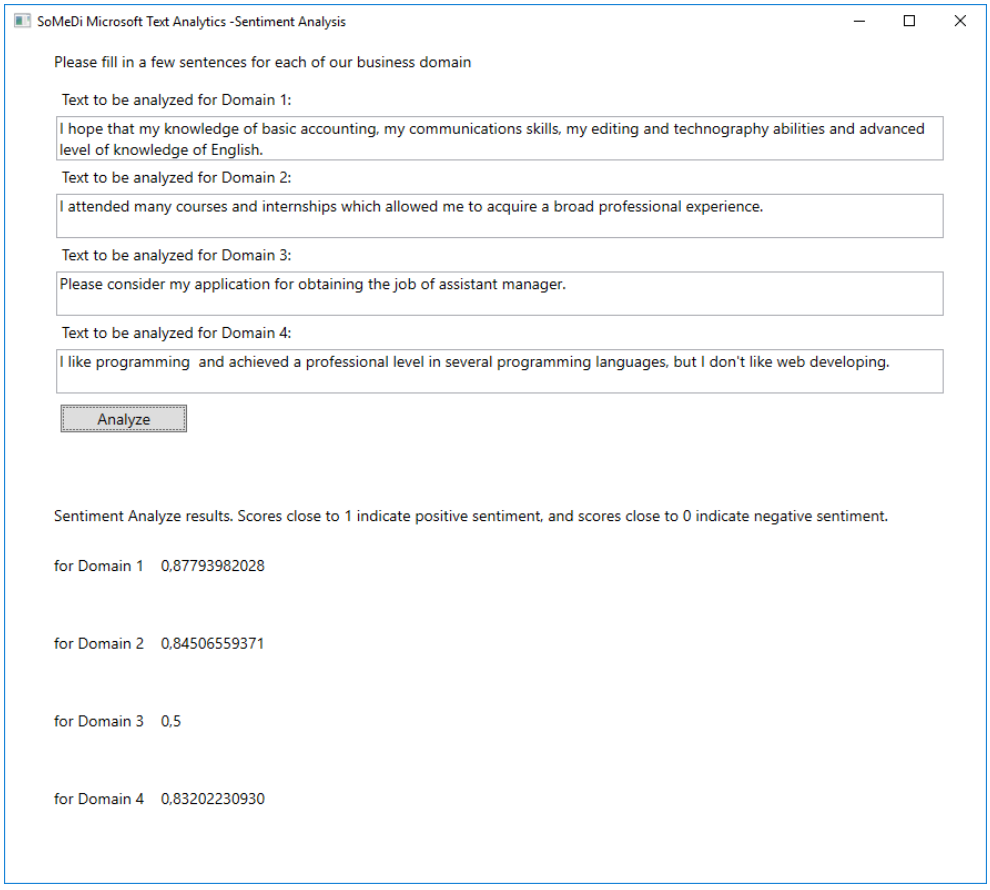


FIGURE 20. (A) MICROSOFT AZURE SENTIMENT ANALYSIS APPLICATION (EN)

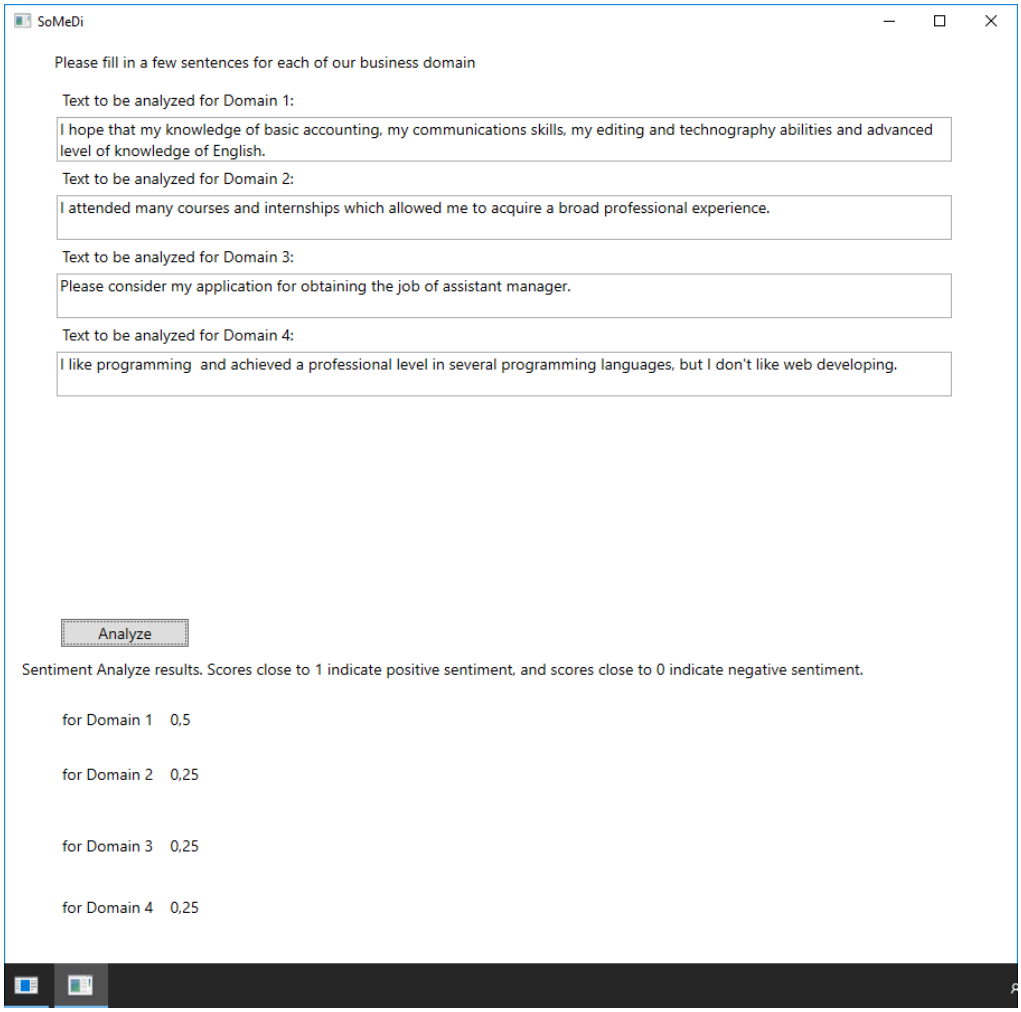


FIGURE 21. (B) STANFORDNLP SENTIMENT ANALYSIS APPLICATION (EN)

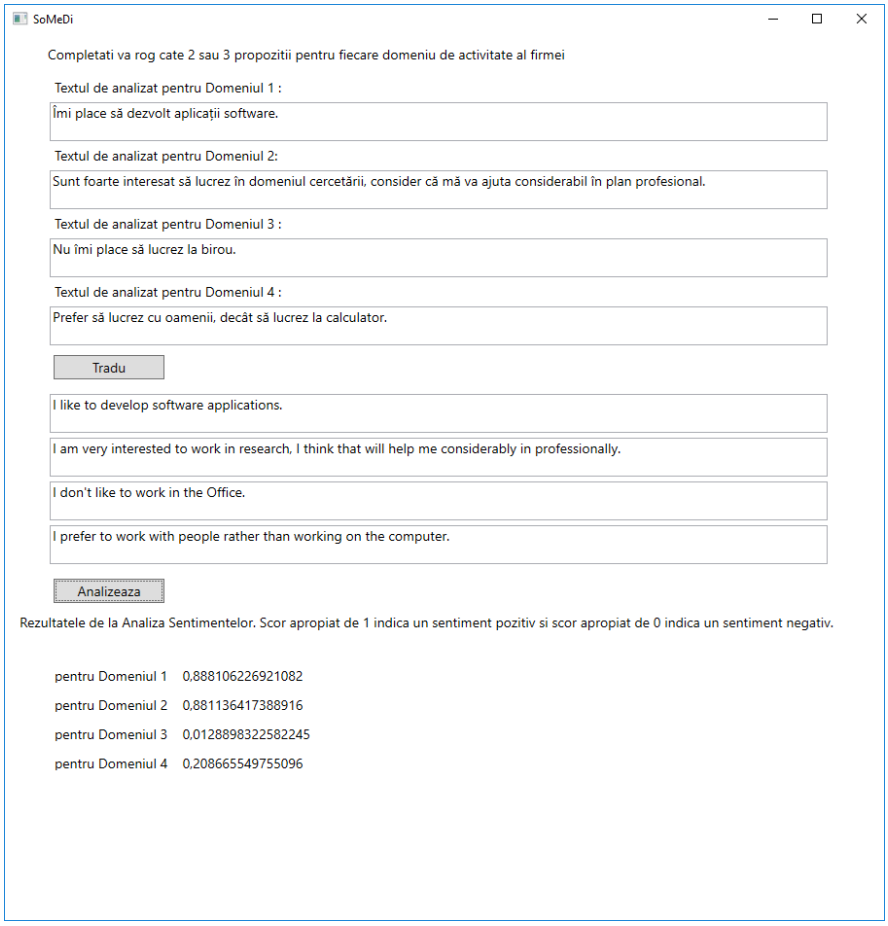


FIGURE 22. (C) MICROSOFT AZURE SENTIMENT ANALYSIS APPLICATION (RO)

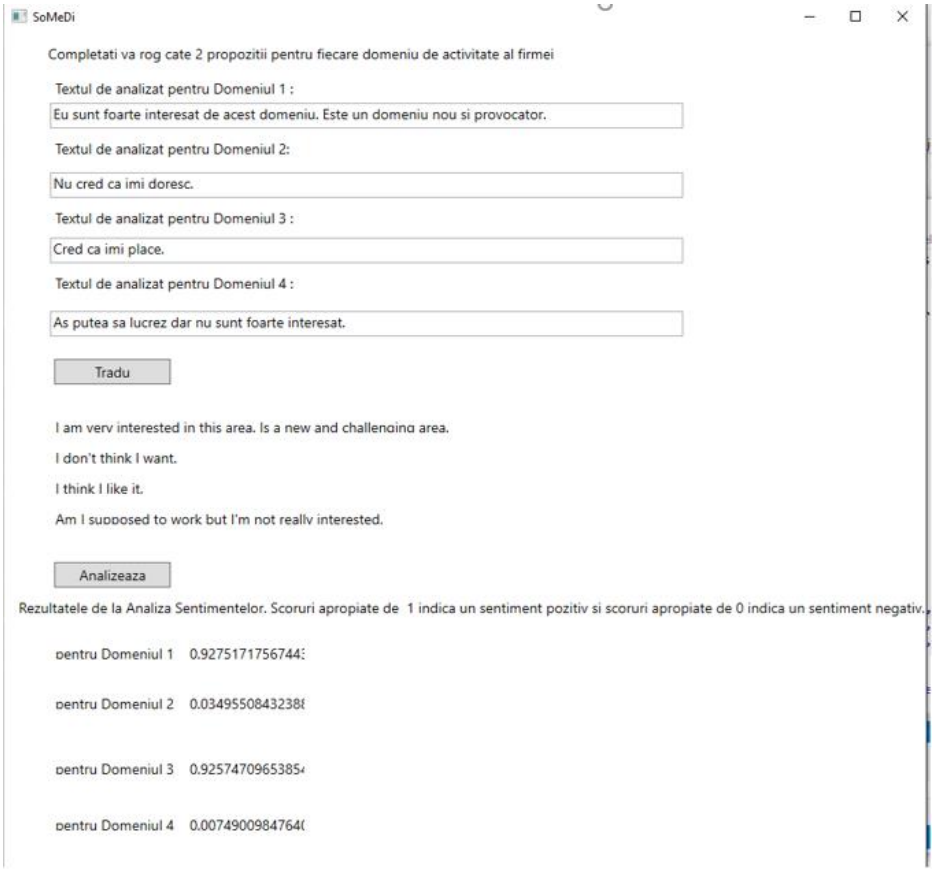


FIGURE 23. (D) STANFORDNLP SENTIMENT ANALYSIS APPLICATION (RO)

3.4. Description of the Stanford CoreNLP Sentiment Analysis Project

Stanford CoreNLP provides a set of human language technology tools. It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences regarding phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

The following Stanford CoreNLP tools are necessary for this project for Sentiment Analysis, in order to: tokenize, ssplit, pos, lemma, parse, sentiment; ner and dcoref are not necessary.

The functions of these tools are:

i. tokenize: Tokenizes the text into a sequence of tokens,
ii. ssplit: Splits a sequence of tokens into sentences,
iii. pos: Labels tokens with their part-of-speech (POS) tag,
iv. lemma: Generates the lemmas (base forms) for all tokens; including Sentiment Class,

v. parse: Provides full syntactic analysis, including, both constituent and dependency representation tokens,
vi. sentiment: Sentiment analysis with a compositional model over trees using deep learning.

Stanford CoreNLP is written in Java. Stanford CoreNLP introduced two new ideas: a) the Stanford Sentiment Treebank and b) a powerful Recursive Neural Tensor Network.

A treebank can be defined as a linguistically annotated corpus (Data Base) that includes some grammatical analysis beyond the part-of-speech level. The Stanford Sentiment Treebank is the first corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in language. It includes labels for every syntactically plausible phrase in thousands of sentences.

Recursive Neural Tensor Networks (RNTN) take as input phrases of any length. They represent a phrase through word vectors and a parse tree and then compute vectors for higher nodes in the tree using the same tensor-based composition function. It is placed on top of grammatical structures. A phrase is composed of a couple of meaning related words/tokens. The tri-gram is used.

The Deep learning model builds up a representation of the whole sentence based on the sentence structure. It computes the sentiment based on how words compose the meaning of longer phases. The activation function is: a) f=tanh() for hidden layers and b) f= softmax() for output layer used for 5-class classification. These 5-class sentiment classifications are: VERY NEGATIVE, NEGATIVE, NEUTRAL, POSITIVE, VERY POSITIVE.

Training of RNTN is performed minimizing the cross-entropy error between the predicted distribution at each node and the target distribution at the same node.

3.5. Software Development

BEIA developed the following software programs in C#,

i. SoMeDi_Sentiment-Analyze_MS-Azure_EN - using MS-Azure Services, in English
ii. SoMeDi_Sentiment-Analyze_MS-Azure_RO - using MS-Azure Services, in Romanian
iii. **SoMeDi_Sentiment-Analyze_StanfordCoreNLP_EN - using StanfordCoreNLP Tools, in English,**
iv. SoMeDi_Sentiment-Analyze_StanfordCoreNLP_RO - using StanfordCoreNLP Tools, in Romanian.

The programs are presented as Installation programs for Windows 10, x64.

Folder: “Doc. Install-3 programs.zip” contains these installation programs.

Notes:

- No shortcuts for programs
- The folder " stanford-corenlp-3.9.1-models" is necessary to be placed on : C:\Program Files\BEIA\SoMeDi_Sentiment-Analyze_StanfordCoreNLP_EN, after installation.
- MS_Azure services need a MS account; the programs do not work without this (remote server returns error).

ANNEX C - METHODOLOGY AND PRACTISE: FROM USER GENERATED & SOCIAL MEDIA DATA INTO DIGITAL INTERACTION INTELLIGENCE – DATA PRESENTATION AND REPORTING SUMMARY

In this Annex we present some findings of SoMeDi in the field of Data Representation and Reporting tools. Some of these were covered in past deliverables of the project, most notably D2.1 and D2.3. Some information and technological descriptions that were deemed confidential for the partners in SoMeDi has been left out.

1. INTRODUCTION

This Annex describes the architecture and development of the Data Analysis and Reporting Tool produced in SoMeDi, with a special focus on the visualization aspects. Given the modular nature of the tool, the process of extending the tool and customizing it to a specific use case is also covered in this document. To illustrate the flexibility of the tool, several use cases are shown, which leverage different data sources and visualization techniques.

The base visualization toolkit and templates to generate a customized data analysis workflow are published at http://lab.cluster.gsi.dit.upm.es/sefarad/sefarad and a specific dashboard covering the SoMeDi use case is available at http://lab.cluster.gsi.dit.upm.es/sefarad/somedi

2. ENABLING TECHNOLOGIES

In this section, we are going to give an insight into techniques used in this project. First, we are going to explain Polymer web components, the technology in charge of the visualization's structure. Secondly, Elasticsearch the technology used to store data.

2.1. Polymer

Polymer[60] is a software library used to define and style web components that was developed by Google. Modern design principles are implemented as a separate project using Google's Material Design principles.

Polymer makes easier to build your very own custom HTML elements. Creating reusable custom elements can make building complex web applications easier and more efficient. By being based on the Web Components API's built in the browser, Polymer elements are interoperable at the browser level, and can be used with other frameworks or libraries that work with modern browsers.

These custom elements are particularly useful for building reusable UI components. Instead of continually re-building a specific navigation bar or button in different frameworks and for different projects, you can define this element once using Polymer, and then reuse it throughout your project or in any future project.

Polymer uses a declarative syntax to make the creation of your own custom elements easier; they use all standard web technologies: HTML is used to define the structure of the element, CSS is for style personalization and you can use JavaScript to make these elements interactive.

In addition, Polymer has been designed to be flexible, fast and close. It uses the best specifications of the web platform in a direct way to simply custom elements creation

2.2. Elasticsearch

Elasticsearch[61] is a search server based on Lucene. It provides a distributed, full-text search engine with an HTTP web interface and schema-free JSON documents. Elasticsearch is distributed, which means that indices can be divided into shards and each shard can have zero or more replicas. Each node hosts one or more shards, and acts as a coordinator to delegate operations to the correct shard(s).

The search API allows to execute search queries and get back search hits that match the query. The query can be provided either using a simple query string as a parameter, or using a request body.

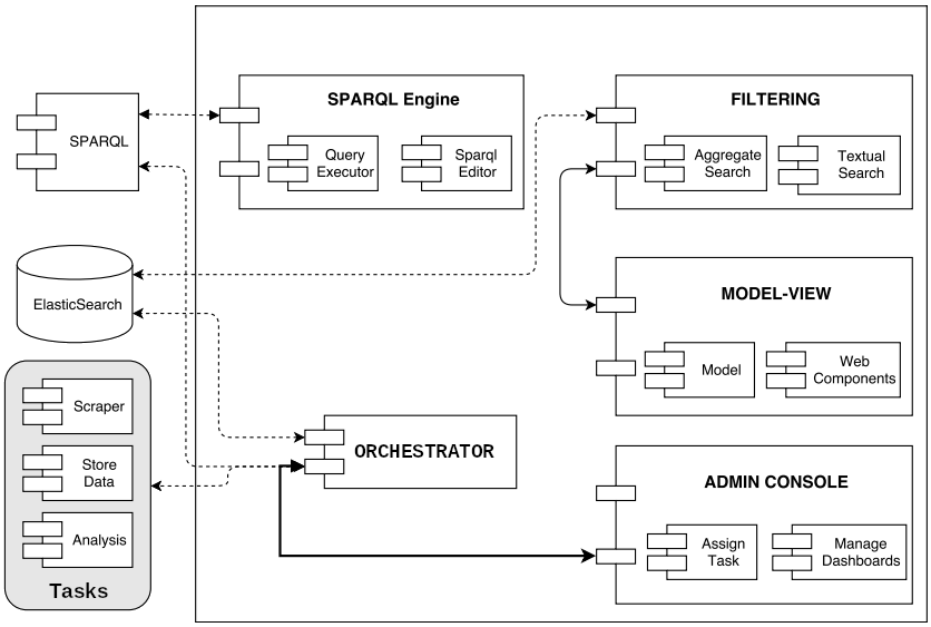
3. ARCHITECTURE

[60] https://www.polymer-project.org/ - Google Polymer website

[61] https://www.elastic.co/ - Elasticsearch website

We chose to follow a Model-View-Controller architecture. Where a single element is in charge of connecting to the data sources, filtering the results, and exposing it to other components, which can then present it. Since this visualization should also be interactive, visualization components will also contain their own set of filters. When interacting with these components, a user may modify the filters, and the component will communicate the change of filters back to the filtering component. The communication between components is achieved through Observers and computed properties, which allow changes to be seamlessly propagated to all components. The result of combining Web Components with these conventions to organize data is Sefarad[62], which is the core of the visualization module in the toolkit. The main function of Sefarad is to represent data in meaningful and interactive ways that lead to insights. This visualization is composed of individual dashboards, which are web pages oriented to display all the collected information.

In turn, these dashboards are further divided into other components (Polymer Web Components) that are connected to present a coherent view.



[62] https://github.com/gsi-upm/Sefarad - Sefarad repository

4. AVAILABLE DASHBOARDS AND WIDGETS

4.1. Dashboards

Sefarad framework has multiple predefined dashboards, which are ready to be used. Due to its high flexibility, it enables to display any kind of information, either user custom data sets or others included by default in the project.

In this section, we are going to explain each dashboard, detailing most user-attractive functionalities using a predefined dataset and presenting the real value of the tool, so anyone can use Sefarad framework by applying his or her customized data.

Due to the modular architecture offered by Sefarad framework, each dashboard has its own structure regardless, being able to interact with the components without influencing the rest. Moreover we are going to focus on explain what can be done in every dashboard through a case study without deepening inside its internal behavior.

Using Sefarad framework is used for developing dashboards for many uses cases, like SPARQL DBPedia, Tourpedia, Financial Twitter Tracker, Footballmood or Aspects. In addition, this framework is used within SoMeDi project.

4.2. Widgets

Sefarad widgets are based in web components. In order to be able to obtain interactive widgets, Polymer events are used to establish a relationship between them. This allows storing which elements have been selected and thus making queries that are more complex to Elasticsearch.

At the time of writing, this is a categorized list of popular widgets in the toolkit:

- Data statistics widgets: These widgets are used to visualize data statistics from an Elasticsearch index at a glance. We include inside this category Google-chart-Elasticsearch, number-chart, spider-chart, Liquid-fluid-d3, wordcloud…
- Sentiment widgets: These widgets are used to visualize sentiment information. We include inside this category chernoff-faces, field-chart, tweet-chart, wheel-chart, Youtube-sentiment…
- NER widgets: These widgets are used to visualize recognized entities from an Elasicsearch index. We include inside this category entities-chart, people-chart, aspect-chart, wheel-chart…
- Location widgets: This group of widgets visualize data geolocated in different maps. Spain-chart, happymap and leaflet-maps are some examples of this kind of widgets.
- Document widgets: Inside this group we can find tweet-chart and news-chart. These widgets are used to visualize all documents within an Elasticsearch index.

- Query widgets: This widgets add more functionalities to Sefarad framework, they are used to modify or ask queries to different endpoints. We include inside this category material-search, YASGUI-polymer, date-slider…

If you want to use some of this components just add them in the dashboard's bower.json file as a dependency. In addition, is necessary to import them inside the dashboard using HTML link tag. Finally, the widget is ready to use using its custom tag.

5. Widget Development

In this section, we will explain how to create new widgets in Sefarad, or import existing ones. For the tutorial, we are going to use number-chart widget mentioned above.

This example is the guideline to create a number-chart widget. First of all you must create a new directory inside `./bower_components`, called number-chart to store all your widget files.

Afterwards, you have to create a new file called `number-chart.html`. If you want to use other widgets inside you have to import them:

```
<link rel="import" href="/bower_components/polymer/polymer.html">
<link rel="import" href="/bower_components/iron-icons/iron-
icons.html">
<link rel="import" href="/bower_components/iron-icon/iron-icon.html">
<link rel="import" href="/bower_components/iron-icons/maps-
icons.html">
<link rel="import" href="/bower_components/iron-icons/social-
icons.html">
```

In addition, you have to define the HTML structure inside `<template>` tag. Sefarad widgets uses Polymer, so variables and data are passed inside curly braces. For more information about Polymer data binding visit Polymer documentation: https://www.polymer-project.org/1.0/docs/devguide/data-binding

```
<dom-module id="number-chart">

<template>
<! -- HERE GOES THE HTML STRUCTURE OF YOUR WIDGET -->
<div class="info-box">
  <div class$="{{stylebg}}">
    <span class="info-box-icon"><iron-icon icon="{{icon}}"></iron-
icon></span>
    <div class="info-box-content">
      <span class="info-box-text">{{title}}</span>
      <span class="info-box-number">
      <span id="number">{{number}}</span>
      <div class="progress">
```

```
        <div class="progress-bar progress-bar-name" id="barprogress"
style="width: 50%"></div>
      </div>
      <span class="progress-description">{{ subtitle }}: {{ total
}}</span>
    </div>
  </div>
</div>
</template>

</dom-module>
```

You may need some CSS rules to style your widget.

Below <template> tag is necessary to add javascript to define your component. Create a Polymer Object with following parts: * is: String that defines the name of the widget. * properties: Object with some widgets properties, add observers if you want to fire a funcion if that property changes. These properties are very useful to store data. * functions: Javascript functions that can be callable by the widget, for example to edit some data or manage requests.

```
Polymer({
  is: 'number-chart',
  properties: {
    icon: {
      type: String,
      value: "trending-up"
    },
    stylebg: {
      type: String,
      value: 'bg-yellow'
    },
    data: {
      type: Object,
      observer: '_dataChanged'
    },
    title: {
      type: String
    }
  },
  _dataChanged: function(){
    var num = this.data.hits.total > 999 ?
(this.data.hits.total/1000).toFixed(1) + 'k' : this.data.hits.total;
    idNum.innerHTML = num;
    this.total = num;
    idBar.style.width = "100%";

  }
});
```

Is also necessary to specify dependencies for this widget using a bower.json file. The structure of this file is like this example:

```
{
  "name": "number-chart",
  "homepage": "https://lab.cluster.gsi.dit.upm.es/sefarad/number-
chart",
  "authors": [
    "GSI-UPM"
  ],
  "description": "",
  "main": "",
  "license": "MIT",
  "dependencies": {
    "iron-icons": "PolymerElements/#iron-icons^1.1.0",
    "polymer": "polymer#*"
}
```

If you want to make your widget installable via bower you can register this package. This requires to have a git repository with all your widget code.

```
$ bower register <my-package-name> <git-endpoint>
```

6. Dashboard Development

Sefarad dashboards are also based in web components. Each dashboard is a collection of Sefarad widgets that displays different data.

Sefarad dashboards are created the same way as Sefarad widgets.

6.1. Fetching data from elasticsearch

The main difference between widgets and dashboards is that dashboards fetch the data passed to widgets. This process require elastic-client component:

```
<elastic-client
  config='{"host": " <!-- ELASTICSEARCH ENDPOINT GOES HERE --> "}'
  client="{{client}}"
  cluster-status="{{myStatus}}">
</elastic-client>
```

After client creation is possible to make queries. Create a new function inside your dashboard Polymer Object.

```
_query: function() {
   var that = this;
     this.client.search({
     // undocumented params are appended to the query string
     index: "<!-- ELASTICSEARCH INDEX -->",
     type: "<!-- ELASTICSEARCH DOCTYPE -->",
     body: {
```

```
        size: 10,
        query: {
          bool: {
            must: [],
          }
        }
      }
      }).then(function (resp) {
        that.data = resp;
        });
  }
```

Elasticsearch results are stored in a Javascript object called data. This data is passed to widgets like this number-chart widget example:

```
<number-chart
    data="{{data}}"
    object="restaurant"
    title="Restaurants"
    icon="maps:local-dining"
    stylebg="bg-yellow">
</number-chart>
```

7. DEPLOYMENT

Sefarad installation is based in Docker containers, only requirement is to have Docker and docker-compose installed.

First of all, you need to clone the Github repository:

```
git clone http://lab.cluster.gsi.dit.upm.es/sefarad/dashboard-
somedi.git
git checkout -b lateral-demo
git pull origin lateral-demo
```

Now the image is ready to run:

```
$ sudo docker-compose up
```

Sefarad visualization server is now running at port 8080

7.1. Loading demo data to visualisation server

Demo data is loaded to ElasticSearch via python script. In a new terminal type:

```
$ docker-compose exec gsicrawler python loaddatatoES.py
```

Finally, check your Sefarad visualization environment visiting different demos available.