# D1.1 SoMeDi Vision

*WP1 Vision, architecture and data integration – T1.1 SoMeDi Vision and context*

| *Delivery Date*: | *Project Number*: | *Responsible partner*: |
|---|---|---|
| M3 - 28/02/2017 | ITEA3 Call2 15011 | HI iberia |

# DOCUMENT CONTRIBUTORS

| Name | Company | Email |
|---|---|---|
| Elena Muelas | HIB | emuelas@hi-iberia.es |
| Inmaculada Luengo | HIB | lluengo@hi-iberia.es |
| Carlos A Iglesias | UPM | |
| George Suciu | BEIA | george@beia.ro |
| Cristina Ivan | Siveco ROMANIA | Cristina.Ivan@siveco.ro |
| Mirela Ardelean | Siveco ROMANIA | mirela.ardelean@siveco.ro |
| Dragos Papatoiu | Siveco ROMANIA | dragos.papatoiu@siveco.ro |
| Emilio Madueño | Innovati | emh@grupoinnovati.com |

# DOCUMENT HISTORY

| Version | Date | Author | Description |
|---|---|---|---|
| 0.1 | 16.12.2016 | HIB | First ToC distribution requesting contributions |
| 0.2 | 30.12.2016 | BEIA | Romanian contributions to SotA |
| 0.3 | 04.01.2016 | Innovati | Contributions to architecture |
| 0.4 | 10.01.2017 | HIB | New version with restructuration of the sections, integration of the different contributions. Also added Beyond SotA contribution to sentiment analysis, Manifesto section and Marketing use case description. |
| 0.5 | 17.01.2017 | UPM | Contribution to Dashboards for social media |
| 0.6 | 18.01.2017 | SIVECO, BEIA | Romanian contribution |
| 0.7 | 18.01.2017 | HIB | Improvement on use cases and SOMEDI Manifesto |
| 0.8 | 19.01.2017 | INNOVATI | Contributions to section 2 and 5 |
| 0.9 | 02.02.2017 | BEIA, SIVECO | Review of contributions |
| 0.10 | 06.02.2017 | Taiger, HIB | Contribution to machine learning and big data technologies/Contribution to sentiment |

| | | | |
|---|---|---|---|
| | | | analysis |
| 0.11 | 20.02.2017 | BEIA, SIVECO | Small contribution on section 2 and 6, add references for section 3 AI |
| 1.0 | 22.02.2017 | HIB | Final version distributed for review |

## Table of Contents

## TABLE OF FIGURES

# References

http://www.fastweb.com/college-scholarships/articles/social-media-internships

http://documents.wfp.org/stellent/groups/public/documents/communications/wfp239384.pdf

http://www.sabinacornovac.ro/joburi-social-media-20/

http://intern.internship-uk.com/how-do-social-media-influence-your-recruitment/

http://cpr.indiana.edu/uploads/AERA14%20Bridge%20or%20Barrier%20Paper.pdf

https://www.strategy-business.com/article/Social-Network-Effects-in-Hiring?gko=33101

https://www.robertwalters.co.uk/content/dam/robert-walters/country/united-kingdom/files/whitepapers/rw-social-media-whitepaper.pdf

# 1. INTRODUCTION

The objective of this deliverable is to compile the SoMeDi Vision to synthesize the overall purpose of the project results and their application to solve the existing problems identified to exploit the value of digital interaction data that can drive and support various business processes and use situations.

This is the first project deliverable, so establish its basis including an overall vision of SOMEDI DID tool (SoMeDi – Digital Intelligence Data Tool ) based on SOMEDI innovations expected for the project:

- Development of a set of advanced mining tools for representing, analyzing and extracting meaningful patterns or topics from social media and digital interaction data.
- Development of a set of improved machine learning algorithms enabling detection, prediction and support for automatic decision making in processes.
- Improved interactive tools to visualize and manage the data

These goals defined the technologies included in the state of the art (

State of the Art of the technologies involved in ) and beyond state of the art (SoMeDi technological innovations - Beyond the state of the art) sections of this deliverable, mainly:

- Machine learning
- Artificial Intelligence
- Opinion mining
- Sentiment analysis
- Dashboards for social media
- Data collection and big data

These technological advances will enable the industry to encompass new or improved business processes that will cover the wide-needs and current pains identified in Industry-wide needs and current pain points. The main SoMeDi objectives on the business processes side are:

SoMeDi will enable companies to improve their social business intelligence,

It will allow for new product or service development from harvesting ideas to quickly taking new products to the market using social media.
It will allow improving the ‘stickiness’ (retention) of customers for content and services.
It will enable tailored recommendation, cross and upsell of offers increasing user satisfaction and economic value.

These objectives will be reflected in the application use cases envisaged for SoMeDi in SOMEDI Preliminary Use Cases: Social Media for marketing purposes and Social Media driven advanced content planning and personalized consumption.

In summary, this document provides a complete vision of SoMeDi scope and objectives through the following sections (a) a SoMeDi manifesto in which the proposed solution is outlined including an overview of what is expected at the end of the project, and an overview of an initial architecture, (b) the summarized State of the Art of the different technologies envisaged in SoMeDi solution, (c) Beyond State of the Art that wants to be reached at the end of the project through the innovations proposed in SoMeDi, (d) the detected industry-wide needs and current pain points, (e) a preliminary set of more detailed application use cases that making used of the technologies and innovations proposed in SoMeDi DID tool solved some of the problems identified in d.

# 2. SOMEDI MANIFESTO

## Objective

SoMeDi has identified value mining of social and other digital user interactions as a viable business model mainly related to business trends such as Software-as-a-Service as well as proliferation of social CRM. The objective is to applied technological innovations in the area of artificial intelligence, opinion mining, big data to exploit the digital user interactions and transform them in Digital Interaction Intelligence.

The innovation areas of the project can be grouped into three areas based on the relation of social media and other types of digital interaction data. The areas are

M = monitoring

I = interaction, and

P = push.

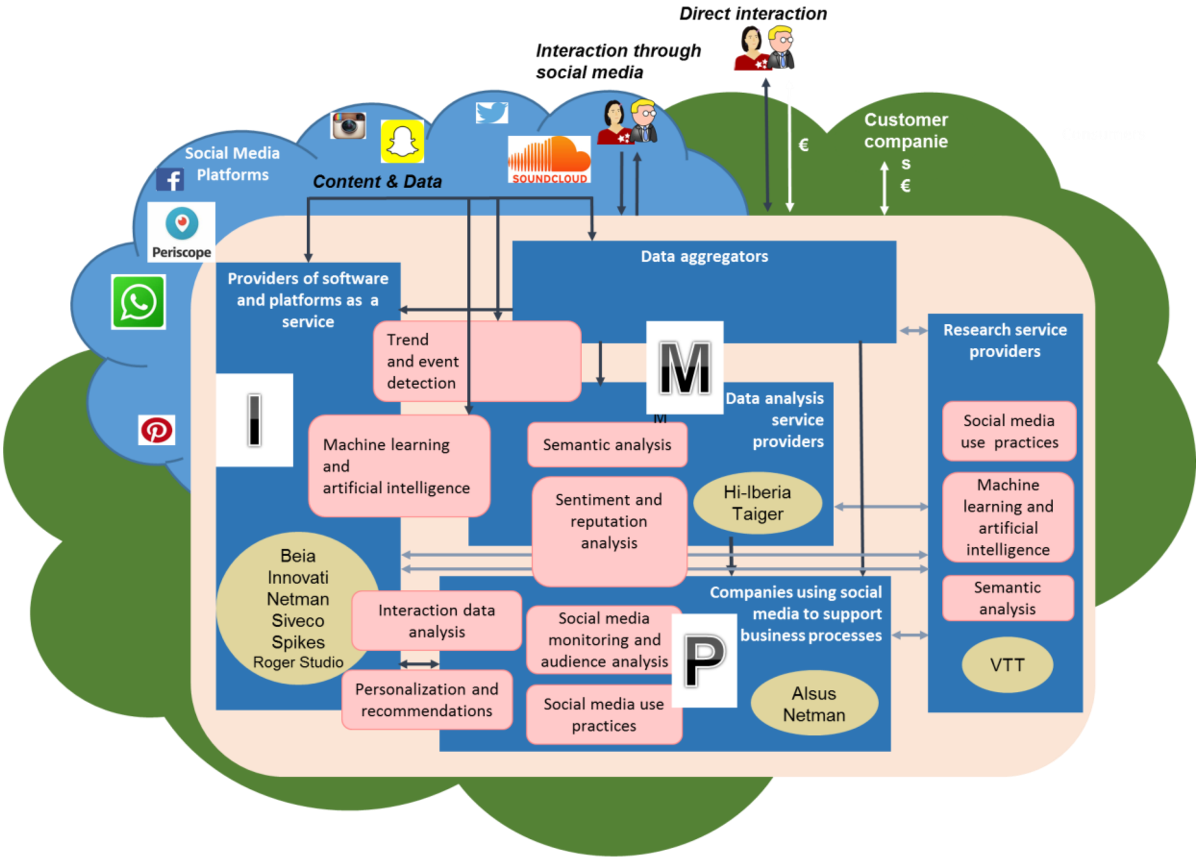

Direct interaction
Interaction through social media
Customer companie s
€
€
Consumers
Social Media Platforms
Periscope
Content & Data
SOUNDCLOUD
Data aggregators
Providers of software and platforms as a service
Trend and event detection
M
I
Data analysis service providers
Research service providers
Machine learning and artificial intelligence
Semantic analysis
Social media use practices
Sentiment and reputation analysis
Hi-Iberia Taiger
Machine learning and artificial intelligence
Beia Innovati Netman Siveco Spikes Roger Studio
Semantic analysis
Companies using social media to support business processes
Interaction data analysis
Social media monitoring and audience analysis
P
VTT
Personalization and recommendations
Social media use practices
Alsus Netman

FIGURE 1 - SOMEDI INNOVATIONS

Monitoring refers to applications where content and data that is being produced and published in different social media and other applications is aggregated and analysed to support various business functions. SoMeDi focuses on being able to process the data more efficiently and with better results for the business processes where conclusions and decisions will be made based on this data. In particular, the project will produce innovation in the following areas:

- Better sentiment analysis and opinion mining i.e. better understanding the informative value of the data and mapping sentiments to different aspects of products, services or companies, and taking into consideration the influence and networks of the opinion expressing people

- Time-series analysis, event detection and prediction based on social network data

Interaction refers to cases where a platform or software application providers closely analyse users' actions while they are using the offered platforms and software, and utilize this data at different time spans to improve the service. SoMeDi focus is on near real-time pattern recognition in multimedia data, on-demand dashboards, recommendation engines combining several data sources, segmentation of user groups, in-depth analysis of user interaction with various types of content or service feature and processing and analysis of multiple sources and clustering of postings.

Push refers to cases where the company actively looks for input through social media. Input from users in social media is used to support and guide decisions in core business processes like innovation, media content production and marketing. Main SoMeDi innovations are focused on two processes:

- *New product or service development* from harvesting ideas to quickly taking new products to the market using social media.

- *Marketing*, where the challenge is to know which channels to use and how to use them in the most effective way, as well as also how to utilize the data that is generated in the social media channels effectively. The data can be used to improve the use of social media in future marketing, and to guide in the future development of the product or service. We aim at getting data-driven support for taking actions on how, when and where to promote advertising and content. The ultimate aim is to be able to automate the interaction needed in social media channels.

## SoMeDi – Digital Intelligence Data Tool

The main output of SoMeDi project will be the DiD tool; the objective of this section is to provide an overview of the functionalities envisaged for the DID tool that will be refined along the project and with the specific requirements of the use cases.

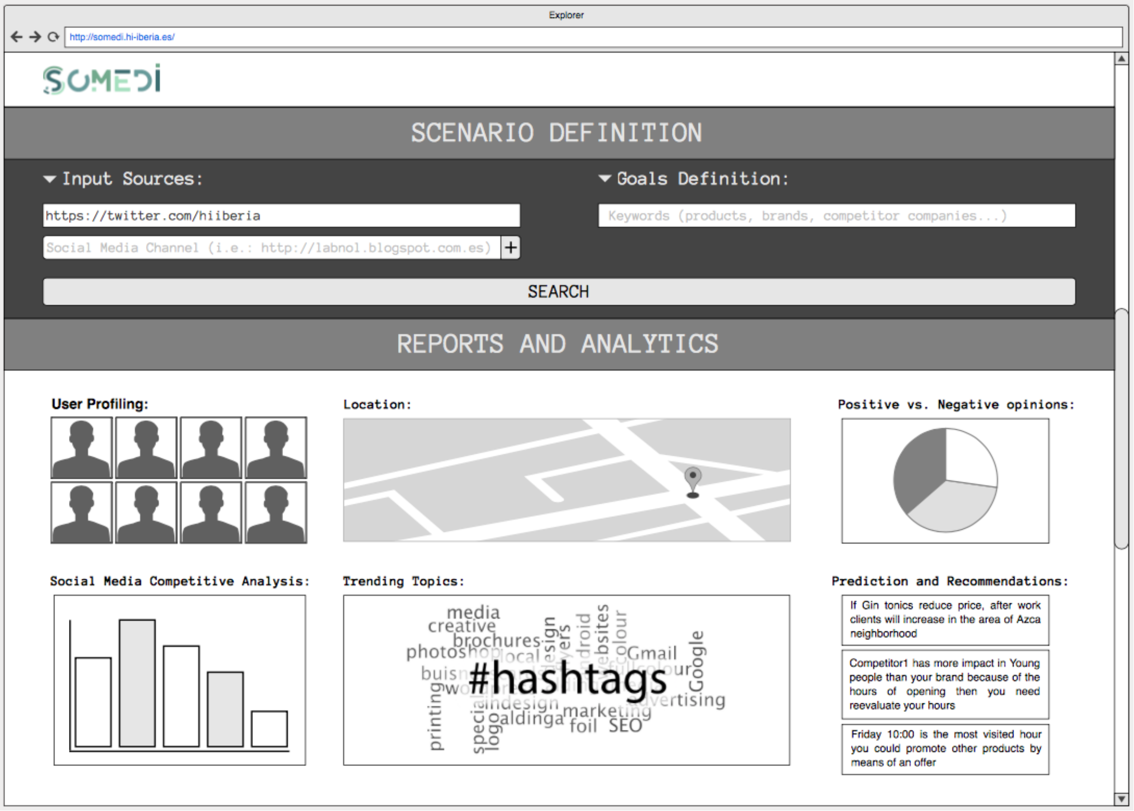
SoMeDi DID tool main user interface is composed by a personalized dashboard that will allow the user to access to different functionalities:

- 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.
  - 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).

- 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?
    - 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?
    - User Profiling: demographic (age, gender, location, etc) study based on tweets and community relationships of the users.
    - Influential users: Identify potential influencers to promote some products or services.
    - 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.
    - 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"

- 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.



http://somedi.hi-iberia.es/

SCENARIO DEFINITION

https://twitter.com/hiiberia

Social Media Channel (i.e.: http://labnol.blogspot.com.es)

Keywords (products, brands, competitor companies...)

REPORTS AND ANALYTICS

User Profiling:

Prediction and Recommendations:

If Gin tonics reduce price, after work clients will increase in the area of Azca neighborhood

Competitor1 has more impact in Young people than your brand because of the hours of opening then you need reevaluate your hours

Friday 10:00 is the most visited hour you could promote other products by means of an offer

FIGURE 2 - SKETCH OF SOMEDI PLATFORM

## Platform architecture envisaged for SoMeDi

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

Regarding the Use case for marketing purposes, the platform will link customer's opinion on social networks with the companies as a feedback in order to improve their marketing performance, so SoMeDi will be included as stated on Figure 3. SoMeDi structure – marketing use case. The addition of this service in the structure will allow companies to increase their influence on the related sector where decisions can be taken inspired from social opinion. SoMeDi will analyze data collected and create a dashboard where this information can be shown. Besides, an integration of this dashboard with the marketing campaign management system of the customer's company will be assessed.

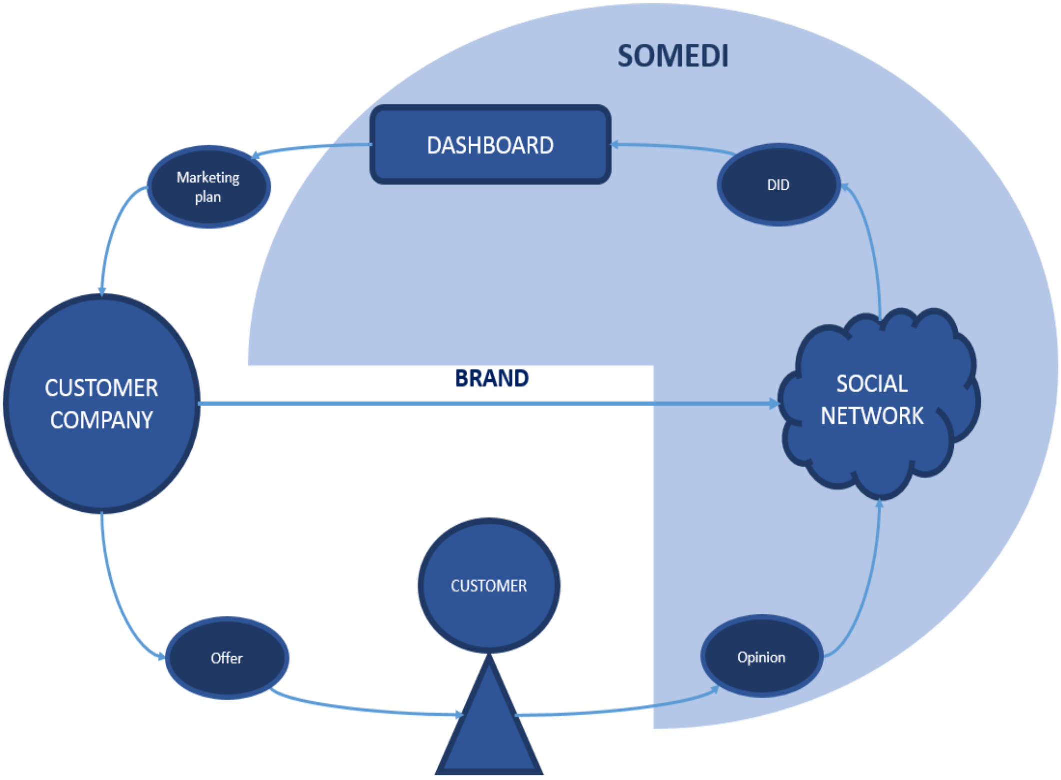


FIGURE 3 - SOMEDI STRUCTURE – MARKETING USE CASE

Also, SoMeDi will develop innovative tools (Figure4. SoMeDi Structure – Recruiting use case) to address the needs from Career Counseling Centers (CCC). Recent studies analyzed the situation of Romanian CCC programs, with the participation of 20.000 students from 24 universities. The research revealed concerning facts about the possibilities of feature graduates to find a job suited to their education. The platform will collect information from social media and recruiting platform and will match this information with candidates' needs and interests.

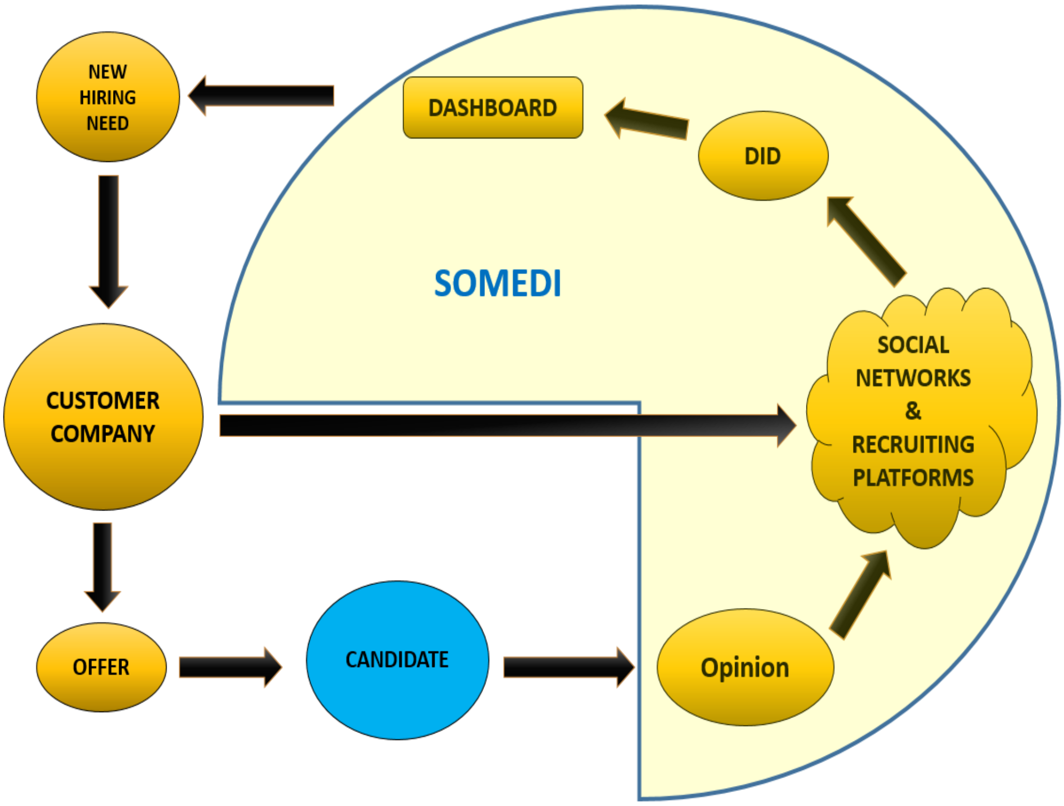


FIGURE 4 - SOMEDI STRUCTURE – RECRUITING USE CASE

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 (Representational State Transfer), which is a set of principles that define how Web standards, such as HTTP and URIs are supposed to be used. It is proposed because it does not leverage much bandwidth, and his lighter weight communications which supports the cloud-based APIs. REST has been used previously on social networking Web applications with satisfactory return. Furthermore, is designed for use over the Open Web which highlights it as a better choice for Web scale applications and cloud-based platforms. For secure connection, SSL Secure Connection has been suggested. SSL is a standard security technology for establishing and encrypted link between a server and a client, so it will be used for server robustness and client safety.

The view layer will be implemented by JavaScript, specifically with AngularisJS, an open-source web application framework maintained by Google and a community of individual developers and corporations to address many of the challenges encountered in application development./**/ The use of JavaScript involve using JSON (JavaScript Object Notation) for compatibility reasons and an easier development. JSON is a syntax for storing and exchanging data derived from the JavaScript scripting language which is the main format for data interchange used on the modern web. The NoSQL database proposed, MongoDB, use also JSON documents in order to store records and offers facilities with the use of REST.

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. Actually, Wildfly is free and open-source software, considered high-end on application runtime and deployment. The main strengths for the project are Wildfly's Web Socket support allow applications the ability to use optimized custom protocols and full-duplex communication with backend infrastructure which is useful in communicating with mobile device, supports the latest standards for REST based data access, including JSON, and it builds on many of best of breed standalone Open Source software, for example Hibernate. With the idea of interoperability with existing and emerging technologies, Spring will be used as application framework which provides a comprehensive programming and configuration model for Java-based enterprise applications – on any kind of deployment platform. It offers core support for dependency injection, transaction management, web applications and data access, among other. In addition, implementation by Spring imply the use of Maven, a software project management and comprehensive tool which make the build process easy, provides quality project information and above all allow transparent migration to new features. All technology proposed on the controller layer ensures the client the proper functioning of a high end application.

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. PHP is a server-side scripting language and a powerful tool for making dynamic and interactive web pages. The scenario definition tool can be totally build by use of these technologies but also incorporating the use of NoSQL databases like MongoDB and REST based data access with JSON message packaging.

As it is said above, a NoSQL database like MongoDB will be used on data access layer, so Hibernate has been proposed. On data issues, Hibernate offers a framework for both, SQL and NoSQL data bases which means more flexibility in case on future develop a different architecture on data access is required. There exist two models: Hibernate ORM (Object Relational Mapping framework), which enables developers to more easily write applications whose data outlives the application process and is concerned with data persistence as it applies to relational databases (via JDBC), and Hibernate OGM, that provides Java Persistence (JPA) support for NoSQL solutions. Hibernate OGM reuses ORM's engine but persist entities into a NoSQL data store instead of a relational database.

# 3. State of the Art of the technologies involved in SoMeDi

The feasibility and viability of value extraction from human digital interaction data has been amply demonstrated by large, mainly US and Asian, companies to the point where traditional services and software offers can be provided for free at the point of audience consumption. Current business trends are enabling the opportunity for smaller companies to also take part in this evolution.

A number of underlying global trends underpin the opportunities for the research in the SoMeDi project. The availability of significant quantities of digital data is resulting from human interactions with and through digital services. This is a direct consequence of the ubiquity of computation, communication and information services brought about both by the Internet and the full digitization of business and consumer services. While this trend is not new, the recent move to Software as a Service (SaaS) business models, Social Media empowered customer engagement and “Over-the-top” media delivery have put global reach in the hands of not just the big international industrial players, but also accessible to Small and Medium European businesses.

The developments in different fields give tools to succeed in this competitive market:

## Machine learning and Big Data

Machine learning aims at giving computers the ability to learn without being explicitly programmed[1]. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data[2]. The process of machine learning is similar to that of data mining as both systems search through data to look for patterns, and machine learning uses that data to improve the program's own understanding.

Supervised systems learn from sample annotations marked up by the user. A problem with these methods is that picking enough good examples is a non-trivial and error-prone task. In order to tackle this problem unsupervised systems employ a variety of strategies to learn how to annotate without user supervision, but their accuracy is still limited.

Next to these specific state-of-the art technological components, several available libraries have emerged and matured in recent years that allow for digital interaction intelligence, both freely available such as Apache Storm, as well as via commercial models, such as IBM Watson or Microsoft Cortana Analytics. Moreover, platforms are being built on top of these analytics libraries making it even easier to create intelligence process with them. An example of this is the Tengu platform, which is available as an experimentation platform.

SoMeDi partners have access to and knowledge on the usage of these platforms and intend to make use of them for improving the social media and digital interaction data insights and strategies.

Some examples of current available tools are:

- Rapid Miner

RapidMiner[3] is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization.

[1] Phil Simon (March 18, 2013). *Too Big to Ignore: The Business Case for Big Data*. Wiley. p. 89. ISBN 978-1-118-63817-0.
[2] Ron Kohavi; Foster Provost (1998). "Glossary of terms". *Machine Learning* 30: 271–274.
[3] https://rapidminer.com/

RapidMiner is developed on an open core model. The RapidMiner (free) Basic Edition is limited to 1 logical processor and 10,000 data rows is available under the AGPL license. Commercial pricing starts at $2,500 and is available from the developer.

RapidMiner is a good solution for quick prototyping and validation of machine learning pipelines. Rapidminer moreover offers a wide set of externally developed plugins that extends the scope of the tool I the field of Linked Data, Semantic Web, Text Mining, Big Data and Deep Learning just to cite a few. However, for a production environment the Open Source license offer very limited possibilities. In the framework of the SOMEDI project Rapidminer can be used to test different machine learning approaches and once selected the more suitable solution implement it using a different framework such as the machine learning libraries developed on top of bigdata platforms.

- Weka

Weka[4] is probably the most famous Machine Learning tool, developed by the University of Waikato, it is a general purpose Machine Learning tool similar to RapidMiner. It is Open source, so it does not have the restrictions of RapidMiner, but it has a less intuitive interface, which requires longer learning process for non-expert users.

Weka implements a considerable amount of the most common machine learning algorithms so that the user can quickly develop a processing pipeline to test. In this sense Weka is similar to RapidMiner. Moreover, starting from version 3.8 Weka provides access to new packages for distributed data mining. The first new package provides base "map" and "reduce" tasks that are not tied to any specific distributed platform. A second package provides Hadoop-specific wrappers and jobs for these base tasks. A third, called distributedWekaSpark, provides Spark-specific wrappers.

- BigML

BigML is an SaaS platform for creation and execution of machine learning pipelines. Similar to RapidMiner, the user can build a processing pipeline, which is executed remotely in the BigML platform.

### Machine Learning and Big Data: Batch Processing

In the big data domain there are nowadays plenty of platforms offering machine learning capabilities, and novel solutions are continuously under developed thanks to an active community of developers. In the case of Big Data we need to differentiate between batch and stream machine learning processes: in the batch processes the data is assumed to be already available: this is the traditional machine learning approach, where data can be divided into training and testing set for validation of the model. This type of analysis was also the first type of processing available in the initial Big Data solutions. On the other hand, Stream processing capabilities made their appearance later in the Big Data ecosystem as a way to process data on real-time (with some exceptions like Spark), consequently some Machine learning algorithms have been adapted to be able to process stream data.

Hadoop (Map/Reduce) and associated ML tools:

The rise of the Big Data research line could be identified in the publication of two seminal papers from Google, one paper[5] describing a distributed file system with fault tolerance capabilities (called Google File System) and another[6] describing a distributed processing framework called Map Reduce.

[4] http://www.cs.waikato.ac.nz/ml/weka/

[5] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. 2003. The Google file system. In Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03). ACM, New York, NY, USA, 29-43.

[6] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), 107-113.

The two systems described in the papers were developed at Google to tackle the problem of storing the index of the crawled webpages and the ability to recalculate the index in a timely fashion (as a batch process running during low traffic hours). The Hadoop ecosystem[7] was initially developed at Yahoo as an extension of Apache Nutch[8] (an open source web crawler) with the intention to implement the systems described in the two Google papers.

In 2006 the Hadoop project became independent from Nutch and was released open source as an Apache incubator project. Initially the distributed file system and the processing framework were part of the Hadoop project, and later on where divided in 2009 as separate subprojects respectively Hadoop Distributed File System (HDFS)[9] and Hadoop MapReduce.

From the release of the first version, a number of companies gradually adopted the framework in their production systems. In some cases driven by real use and in some other cases driven by the hype that was built around the Big Data buzzword[10], expecting it to be the panacea for all their illnesses.

Nowadays HDFS is the file system underlying most of big data projects. HDFS is a distributed file system developed in Java able to provide a reliable file system on commodity hardware. HDFS is scalable and fault-tolerant: in case more space is required it is possible to add nodes to the HDFS clusters; in case that one or more nodes fail, the HDFS is able to provide the portion of data that became isolated by maintaining distributed replicas of data.

The architecture of a typical HDFS cluster is represented in **¡Error! No se encuentra el origen de la referencia.**: there are two types of nodes: Data Nodes and Name Nodes: the first is used to store physically the data, while the latter is used to maintain a file table that is used to access the files.

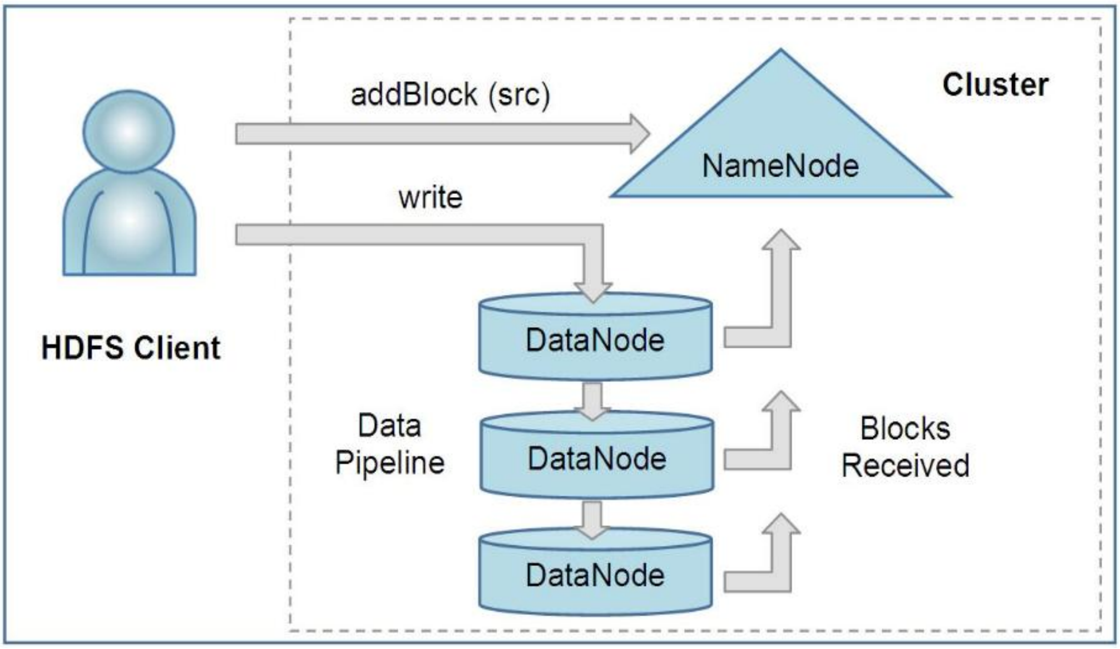


FIGURE 5 - ARCHITECTURE OF AN HDFS CLUSTER

[7] http://hadoop.apache.org/

[8] http://nutch.apache.org

[9] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. 2010. The Hadoop Distributed File System. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST) (MSST '10). IEEE Computer Society, Washington, DC, USA, 1-10.

[10] https://hbr.org/2013/12/you-may-not-need-big-data-after-all

The developers of HDFS have improved the file system over time with the addition of several features described in **¡Error! No se encuentra el origen de la referencia.**:

| Feature | Description |
|---|---|
| **Rack awareness** | Considers a node's physical location when allocating storage and scheduling tasks. |
| **Minimal Data Motion** | Hadoop moves compute processes to the data on HDFS and not the other way around. Processing tasks can occur on the physical node where the data resides, which significantly reduces network I/O and provides very high aggregate bandwidth. |
| **Health Utilities** | Dynamically diagnose the health of the file system and rebalance the data on different nodes. |
| **Rollback mechanism** | Allow operators to bring back the previous version of HDFS after an upgrade, in case of human or systemic errors. |
| **Minimal Intervention** | HDFS requires minimal operator intervention, allowing a single operator to maintain a cluster of 1000s of nodes. |
| **High Availability** | Provides redundancy of the Name Node to supports high availability (HA). |

TABLE 1 HDFS FEATURES

The main issue with HDFS is that does not perform well with small files, it has been designed for storage of big sequential indexes originally and this limitation results from this initial assumption.

Regarding distributed processing and analytics frameworks, a survey of available tools and libraries is presented in a survey[11].

Hadoop Map Reduce is the Open Source implementation of the Map Reduce framework described in the Google paper. As the name suggests, it is composed by two phases: Map and Reduce.

The Map phase takes as input a set of objects and generates a set of Key Values pairs. The criteria by which these Key Values pairs are created depend on the implementation and on the problem to solve. As an example, we consider the typical word count problem represented in **¡Error! No se encuentra el origen de la referencia.** in the Map phase each node takes as input a set of documents (input), each document is tokenized into single words (keys) and for each word the Map phase counts how many times it appears in the documents processed by the node (values).

After the Map phase there is an intermediate phase, which is called Shuffling, where all the outputs of the Map processes are forwarded to the Reduce phase. This shuffling process is necessary because it forwards the elements of the Key Value list by creating new lists composed of all the elements with the same Key. Then these new lists are forwarded each one to a different node and the Reduce phase can be performed.

The Reduce phase takes as input a list of Key Value pairs and creates a new list of Key Values pairs. In the case of the word count problems, the Reduce phase takes the values associated with each Key and returns the sum.

[11] A survey of open source tools for machine learning with big data in the Hadoop ecosystem, Sara Landset, Taghi M. Khoshgoftaar, Aaron N. RichterEmail author and Tawfiq Hasanin. Journal of Big Data20152:24

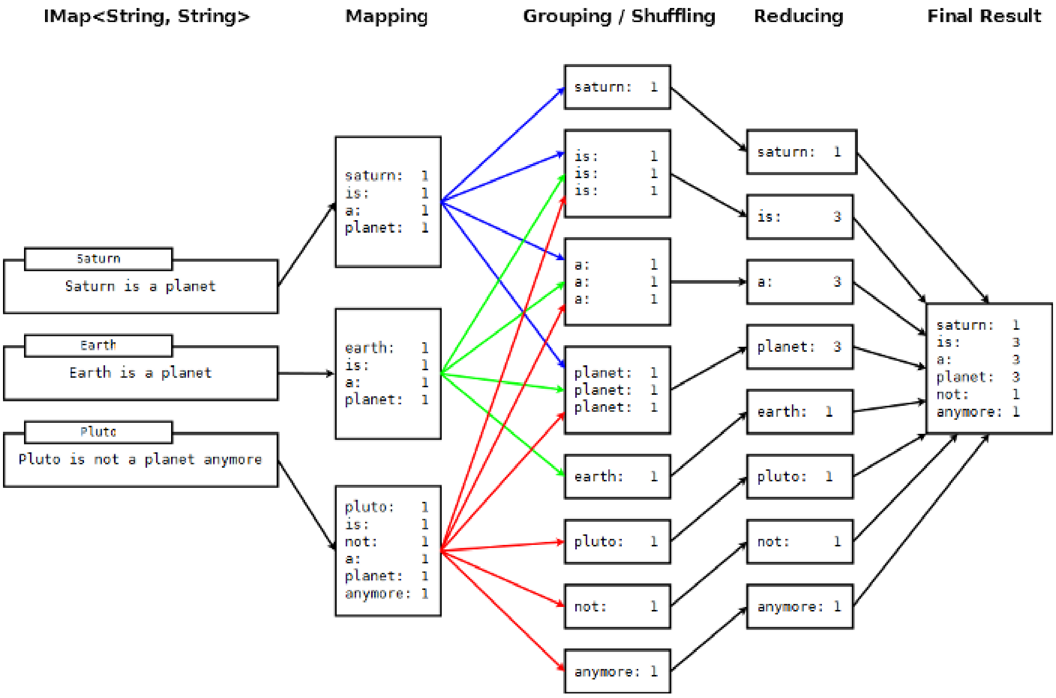


FIGURE 6 - THE WORD COUNT PROBLEM SOLVED WITH MAP REDUCE

The main issue with the Hadoop implementation of the Map Reduce process is that the output of the Map phase is stored physically in the HDFS resulting in a bottleneck.

These frequent disk intensive operations can become very expensive in terms of latency, computational resources, and network bandwidth. These issues become more apparent in cases where it is necessary to update models with new data, which is normally the case in real-world ML applications.

This makes Hadoop processing framework outdated or less performant with respect to novel systems such as Spark, Flink and H2O.

A set of frameworks for ML have been developed on top of Hadoop Map Reduce aiming at efficient distribution of ML tasks.

In this group we can identify Apache Mahout[12] as the representative tool for ML with Hadoop; even if, it is worth to notice that from April 2015, the development team behind Mahout is gradually removing support to algorithms running on native Map Reduce environments such as Hadoop, due to the inherent inefficiencies of the framework, moving towards a math environment called Samsara[13] which includes statistical and algebraic operations. This change on the focus has been motivated by the intention to provide a platform allowing building distributed processing algorithms, rather than providing a set of ready-made ones.

Apache Mahout is both a single machine and a distributed machine learning library that can be integrated with different processing frameworks: Spark, H2O and Flink.

Previously to its change of focus, Mahout provided a set of ML algorithms that are translated (where possible) into distributed processes for the underlying processing framework, however the set of ready-to-use algorithms depends of the processing engine. Moreover, the set of ready-to-use

[12] https://mahout.apache.org/

[13] https://mahout.apache.org/users/environment/out-of-core-reference.html

algorithms released with Mahout has become smaller with respect to the previous versions, resulting in a ML tool that can address only specific tasks out of the box. The set of algorithms that were supported when Mahout was still supporting Hadoop were able to address classification, collaborative filtering, clustering, dimensionality reduction, topic modelling, Tf-Idf (Term frequency - Inverse document frequency), and others while now the only types of algorithms supported are collaborative filtering, naive Bayes classification and dimensionality reduction. On the other hand it offers a more flexible solution to specific machine learning problems, allowing the composition of distributed algorithms from a set of basic operations. Critics to Mahout were raised due to the difficulty with configuration, integration and development of new algorithms [14][15][16], however, older releases of Mahout were applied successfully in several production environments[17][18] [19]. At the moment of writing, due to the already mentioned change of focus, it may be early for a comprehensive analysis of the ML potential of the Mahout tool. However, it needs to be kept into consideration due to its flexibility in the creation of customized ML algorithms.

### Spark and associated ML tools

Spark[20] was developed initially at the University of California, Berkeley[21] moving later on to become an Apache project. The framework is based on Map Reduce. The main motivation behind the creation of Spark was to address the disk intensive inefficiencies of Hadoop for many of the common cases. Supporting iterative computation and improving on speed and resource utilization by adopting an in-memory computation model have managed to solve this issue in most cases; in particular in the case of ML tasks which iterate over the data.

The main data abstraction used in Spark is called Resilient Distributed Datasets (RDD), which consists of a read-only distributed in-memory storage providing naturally a fault tolerance mechanism without the need for physical replication of data on disk.

RDDs can be the result of an input process, as for example importing data from HDFS file system or the result of a transformation process from others RDDs to new RDDs. Spark allows another type of operation on RDDs that is called action: this type of operation is used to produce final results from a RDD.

The architecture of Spark is composed by a Driver node and a set of Worker nodes as shown in **¡Error! No se encuentra el origen de la referencia.**: the Driver node is where the program logic is executed, while the worker nodes store and perform operations on the RDDs. Whenever is possible, there will be no data exchange between workers nodes. This can be guaranteed in some cases: for this reason the transformations are divided into two groups narrow and wide transformations, the first set of transformations guarantees that no data between the worker nodes is exchanged (filter,

[14] Wegener D, Mock M, Adranale D, Wrobel S. Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters. In: 2009 IEEE International Conference on Data Mining Workshops; 2009. pp. 296–301

[15] Lin J, Kolcz A. Large-scale machine learning at twitter. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data; 2012. pp. 793–804.

[16] Zeng C, Jiang Y, Zheng L, Li J, Li L, Li H, Shen C, Zhou W, Li T, Duan B, Lei M, and Wang P. FIU-Miner: A Fast, Integrated, and User-Friendly System for Data Mining in Distributed Environment. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013. pp. 1506–9

[17] Jack K. Mahout becomes a researcher: Large Scale Recommendations at Mendeley. In: Big Data Week, Hadoop User Group UK; 2012

[18] Sumbaly R, Kreps J, Shah S. The big data ecosystem at LinkedIn. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13); 2013. pp. 1125–34

[19] Metz C. Mahout, There It Is! Open Source Algorithms Remake Overstock.com. Wired Magazine. 2012. http://www.wired.com/2012/12/mahout/. Accessed 18 Dec 2014

[20] https://spark.apache.org/

[21] Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: Cluster Computing with Working Sets. In: Proceedings of the 2nd USENIX conference on hot topics in cloud computing; 2010

map, sample,...), while for the second type of transformations there is the need for exchanging data between worker nodes (sort, group by, join,...).

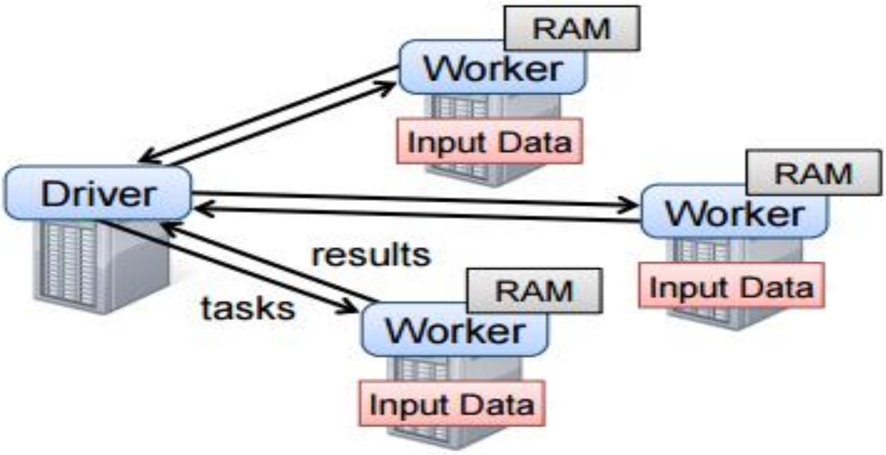


FIGURE 7 - SPARK ARCHITECTURE

Using this approach, the intermediate results of the Map Reduce computations, which are the reason for the main performance issues with Hadoop systems, are stored in the distributed memory, significantly cutting down on the number of read and write operations on the file system and data exchange between nodes, as represented in **¡Error! No se encuentra el origen de la referencia.**.

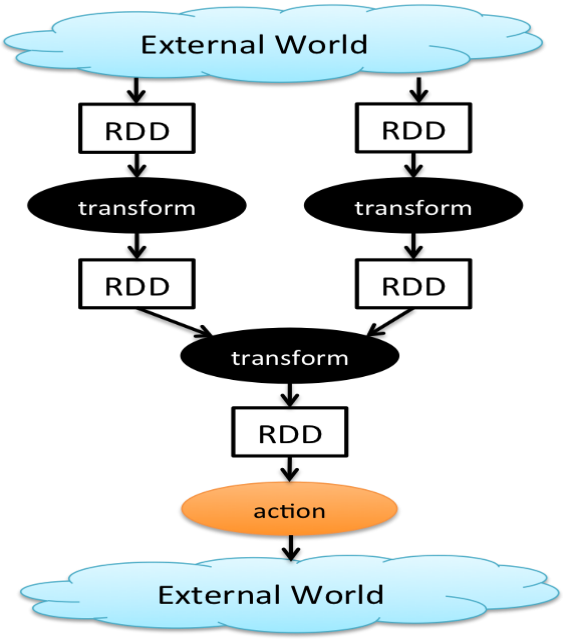


FIGURE 8 - RDD PROCESSING PIPELINE

Spark, additionally to Hadoop adapts to different programming languages (Java, Python, Scala and R), making it easier to develop and adapt to different environments.

However Spark is not perfect either. Some concerns about Spark's approach have been identified in the distribution of data across nodes[22] . Data transfers take place throughout the network, and because of the job isolation mechanism present, only one driver can serve requests to all of its RDDs, potentially leading to a bottleneck within the network when there are multiple requests to multiple nodes.

Besides the already mentioned Mahout library, Spark also has its own machine learning library called MLlib[23].

MLlib is a library similar to the early versions of Mahout (before change of focus towards Samsara): it provides the same set of ML algorithms plus topic modelling and frequent patterns mining. MLlib takes advantage of Spark architecture for iterative batch processing, stream (micro-batch) processing and in-memory caching of intermediate results and it is able to deliver much better performances than Mahout.

The advantages of MLlib also consist of its easy configuration and deployment, due to the fact that is tied to the RDD model and therefore does not require all the adaptation and integration required by Mahout.

Some critics on the library highlight the fact that the behavior in non-ideal situations (very big or very small vectors of data) have raised some issues[24] . Other studies[25] point out the fact that the tool has performance issues due to slow convergence of some algorithms. However, even if it is younger than Mahout, MLlib has been deployed in many production environments[26].

Additionally, Spark framework provides a concept called ML pipeline[27], built on top of Spark-SQL[28] library, allowing users to set up, build and execute Machine Learning processing pipelines.

MLlib provides standardized APIs for machine learning algorithms which makes easier to combine MLlib algorithms into a processing pipeline. ML pipelines also allow creating non-linear processing pipelines, as long as they are defined as Directed Acyclic Graphs (DAG).

Alike Mahout, there has been a shift of focus in the latest releases of the MLlib library, moving from RDD data models to DataFrames. However, support for RDD models is still provided, but no new algorithms will be developed to support RDD. The development of new algorithms will focus only on DataFrames, supporting this way the integration with ML pipelines.

### FLINK AND ASSOCIATED ML TOOLS

A relatively new approach supporting stream and batch processing is Apache Flink[29] originated from the Stratosphere research project from University of Berlin[30].

[22] Singh J. Big Data Analytic and Mining with Machine Learning Algorithm. Int J Inform Comput Technol. 2014;4(1):33–40

[23] http://spark.apache.org/mllib/

[24] Koutsoumpakis G. Spark-based Application for Abnormal Log Detection. Thesis, Uppsala University; 2014

[25] Zheng J, Dagnino A. An initial study of predictive machine learning analytics on large volumes of historical data for power system applications. In: 2014 IEEE International Conference on Big Data; 2014. pp. 952–59

[26] http://spark.apache.org/powered-by.html

[27] http://spark.apache.org/docs/latest/ml-pipeline.html

[28] http://spark.apache.org/docs/latest/sql-programming-guide.html

[29] http://flink.apache.org/

[30] Alexandrov A, Bergmann R, Ewen S, Freytag JC, Hueske F, Heise A, Kao O, Leich M, Leser U, Markl V, Naumann F, Peters M, Rheinländer A, Sax MJ, Schelter S, Höger M, Tzoumas K, Warneke D. The Stratosphere platform for big data analytics. VLDB J Int J Very Large Data Bases. 2014;23(6):939–64

Flink combines database, stream processing and Map Reduce technology to keep on one side the processing of user defined functions, complex data types, and the scalability of map reduce systems, and to offer on the other side the declarative, independence, and automatic optimization of database technology, allowing data analytics engineers to focus on the analytic problem and depart from programming issues. Apache Flink implements an architecture, basic programming models, stream programming models and advanced concepts, such as iterations and automatic optimization. It offers capability for both batch and stream processing, thus allowing for the implementation of hybrid (stream and batch) processing architectures (such as Lambda-architecture[31] and the most recent Kappa-architecture[32]).

Similarly to Spark architecture as represented in **¡Error! No se encuentra el origen de la referencia.**, Flink is composed by a JobManager and many TaskManager nodes: the JobManager executes the program logic, while the TaskManager executes the atomic operations on data.

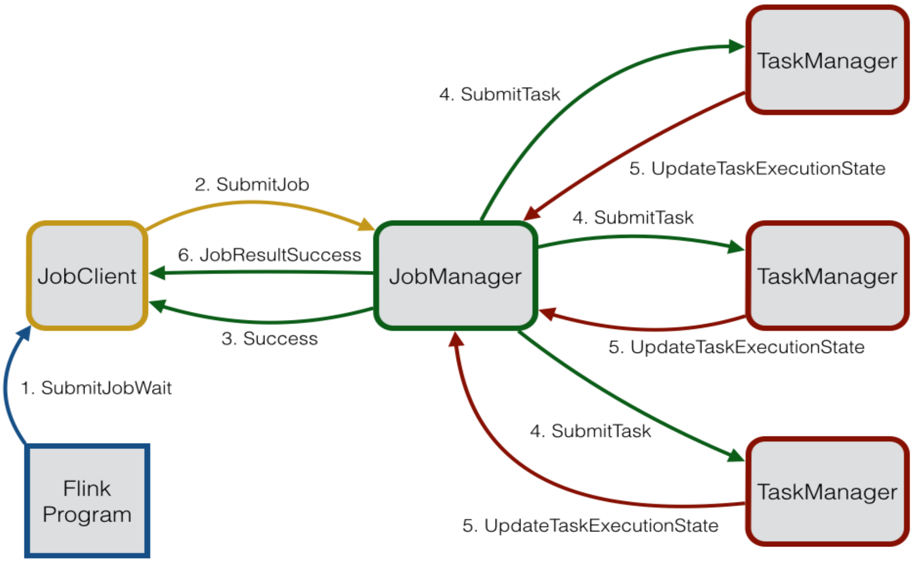


FIGURE 9 - FLINK'S ARCHITECTURE

Flink can be both integrated in the Resource Manager of Hadoop or can run independently. Flink's processing framework is based on transformations that are applied to the data collections in the nodes. These transformations can be generic Map and Reduce functions (allowing execution of Flink code using traditional Map Reduce approach without modification), but also some more specialized functions such as join, group, and iterate. Flink execution model includes a cost-based optimizer that automatically selects the best execution strategy for each process.

Even if Flink is relatively young with respect of Hadoop and Spark, some Machine Learning tools are available for the framework. As already mentioned a portion of the algorithms available in Mahout has been implemented to support Flink.

However Flink project has its own machine learning library, called Flink-ML[33].

[31] Marz N, Warren J. Big data: principles and best practices of scalable realtime data systems. Manning Publications; 2015

[32] Jay Kreps Questioning the Lambda Architecture, O'Reilly July 2014 https://www.oreilly.com/ideas/questioning-the-lambda-architecture

[33] https://ci.apache.org/projects/flink/flink-docs-release-1.2/dev/libs/ml/index.htm

With respect to Mahout and MLlib, the list of algorithms is small: Support Vector Machines (SVM) and multiple linear regression are the two only supervised algorithms available, while K-nearest neighbours is the only unsupervised algorithm implemented.

An Alternating Least Squares (ALS) recommendation algorithm is also available as well as some data pre-processing methods and support utilities such as cross validation.

Similarly to Spark, the library provides the feature of building processing pipelines, in a similar fashion than the well-known scikit-learn[34] library developed for Python programming language.

On the roadmap for future implementation there are additional ML algorithms that will become available gradually.

Since the Flink processing framework and the related machine learning libraries are relatively young, there are not extensive performance analysis. An old report is available[35], where the authors compared Spark with the Stratosphere project (former name of FLink), by performing an analysis both from a theoretical perspective and a practical one. As a result, Spark was found to be performing better in the areas of fault tolerance and handling of iterative algorithms, while Flink demonstrated better optimization mechanisms and better integration with other projects.

In terms of practicality, Flink is more resources intensive, but faster than Spark.

The Flink team also published benchmark results[36] on some machine learning tasks, such as Page Rank, reporting that Flink's execution was significantly faster than Spark.

However, as already mentioned, these reports are already out-dated and some more independent comparisons are needed.

### H2O AND ASSOCIATED ML TOOLS

H2O[37] is an open source project (with additional support for Enterprise editions) that, more than a distributed processing framework, can be considered as a complete analytical product on its own: it provides a distributed processing engine, but also data pre-processing, analytics, math, machine learning libraries and evaluation tools. As well as Spark, it offers support for Java, R, Python, and Scala languages and it is able to execute Spark processes by providing integration with Spark processing framework through its Sparking Water[38] library; H2O is also able to execute its own processing models on top of Spark and Storm.

H2O's processes data in-memory using multiple execution methods, an approach similar to Flink.

The generic approach to deploy a job in H2O is called Distributed Fork/Join which is a divide and conquers technique adapting well to massively parallel tasks. This approach breaks down a processing job into smaller jobs, which are executed in parallel, resulting in dynamic fine-grain load balancing for Map Reduce jobs as well as graphs and streams.

Regarding Machine Learning aspects, Mahout Library offers implementation of most of its machine library algorithms for H2O.

[34] http://scikit-learn.org/stable/

[35] Ni Z. Comparative Evaluation of Spark and Stratosphere. Thesis, KTH Royal Institute of Technology; 2013

[36] Metzger R, Celebi U. Introducing Apache Flink—A new approach to distributed data processing. In: Silicon Valley Hands On Programming Events; 2014

[37] http://www.h2o.ai/h2o/

[38] http://www.h2o.ai/sparkling-water/

However, H2O platform is shipped with a ready-made machine learning module that in addition to traditional machine learning algorithms offer a set of tools for deep neural networks, which is nowadays a hot topic due to the important advances in several machine learning problems and the enormous hype that the media generated as a result of it.

As already mentioned, H2O is able to run code written in Java, Python, R and Scala. However, users with no programming expertise can still define processing pipelines via the Web interface.

At of the time of writing, the ML tools offered with H2O are able to address a relatively wide set of ML tasks, including classification, clustering, generalized linear models, statistical analysis, ensembles, optimization tools, data pre-processing options and deep neural networks. On the roadmap for future implementation there are additional algorithms and tools from the categories already mentioned as well as recommendation and time-series analysis and prediction.

### Machine Learning and Big Data: Stream Processing

Stream processing in Big Data systems is a different problem with respect to traditional batch analytics. While batch processing focuses on the offline analysis of huge and almost static datasets, stream processing focuses on the fast distributed analysis of often smaller but highly dynamic data sets.

A stream analytical process is traditionally defined as a Directed Acyclic Graph (DAG) where each node represents a processing step that receives a set of events, transform them according to some programming logic and forward the results to the next processing step (node) in the graph.

In a stream processing architecture, typically, there is a listener that waits for events submitted into a message queue, a file system a database or other event forwarding mechanism. Once the listener receives a new event to analyse, it submits the event to the nodes that are performing the computation; once concluded the computation, these nodes forward the results to the following nodes and so on until the DAG has reached the final nodes. At this stage the processing pipeline is concluded and the resulting data is obtained.

In a theoretical stream processing framework, as soon as an event is received it is processed in a continuous fashion. However, in reality this is not always the case: there are two distinct approaches in developing a stream processing mechanism. The first approach is called run-time stream processing and follows a process similar to the one described above: an event is processed upon its arrival, while the second is called micro-batch processing. In this second approach, while receiving the new data, the listener bundles together a bunch of events into a data model and forwards it to the processing pipeline for a small batch processing of all the events in the bundle.

Both approaches have advantages and disadvantages. Real-time streaming takes the stream as it is and processes the events upon their arrival without delay, reaching a performance in latencies which is normally better then micro-batches based approaches. As a disadvantage, real-time streaming systems have usually lower throughput and fault-tolerance is more expensive since it has to consider every event. Moreover, load balancing can become an issue, resulting in the overload of some nodes (as an example nodes taking long time to process events coming at high speed, resulting in a bottleneck originated from the slow nodes).

Dividing streams into micro-batches reduces system expressiveness: some operations are hard to implement with respect to pure real-time analysis because they need to operate on the entire micro-batch instead of on a single event. On the other hand, fault tolerance and load balancing are much simpler as it has to be applied punctually at micro-batch level rather than every single event.

Also, it is an important remark to point out that every real time stream-processing framework can behave like a micro-batch streaming framework, but not the other way around.

There are many practical use cases, where the need for batch processing and stream processing are both required: sometimes motivated by the performance of pure stream processing framework or because of the need for combining stream data with historical data. The most well-known architectures designed to combine batch and stream analysis are the Lambda-architecture[39] and the more recent Kappa-architecture[40]. These architectures combine stream and batch processing layers focusing on the synchronization of the results of the processing pipelines.

### Spark and associated ML tools

Spark originally was developed to overcome some of the limitations of the Hadoop architecture and implementations, especially in the case of data shuffling and storage of intermediate results, which has been proved to be inefficient in Hadoop[41].

The main data abstraction used in Spark is called Resilient Distributed Datasets (RDD), which consists of a read-only distributed in-memory storage providing naturally a fault tolerance mechanism without the need for physical replication of data on disk.

RDDs can be the result of an input process, as for example importing data from HDFS file system, from listening on a message queue, or the result of a *transformation* process from others RDDs to new RDDs. Spark allows another type of operation on RDDs that is called *action*: this type of operation is used to produce final results from a RDD.

During the development of the project, a streaming library, called Spark streaming[42], has been added to the Spark ecosystem in order to provide a tool to support stream-processing tasks.

Spark approach to stream processing is defined as an extension of the Spark framework allowing data to be ingested from many sources like Kafka[43], Flume[44], Kinesis[45], or TCP sockets.

Spark streaming receives the data as input and creates as output a set of micro-batches of data that are processed using traditional Spark batch processing framework. In order to represent the stream of data, Spark Streaming provides an abstraction called discretized stream or DStream. Internally, a DStream is represented as a sequence of traditional Spark's RDDs so that they can be processed seamlessly with the Spark framework.

Obviously, this approach cannot be considered as pure real-time stream processing, since the streams are converted into batches of data and then processed in a way similar to the traditional batch processing fashion. However, this approach is still suitable in many real-world situations, such as clickstream analysis or event detection.

As described in a Berkeley technical report[46] the use of micro-batches makes load balancing easier and is more robust to node failures. In the same survey, the authors point out that even if the

[39] Sds Marz N, Warren J. Big data: principles and best practices of scalable realtime data systems. Manning Publications; 2015

[40] Jay Kreps Questioning the Lambda Architecture, O'Reilly July 2014 https://www.oreilly.com/ideas/questioning-the-lambda-architecture

[41] Jiang, D., Ooi, B.C., Shi, L. and Wu, S., 2010. The performance of MapReduce: an in-depth study. Proceedings of the VLDB Endowment, 3(1-2), pp.472-483

[42] http://spark.apache.org/streaming/

[43] https://kafka.apache.org/

[44] https://flume.apache.org/

[45] https://aws.amazon.com/es/kinesis/streams/

[46] Zaharia M, Das T, Li H, Hunter T, Shenker S, Stoica I. Discretized Streams: A Fault-Tolerant Model for Scalable Stream Processing. University of California at Berkeley Technical Report No. UCB/EECS-2012-259; 2012

approach is slower than others true stream processing frameworks, the latency can be minimized enough to be acceptable for most real use cases.

Spark streaming supports also windowed computations, allowing converting discretized streams into windows of streams as result of a user-defined transformation.

Spark also provides tight integration between its streaming and batch layers, making it naturally fit for hybrid architectures such as the previously mentioned Lambda-architecture.

Regarding machine learning tasks, Spark streaming can take advantage of the algorithms present in the MLlib[47] library: streaming machine learning algorithms (e.g. Streaming Linear Regression, Streaming KMeans, …), which can be used for simultaneously learn from the streaming data.

Also hybrids models are allowed, permitting to use MLlib algorithms designed for batch processing: it is possible to learn a model under a batch processing fashion and then apply the model on the incoming stream of data.

MLlib provides standardized APIs for machine learning algorithms which makes easier to combine MLlib algorithms into a processing pipeline. ML pipelines also allow creating non-linear processing pipelines, as long as they are defined as Directed Acyclic Graphs (DAG).

There has been a shift of focus in the latest releases of the MLlib library, moving from RDD data models to DataFrames. However, support for RDD models is still provided, but no new algorithms will be developed to support RDD, the development of new algorithms will focus only on DataFrames, supporting this way the integration with ML pipelines.

Spark framework provides also a concept called ML pipeline, built on top of Spark-SQL[48] library, allowing users to set up, build and execute Machine Learning processing pipelines.

### FLINK AND ASSOCIATED ML TOOLS

A relatively new approach supporting stream and batch processing is Apache Flink[49] originated from the Stratosphere research project from University of Berlin[50], combining database technology, stream processing technology and Map Reduce technology, to keep on one side the processing of user defined functions, complex data types, and the scalability of map reduce systems, and to offer on the other side the declarativity, independence, and automatic optimization of database technology, allowing data analytics engineers to focus on the problem and depart from programming issues. Apache Flink implements an architecture, basic programming models, stream programming models and advanced concepts, such as iterations and automatic optimization. It offers capability for both batch and stream processing, thus allowing for the implementation of hybrid architectures (such as the already mentioned Lambda and Kappa architectures).

Like Spark, Flink also offers iterative batch as well as pure stream processing options, via their streaming API which is based on individual events, rather than the micro-batch approach that Spark uses. This is a similar model that other frameworks (such as Storm) use for true real-time processing. Connectors are offered which allow for processing data streams from Kafka, RabbitMQ, Flume, Twitter, and user-defined data sources. Results are provided to a sink element, which decides what to do with the resulting data (e.g. store the results in HDFS).

[47] http://spark.apache.org/mllib/

[48] http://spark.apache.org/docs/latest/sql-programming-guide.html

[49] http://flink.apache.org/

[50] Alexandrov A, Bergmann R, Ewen S, Freytag JC, Hueske F, Heise A, Kao O, Leich M, Leser U, Markl V, Naumann F, Peters M, Rheinländer A, Sax MJ, Schelter S, Höger M, Tzoumas K, Warneke D. The Stratosphere platform for big data analytics. VLDB J Int J Very Large Data Bases. 2014;23(6):939–64

Flink provides windowed data abstractions such as aggregations and data-driven windows that are activated by the evaluation of a condition (e.g. a threshold to decide the activation of the windowing process) as well as other operators such as joins on windows of data.

A basic concept in Flink used to define stream processing pipelines is called Data Stream programs. These are simple programs embedding the processing logic, taking as input a set of events from a data source, applying transformations on the input and writing the results into data sinks.

Flink project comes with its own machine-learning library, called Flink-ML[51].

Compared to other ML libraries such as Spark MLlib, the list of available algorithms is relatively small: Support Vector Machines (SVM) and multiple linear regressions are the two only supervised algorithms available, while K-nearest neighbours is the only unsupervised algorithm implemented.

An Alternating Least Squares (ALS) recommendation algorithm is also available as well as some data pre-processing methods and support utilities such as cross validation.

Similarly to Spark, the library provides the feature of building processing pipelines, in a similar fashion than the well-known scikit-learn[52] library developed for Python programming language.

Additionally, an adapter is available for the Apache SAMOA[53] library (which will be discussed later), which offers learning algorithms for stream processing.

### Storm and associated ML tools

Apache Storm[54] is a framework initially developed by BackType, a company specialized in the analysis of social media streams, and continued at Twitter after BackType acquisition. In 2014, Twitter released the project under the open source Apache license. The motivations behind the creation of Storm can be identified with the intention to overcome limitation of existing distributed systems in processing social data streams of data. It is important to highlight that the authors of Storm are the same people that invented the Lambda architecture term, therefore Storm also is able to adapt to this architecture.

As represented by **¡Error! No se encuentra el origen de la referencia.**, Storm architecture consists of two main components: *spouts* and *bolts.*

A spout represents the input stream under analysis (e.g. a connector to the Twitter streaming API), while bolts represent the programming logic. A bolt is used to transform the data from either a spout or other bolts, but it can also be used as data sink when the processing pipeline reaches a final node in the processing pipeline.

A stream-processing task is Storm is defined as a DAG with spouts and bolts as nodes; this DAG is called a *topology* in Storm terminology.

Storm has been designed with the intention to process real-time streaming, but it also offers micro-batch processing capabilities via its Trident API[55].

[51] https://ci.apache.org/projects/flink/flink-docs-release-1.2/dev/libs/ml/index.htm
[52] http://scikit-learn.org/stable/
[53] https://SAMOA.incubator.apache.org/
[54] http://storm.apache.org/
[55] http://storm.apache.org/releases/2.0.0-SNAPSHOT/Trident-tutorial.html

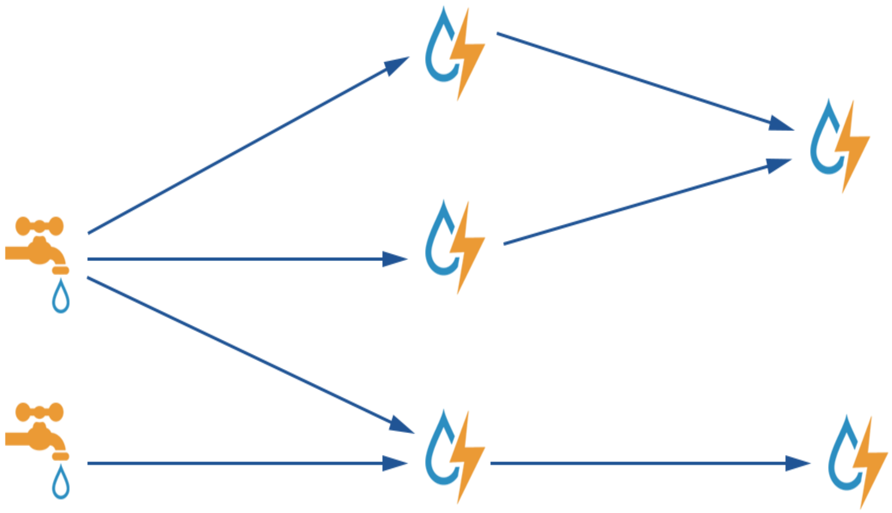


FIGURE 10 - STORM STREAM PROCESSING ARCHITECTURE

Storm has been designed as a system independent from Hadoop ecosystem or other distributed processing frameworks, however, since Hadoop moved to its new resource negotiator YARN[56], some effort has been dedicated in integrating the two projects.

Regarding ML capabilities, Storm does not comes with a machine learning library, but Apache SAMOA currently has implementations for classification and clustering algorithms running on Storm.

### Apache SAMOA

Apache SAMOA[53] is an incubator project from Apache with the intention to provide a platform for machine learning from streaming data.

It was originally developed at Yahoo! Labs in Barcelona in 2013 and moved to become an Apache incubator project since 2014. SAMOA is a flexible framework that can be run locally or few stream-processing engines, including Storm and Flink.

At the moment of writing, SAMOA offers only a small set of implemented machine-learning algorithms, which are also represented as DAG, internally called topologies (as in Storm). So far, there are only a few learning algorithms implemented, but they cover the most common tasks. They can be applied for classification (Vertical Hoeffding Tree, which utilizes vertical parallelism on top of the Very Fast Decision Tree, or Hoeffding Tree, the standard decision tree algorithm for streaming classification tasks), clustering (CluStream), regression (implemented as Adaptive Model Rules Regressor for both vertical and horizontal parallelism), and frequent pattern mining (Distributed Stream Frequent Itemset Mining), along with boosting, and bagging for ensemble creation of classifiers.

Additional learning algorithms can be integrated from Moa[59] with a plugin called SAMOA-Moa[57], which allows integrating Moa classifiers and clustering algorithms inside the SAMOA framework.

However, it is important to highlight that the underlying implementation of Moa's algorithms is not distributed; therefore it could not be considered as true distributed stream processing.

In case of customisation of existing algorithms, or implementation of new algorithms, SAMOA provides a common platform and a framework for the user to write their own distributed streaming

[56] http://hadoop.apache.org/docs/stable2/hadoop-yarn/hadoop-yarn-site/YARN.html
[57] https://github.com/SAMOA-moa/SAMOA-moa

algorithms. At the moment, SAMOA does not have a community of users and developers comparable to other ML libraries such as MLlib or Mahout, but it offers an extensive documentation.

The platform is targeted to process big data streams that are constantly being updated. One of the benefits of SAMOA is that it easily allows the user to implement new algorithms that will automatically run on any processing engine that is able to plug into SAMOA. Evaluation of the platform is not extensive: in a relatively recent survey[58] the authors provide an analysis in comparing SAMOA with Moa[59] in topic modelling pointing out significant higher throughput on SAMOA and highlighted the fact that the framework is robust and stable enough for production environments.

Concluding, SAMOA is a very young project and new tools are continually being developed to expand the library making it a potential protagonist in the stream processing community. However, it offers already a solid base for development of user defined algorithms making it an interesting solution.

### Artificial Intelligence

Artificial Intelligence (AI) is “the branch of computer science concerned with making computers behave like humans”. The term was first used by John McCarthy[60] in 1950.

There are two types of AI:

- Strong AI – Is an artificial intelligence based on a computer that can "think" and is "self-conscious"
- Weak AI - Is an artificial intelligence that does not pretend to think, but can solve a certain class of problems in a more or less "smart" way, for example by applying a set of rules.

Strong AI development does not record a major progress during the time, but there is progress in the field of weak artificial intelligence, such as verbal and written recognition, automatic translation from one language to another, chess gameplay.

One principles of artificial intelligence is that human reasoning is a form of calculation that can be identified, formalized, and consequently automated.

An attempt to simulate human intelligent behavior has led to the investigation of some types of problems that involve a high degree of human expertise, such as medical diagnosis, management, designing and solving general engineering problems. These problems require shaping of large amounts of information that are difficult to structure like experience and intuition.

Starting from the importance of knowledge in intelligent systems, in the late 1970s emerged a branch of artificial intelligence called knowledge engineering. Knowledge Engineering deals with the study of methods and techniques for acquiring, representing and organizing knowledge in knowledge-based systems

Another branch of AI is machine learning. The objective of machine learning is getting computers to act without being explicitly programmed, objective that can be achieved by using algorithms that identify models on received data and make decisions or predictions based on these models.

Current machine learning methods are increasingly sophisticated, allowing them to be integrated into complex medical applications such as genome analysis or depression diagnosis.

[58] Romsaiyud W. Automatic Extraction of Topics on Big Data Streams through Scalable Advanced Analysis. In: 2014 International Computer Science and Engineering Conference (ICSEC); 2014. pp. 255–60

[59] http://moa.cms.waikato.ac.nz/

[60] http://airesearch.com/tag/john-mccarthy/

Deep learning is a new area of machine learning and is, at present, the most advanced field of artificial intelligence. The main objective of deep learning is to give computers the ability to think and learn as close as possible like humans. Deep learning uses a model of computing that's very much inspired by the structure of the human brain.

The following figure, published by NVIDIA[61] in 2016, presents the relationship between AI, machine learning and deep learning.

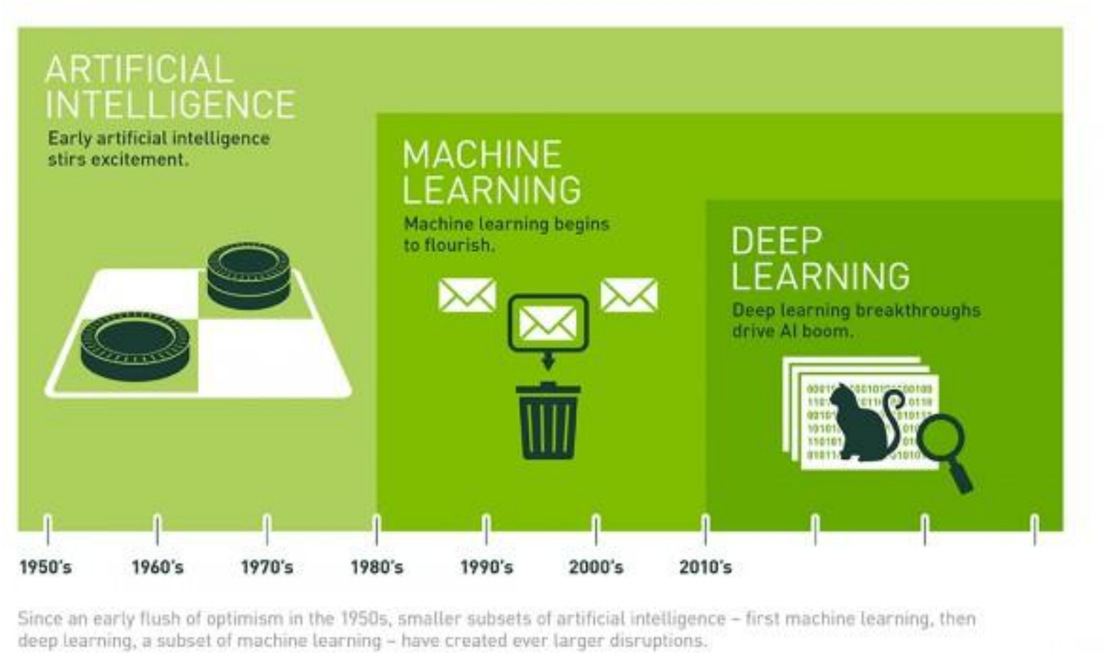

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

FIGURE 11 - RELATION AMONG AI, MACHINE LEARNING AND DEEP LEARNING

Artificial intelligence techniques have the role to organize and to use knowledge in an effective way so that:

- can be perceived by the people who provided that knowledge
- can be easily modifiable to correct errors
- can be used in many situations, although they are incomplete or inaccurate

Artificial intelligence techniques can be used in the following fields:

- sorting information
- voice and image recognition
- image, writing and voice analysis
- automatic translation
- smart games (chess, bridge, go)
- medical diagnosis
- task planning
- robots handling

**Searching** represents the universal technique of problem solving in artificial intelligence. The solution to many problems can be described by defining sequences of actions that lead to a result. Each of these actions changes a state; the purpose is to identify the sequence of actions and states

[61] https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

that lead from the initial state to the final state and that means solving the problem (or reaching the proposed goal).

- Search techniques can be grouped into the following 3 categories:
- Finding a way from the initial state to the final state
- Finding the right way
- Finding an acceptable way against an opponent (the opponent has his own goal)

There are several search techniques; each materializes through an algorithm that will solve the problem. You have to choose the one that is best suited to the typology of the problem to be solved.

Some of the search techniques are listed below:

- Breadth-First Search
- Depth-First Search
- Bidirectional Search
- Uniform Cost Search
- Iterative Deepening Depth-First Search
- A * Search
- Hill-Climbing Search
- Local Beam Search

In September 2016 the following top 15 open source tools for AI were published:

1. Caffe[62] - Is a deep learning framework based on expressive architecture and extensible code
2. CNTK[63] (Computational Network Toolkit) - Is one of Microsoft's open source artificial intelligence tools. It is used speech recognition, machine translation, image recognition, image captioning, text processing, language understanding and language modeling.
3. Deeplearning4j[64] - is an open source deep learning library for the Java Virtual Machine (JVM). It makes it possible to configure deep neural networks.
4. Distributed Machine Learning Toolkit[65] (DMTK) - is one of Microsoft's open source artificial intelligence tolos, designed for use in big data applications
5. $H_2O$[66] - It is focused on enterprise uses of AI, it can be used for predictive modeling, risk and fraud analysis, insurance analytics, advertising technology, healthcare and customer intelligence
6. Mahout[67] - is an open source machine learning framework..
7. MLlib[68] - is Apache Spark's scalable machine learning library
8. NuPIC[69] - Is an open source artificial intelligence project based on a theory called Hierarchical Temporal Memory (HTM).
9. OpenNN[70] - is a C++ programming library for implementing neural networks
10. OpenCyc[71] - It provides access to the Cyc knowledge base and commonsense reasoning engine. It is useful for rich domain modeling, semantic data integration, text understanding, domain-specific expert systems and game AIs.

[62] http://caffe.berkeleyvision.org/
[63] https://www.microsoft.com/en-us/cognitive-toolkit/
[64] https://deeplearning4j.org/
[65] https://www.dmtk.io/
[66] https://www.h2o.ai/
[67] http://mahout.apache.org/
[68] http://spark.apache.org/mllib/
[69] https://leanpub.com/realsmartmachines/read
[70] http://www.opennn.net/

11. Oryx 2[72] - is a specialized application development framework for large-scale machine learning. It also includes some pre-built applications for common big data tasks like collaborative filtering, classification, regression and clustering.
12. PredictionIO[73] - It helps users create predictive engines with machine learning capabilities that can be used to deploy Web services that respond to dynamic queries in real time.
13. SystemML[74] - It offers a highly-scalable platform that can implement high-level math and algorithms written in R or a Python-like syntax.
14. TensorFlow[75] - It is one of Google's open source artificial intelligence tools. It offers a library for numerical computation using data flow graphs.
15. Torch[76] - is "a scientific computing framework with wide support for machine learning algorithms that puts GPUs first."

## Using social media

Social media is an important source of Digital interaction data. Social media channels have evolved and are evolving and change the way people communicate with each other and also how they interact with companies. People may share any kind of content, text, photos, as well as audio and video. Some services have gained a huge user base, Facebook in particular, but the field is continuously evolving with new services are continuously being introduced Snapchat and Periscope are good examples of new services with quickly growing user base and special capabilities, such as short-lived content and real-time streaming. Companies wanting to reach people through different channels must all the time be able to monitor the field and adapt and innovate their communication in these new channels.

Different examples of current social media tools: Crystal, Live Stream, Welodias, One Riot, Collecta, LinkedIn, Plaxo with simply hired, Twitter with URL Blog or LinkedIn, Jobster, Facebook, Craiglist, MyWorkster, VisualCV, JobFox, Ecademy, Breezy HR, Job Adder, The Ladders, Hiring Solved, Connectifier, Data.com, SocialMention, BuzzSumo, SearchWiki (SocialSearch.com), Google Analytics, Yoast SEO, Smart Recruiters, Stack Overflow, GitHub, TalentStream Recruit, Handshake, 48ers.

## Opinion mining

Opinion mining, also known as “Sentiment analysis,” crawls the information from numerous text forms such as reviews, news, and blogs, advancing a polarity classification, listing them as positive, negative or neutral.

A lot of the data in social media is text giving vast possibilities for opinion mining and otherwise detecting trends and developments. Opinion mining involves analyzing people's opinions, sentiments, and attitudes expressed in written language. Studying the opinion mining tools, we distinguished three categories:

- Document level – queries that determine whether a whole opinion document denotes a positive or negative sentiment. Think of a product review, with these technologies we can conclude whether the post covers a positive or negative review.

[71] http://www.opencyc.org/
[72] http://oryx.io/
[73] https://predictionio.incubator.apache.org/
[74] https://systemml.apache.org/
[75] https://www.tensorflow.org/
[76] http://torch.ch/

- Sentence level - at this level we examine factual (relates to exact information) and subjective (personal views) sentences, and achieve a deeper understanding of the sentiment analysis process.
- Entity and Aspect level: known as feature level.

The most widespread use of sentiment analysis is in the area of reviews of consumer products and services. There are numerous data sources available for Opinion mining, namely:

- Blogs: http://indianbloggers.org/, http://www.bloggersideas.com/, http://www.digitaltrends.com/, http://thoughts.com/free-blog, http://blog.com/, http://blog.hubspot.com/, https://wordpress.com/;
- Microblogs:https://tumblr.com/ (Tumblr), http://friendfeed.com/ (Friendfeed), http://www.plurk.com/top/ (Plurk), https://twitter.com/ (Twitter), http://www.jaiku.com / (Jaiku), http://www.qaiku.com/ (Quiku), https://www.identi.ca/ (Identica), http://www.spotjots.com/ (S potjots), http://www.meetme.com/ (Meet me);
- Online posts: https://www.facebook.com/ (Facebook), https://myspace.com/ (MySpace), http://www.skype.com (Skype), https://www.linkedin.com/ (Linkedin), https://diasporafoundation.org/ (Diaspora), https://plus.google.com/ (GooglePlus), https://www.whatsapp.com/ (Whatsapp), https://www.snapchat.com;
- Forums: http://www.forums.mysql.com, http://www.forums.cnet.com, http://www.forum.joomla.org, https://forums.digitalpoint.com, http://www.bookforum.com, http://www.myspace.com/forums, http://tsrmatters.com/ (The Student Room), http://ubuntuforums.org, https://stackoverflow.com/;
- Review sites and news feeds.

As mentioned, there are various data sources available on web and mining the vast amount of data is rather challenging. The primary challenge is the extraction of emotions, structure of the text, a form of data, either image or text, the language used on the internet for communication varies from person to person or state to state.
We present below some ready to use tools for opinion mining for various purposes like data preprocessing, classification of text, clustering, opinion mining, sentiment analysis, etc.

1. STANFORD CORENLP[77] POS tagging, Named entity recognizer, Parsing, Sentiment analysis, Bootstrapped pattern learning;
2. WEKA[78] - Machine learning algorithm for Data Mining, Data pre-processing, Classification, Regression, Clustering, Association rules;
3. NLTK[79] - Classification, Tokenization, Stemming, Tagging, Parsing, Semantic reasoning, Provides lexical resources such as WordNet;
4. APACHE OPENNLP[80] - Tokenization, Sentence segmentation, Part-of speech tagging, Named entity extraction, Chunking, Parsing, Coreference resolution;
5. GATE[81] - Tokenizer, Gazetteer, Sentence splitter, POS tagging, Named entities transducer, Coreference tagger;
6. Pattern[82] - Data mining, POS tagging, N-gram search, Sentiment Analysis, WordNet, Machine learning, Network Analysis, Visualization;
7. Robust Accurate Statistical Parsing[83] - Statistical Parser, Tokenization, Tagging, Lemmatization and Parsing.

[77] http://nlp.stanford.edu/software/corenlp.html
[78] http://www.cs.waikato.ac.nz/ml/weka/
[79] http://www.nltk.org/
[80] https://opennlp.apache.org/
[81] https://gate.ac.uk/
[82] http://www.clips.ua.ac.be/pattern
[83] https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/parser/charniak/CharniakParser.html

Sentiment analysis proved as a very useful tool for campain managers, offering great insight for candidates running for various positions. Campaign manager can feel the voters view on certain issues and how to adjust the candidate speech and direct their future actions. An analysis of tweets over the 2010 campaign is available at http://www.nytimes.com/interactive/us/politics/2010-twitter-candidates.html.

Also, sentiment analysis technologies have been well integrated into the financial market domain, grasping the copious news items, articles, blogs, and other online posts about each public company. A sentiment analysis system can employ these multiple sources to find articles that discuss the companies and aggregate the view about them as a single score that can be used by an automated trading system. One such system is The Stock Sonar[84], developed by Digital Trowel, shows the daily positive and negative sentiment graphically, about each stock alongside the graph of the price of the stock.

Of course, when engaging in online marketing, the use of several monitoring tools (Google Analytics, Radian6, etc.) is the best choice to listen to the audience and identify their views and adapt your product and services to their desires.

We present below a shortlist with several practical tools that can track user sentiment:

- Meltwater: Evaluates the tone of the review as a proxy for brand reputation and uncovers new insights that help to identify the target audience.
- Google Alerts: A simple and very useful tool to monitor the search queries. It's used to track "content marketing" and get regular email updates on the latest relevant Google results. Also, it's considered a good starting point for tracking influencers, trends, and competitors.
- Tweetstats: allows to graph Twitter stats, having a very intuitive interface.
- Facebook Insights: With this tool we can view the total page Likes, the number of fans, daily active users, new Likes/Unlikes, Like sources, demographics, page views and unique page views, tab views, external referrers, and media consumption.
- Pagelever: This is another tool for measuring Facebook activity. Pagelever offers the ability to precisely measure each stage of how content is consumed and shared on the Facebook platform.
- Social Mention: The social media equivalent to Google Alerts, this is a useful tool that allows to track mentions for identified keywords in video, blogs, microblogs, events, bookmarks, comments, news, Q&A, hash tags and even audio media. It also indicates if mentions are positive, negative, or neutral.
- Marketing Grader: Hubspot's Marketing Grader is a tool for grading an entire marketing funnel.
- People Browser: Offers reports with all mentions of a brand, industry, and competitors and analyze sentiment. This tool allows analyzing the volume of mentions before, during and after the marketing campaigns.
- Google Analytics: Is a great tool for determining which channels influenced the subscribers and buyers. It allows to generate custom reports, annotations to keep uninterrupted records of the marketing and web design actions.

### Sentiment analysis

In recent years, with the growth of the World Wide Web and the advent of Internet, people actively express their opinions about products, services, events, political parties, etc., in social media, blogs, and website comments so there is a big amount of texts available online. In particular, on the one hand the opinions about different topics influence each other's decisions by communicating their sentiments towards a brand which reverberates in the companies. On the

[84] http://www.thestocksonar.com

other hand, opinions are crucial for companies to know about the quality of their products and services.

The majority of current sentiment analysis systems address a single language, usually English; however, with the growth of the Internet around the world, users write comments in different languages.

One of the main problems in multilingual sentiment analysis is a significant lack of resources. Thus, sentiment analysis in multiple languages is often addressed by transferring knowledge from resource-rich to resource-poor languages, because there are no resources available in other languages. The majority of multilingual sentiment analysis systems employ English lexical resources such as SentiWordNet.

Another approach is to use a machine translation system to translate texts in other languages into English: the text is translated from the original language into English, and then English-language resources such as SentiWordNet are employed. Translation systems, however, have various problems, such as sparseness and noise in the data. Sometimes the translation system does not translate essential parts of a text, which can cause serious problems, possibly reducing well-formed sentences to fragments.

The main techniques considered for the sentiment analysis systems can be classified into corpus-based approaches using machine learning, lexicon-based approaches, and hybrid approaches. Corpus-based methods use labelled data; lexicon-based methods rely on lexicons and optionally on unlabelled data; and hybrid methods are used based on both labelled data and lexicons, optionally with unlabelled data.

Sentiment lexicons have been used in a number of approaches to multilingual sentiment analysis in order to improve the performance of classification. Sentiment lexicons are used mainly in lexicon-based sentiment analysis.

- SenticNet is a lexical resource based on a new multi-disciplinary approach proposed by Cambria et al.[85] to identify, interpret, and process sentiment in the Internet. SenticNet is used for concept-level sentiment analysis. It is also used to evaluate texts basing on common-sense reasoning tools that require large inputs. However, it is not capable of analysing text with sufficient level of granularity. Sentic computing methodology is used, in particular, to evaluate texts at the page or sentence level. The purpose of SenticNet is to build a collection of concepts, including common-sense concepts, supplied with polarity labels, positive or negative. Unlike SentiWordNet, SenticNet does not assume that a concept can have neutral polarity. SenticNet includes a simple and clear API for its integration in software projects. It can be used with the Open Mind software. It guarantees high accuracy in polarity detection. Multilingual tools are available for SenticNet[86].
- SentiWordNet is a lexical resource that assigns WordNet synsets to three categories: positive, negative, and neutral, using numerical scores ranging from 0.0 to 1.0 to indicate a degree to which the terms included in the synset belong to the corresponding category. SentiWordNet was built using quantitative analysis of glosses for synsets [87]. While SentiWordNet is an important resource for sentiment analysis, it contains much noise. In addition, it assigns polarity at the syntactic level, but it does not contain polarity information

[85] Cambria, E, Olsher D, Rajagopal, D. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In: AAAI, 2014, p. 1515-1521, Quebec City

[86] Xia Y, Li X, Cambria E, Hussain A. A localization toolkit for SenticNet. In: 2014 IEEE international conference on data mining workshop (ICDMW). 2014, p. 403–8

[87] Singh VK, Piryani R, Uddin A, Waila P, et al. Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches. In: 2013 5th international conference on knowledge and smart technology (KST). IEEE; 2013, p. 122–27

for phrases such as “getting angry” or “celebrate a party”, which correspond to concepts found in the text to express positive or negative opinions[88].

- SEL is a Spanish emotion lexicon that presents 2036 words supplied with the Probability Factor of Affective use (PFA) as the measure of their expression of basic emotions: joy, anger, fear, sadness, surprise, and disgust, on the scale of null, low, medium, or high. The lexicon was developed manually by 19 annotators, which had to agree above certain threshold for a label on the word to be included in the lexicon. The measure called Probability Factor of Affective use (PFA) was developed by the authors of this lexicon to incorporate agreement between annotators in decision-making on labelling the words: the greater the agreement, the stronger the expression of the emotion by the given word. The lexicon, freely available for download, has been used in opinion mining tasks on Spanish tweets[89].
- R Sentiment Analysis. Machine learning makes sentiment analysis more convenient. In the landscape of R, the sentiment R package and the more general text mining package have been well developed by Timothy P. Jurka90. Text analysis in R has been well recognized (see the R views on natural language processing). Part of the success belongs to the tm package: A framework for text mining applications within R. It did a good job for text cleaning (stemming, delete the stopwords, etc) and transforming texts to document-term matrix (dtm). There is one paper about it. The most important part of text analysis is to get the feature vectors for each document. The word feature is the most important one. Of course, you can also extend the unigram word features to bigram and trigram, and so on to n-grams.

### Dashboards for social media

Nowadays, Twitter has become an interesting scenario for social analysis. We can find in this micro-blogging site millions of interactions between users. The main and most popular feature is its simplicity and synthesizing, that is because in most micro-blogging systems, user’s messages are delimited by the number of characters. In those characters you can do many things such as talking about what you are doing, interact with other users, etc.

In this project we are focusing in one thing you can also do with these micro-blogging systems, give your own opinion. This means, Twitter is a source of many varied opinions about a topic.

The topic we are managing is football, because is known that sporting events evoke strong emotions amongst fans. The idea is to relate these sentiments generated in Twitter with these football events in order to analyze correlations between them. In particular, this final graduate work will be focused on the application of visualisation techniques for providing insight about these relationships.

To do so, we will use visualisation techniques based on interactive widgets to provide easy access to data.

### Large Scale Data collection

Large-scale user data collection from Internet and recently from social networking services mainly depends on the functionality provided by the analyzed system. Possible solutions includes methods

[88] Cambria, E, Olsher D, Rajagopal, D. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In: AAAI, 2014, p. 1515-1521, Quebec City

[89] Sidorov G, Miranda-Jime´nez S, Viveros-Jime´nez F, Gelbukh A, Castro-Sa´nchez N, Vela´squez F, Dı´az-Rangel I, Sua´rezGuerra S, Trevin˜o A, Gordon J. Empirical study of opinion mining in Spanish tweets. MICAI 2012. Lect Notes Comput Sci. 2012;7629:1–14

[90] https://www.r-bloggers.com/sentiment-analysis-with-machine-learning-in-r/

that use of a system's APIs, as in Google+ and Twitter[91] [92], and the integration of their own applications to attract people and users to provide access to their profile information, as in Facebook[93]. In addition, advanced web crawling methods have been used to analyze Facebook[94] [95] and other platforms, e.g., Myspace, Flickr, and YouTube[96]. Usually, web crawling is applied in cases where the required data cannot be accessed via an API, or the target sources provide more data from the web portal than the API. The research community uses the collected data to analyze social media traffic and public information from user profiles[97] [98].

The content that is generated by users in Social Network platforms is of high interest for companies that wants to monitor their online reputation and reception. However, the collection of a large-scale dataset is not a trivial task due to several challenges that are introduced mostly by the resources' limitations[99][100].

A large-scale data collection process can be either Resource Specific or Real-Time. The first case provides retrieval services for a specific resource (e.g a user's profile, a tweet or post), while the latter enables the sample collection of real-time information that is being published in the Social Network.

An example of a resource specific large-scale data indexing platform is Exalead Cloudview[101] : an innovative search engine that comes with the important function of indexing and correlating information from multiple data sources unstructured, semi-structured and structured display of comprehensive results. The solution has advanced search features and different types of sequencing results. Cloudview main features and advantages are:

- Allow processing, indexing and accessing records of any kind (databases, documents, images, videos and audio, etc.);
- Can analyze large volumes of structured data (XML records, GPS tracking units, MySQL tables) and unstructured (documents, emails, audio and video);
- Can access multiple sources through a wide range of connection points.

[91] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas, "Google+ or google-?: dissecting the evolution of the new OSN in its first year," in WWW, 2013.

[92] Ratan Dey, Yuan Ding, and Keith W. Ross. 2013. "Profiling high-school students with facebook: how online privacy laws can actually increase minors' risk". In Proceedings of the 2013 conference on Internet measurement conference (IMC '13).

[93] A. Nazir, S. Raza, and C.-N. Chuah, "Unveiling Facebook: A Measurement Study of Social Network based Application," in Proc. of the ACM SIGCOMM Conference on Internet Measurement Conference, 2008, p. 43–56.

[94] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs," in Proc. IEEE INFOCOM, 2010, p. 1–9.

[95] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical Recommendations on Crawling Online Social Networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 9, p. 1872–1892, 2011.

[96] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "I tube, you tube, Everybody tubes: Analyzing the World's Largest Uer Generated Content Video System," in Proc. of the ACM SIGCOMM Conference on Internet Measurement, 2007, p. 1–14.

[97] Reza Farahbakhsh, Xiao Han, Angel Cuevas and Noel Crespi, "Analysis of publicly disclosed information in Facebook profiles", IEEE/ACM ASONAM, Niagara fall, Canada, 2013.

[98] Gabriel Magno, Giovanni Comarela, Diego Saez-Trumper, Meeyoung Cha, and Virgilio Almeida. 2012. "New kid on the block: exploring the google+ social graph". In Proceedings of the 2012 ACM conference on Internet measurement conference (IMC '12)

[99] A. Black, C. Mascaro, M. Gallagher, and S. P. Goggins. Twitter zombie: Architecture for capturing, socially transforming and analyzing the twittersphere. In Proceedings of the 17th ACM International Conference on Supporting Group Work, GROUP '12, pages 229–238, New York, NY, USA, 2012. ACM

[100] K. Tao, C. Hauff, G. J. Houben, F. Abel, and
G. Wachsmuth. Facilitating twitter data analytics: Platform, language and functionality. In Big Data (Big Data), 2014 IEEE International Conference on, pages 421–430, Oct 2014.

[101] www.exalead.com

In this line of investigation, several research works can be used as reference material: Regarding the collection of data, we mention a work[102] that studies the design of an effective web crawler. In the paper, the authors present several problems in crawling procedure created by the rapid increment on the size of the data available in the Web. The authors propose multiple architectures for parallel distributed crawling framework and identify the challenges in large-scale data crawling, a similar approach is described also in another paper[103].

Several challenges can also be identified in the design and implementation of an effective large-scale dataset collection framework, as the increasing quantity of information that is published in Social Network platforms introduce relevant problems. The majority of these platforms maintain monitoring services to control the data throughput, introducing several additional challenges to the parallel data collection campaigns.

A major problem in a data collection campaign is the one of requests rate limiting policies of Social Network providers. In the recent years major Social Network platforms have used IP-based policies, which restrict a single machine to perform a certain number of requests[104]. The solution on addressing this challenge was straightforward: a distributed data collection procedure was able to effectively overcome this limitation. A research work[105] present the different categories of challenges for building a crowd crawling system, highlighting the resource diversity of the different parts, the different rate limiting policies from Social Network providers, and the data fidelity. They propose a framework of crowd crawling; where a team of multiple groups shares resources in order to efficiently collaborate in a data crawling procedure. Their prototype is implemented over Planetlab[106], from which they take advantage of the availability of different IP addresses used for the crawling process.

Coalmine[107] is a social network data-mining system, which implements its own mechanism for collecting the data from Twitter and is able to retrieve data using the official API. Its overall architecture is based on distributed principles, where multiple IP addresses are used. In another work[108], the authors propose another similar framework for large-scale dataset collection from Facebook. They design and implement a distributed tool, which is able to overcome IP-based limitations and collect a large sample.

However, recently the most famous Social Network platforms, such as Twitter, have changed the traditional IP-based limitation policy to Application-based limitation; effectively restricting a single application from performing a large number of requests. This recent change of policies makes large-scale dataset collection procedure more complex, as the distributed design in the proposed fashion is not functional. In a recent work[109], the authors propose a framework, which is able to overcome the newly introduced challenges in the large-scale data collection field and perform a large-scale data collection campaign by using multiple Social Networks accounts. The proposed solution enables the collection of more than 2M complete user information in one day, while previous works require a much more resource-demanding configuration to achieve similar performance.

[102] J. Cho and H. Garcia-Molina. Parallel crawlers. In Proceedings of the 11th International Conference on World Wide Web, WWW '02, pages 124–135, New York, NY, USA, 2002. ACM.

[103] M. D. Dikaiakos and D. Zeinalipour-Yazti. A distributed middleware infrastructure for personalized services. Computer Communications, 27(15):1464 – 1480, 2004

[104] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM

[105] C. Ding, Y. Chen, and X. Fu. Crowd crawling: Towards collaborative data collection for large-scale online social networks. In Proceedings of the First ACM Conference on Online Social Networks, COSN '13, pages 183–188, New York, NY, USA, 2013. ACM

[106] http://www.planet-lab.org/

[107] J. S. White, J. N. Matthews, and J. L. Stacy. Coalmine: an experience in building a system for social media analytics, 2012

[108] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks.Selected Areas in Communications, IEEE Journal on, 29(9):1872–1892, October 2011

[109] H. Efstathiades, D. Antoniades, G. Pallis and M. D. Dikaiakos, "Distributed Large-Scale Data Collection in Online Social Networks," 2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC), Pittsburgh, PA, 2016, pp. 373-380.

# 4. SoMeDi technological innovations - Beyond the state of the art

## Introduction

The objective of this section is to introduce the progress and technological innovations that SoMeDi approach proposed in the different main domains identified in the previous sections. Our objective is to develop in SoMeDi innovations to cover the gap among the current state of the art and the beyond state of the art we want to achieve after the project finalization.

## Machine Learning

SoMeDi will advance machine learning applications on social media and DII tools, integrated within software modules capable of time-series analysis and prediction, iterations run over samples of hundreds of social network data streams.

The results of the analysis of these sets of data will consist in viable knowledge on the input streams (in batch or streaming mode) and gain insight over future values of the monitored data series to allow preventive actions / decisions.

SoMeDi will deploy functions that allow the analysis of multimedia information for near-real time pattern analysis.

The state of the art technologies involved in developing these features will define new methods for user profiling, based on conventional clustering algorithms.

Clustering is an unsupervised learning problem whereby we aim to group subsets of entities with one another based on some notion of similarity. Clustering is often used for exploratory analysis and/or as a component of a hierarchical supervised learning pipeline (in which distinct classifiers or regression models are trained for each cluster).

Spark: As an in-memory cluster computing framework for iterative and interactive applications.

Spark, one of the most active Apache open-source projects, is a parallel dataflow system implemented in Scala and centered around the concept of Resilient Distributed Datasets (RDDs).

The spark.mllib package supports the following models:

- K-means
- Gaussian mixture
- Power iteration clustering (PIC)
- Latent Dirichlet allocation (LDA)
- Bisecting k-means
- Streaming k-means

## Artificial Intelligence

Artificial intelligence (AI) is the ability of computers to understand certain aspects of the natural world, and ultimately, use that understanding to complete tasks normally requiring human intellect and effort.

In this white paper [110], DemandGen Report presents findings from a recent poll of B2B marketers showing that those companies who have successfully deployed lead nurturing programs showed an "average 20% increase in sales opportunities".

With AI's promise to change the game for marketers seeking better behavioral targeting methods, more software with built-in AI capabilities will begin to appear.

For example, a marketing stack that employs AI algorithms might learn that a certain buyer who habitually uses Twitter on Thursday and Friday mornings has recently spearheaded and executed a major project and constantly video-conferences with colleagues across locales. The software can then suggest (or even create) targeted Tweets to be published on the days and times that she'll see them: one that congratulates her on her successful project and one linking to a blog post about how your product brings multi-regional corporate employees closer together.

A complete view of individual customers is priceless and promises to make your social media marketing, content distribution and all other campaign elements more effective.

The use of AI in marketing is mostly about increasing personalization. Strategies for products / services personalization:

- Conform the content to buyer persona insights as closely as possible;
- Mine the CRM tools for rich customer insights. AI tools facilitate to conduct a deep review of customer details within the CRM platform. But if there is available a large-enough dataset of customer interactions, it can be deployed a semantic analysis to understand the level of buying intent behind the words your qualified prospects use;
- Generate viable content / posts that encourage engagement and maintain an active presence in social media discussion hubs.

### Opinion mining and Sentiment Analysis

By using natural language processing tools we can enter another era of social media. Opinion mining, or sentiment analysis technologies allow the evolution of traditional emotion recognition techniques, usually engaged through machine learning approaches.

Social Media Monitoring is nowadays a common process, "monitoring" conversations on social media channels is a must-have technology for important brands and companies.

An interesting approach would be to turn the attention towards social media platforms like LinkedIn, to grasp a deeper understanding of the trends and maybe even change the learning processes to better suit the work market.

Realizing that social media users are connected and influence each other and this influence and emotions generated suggest a very dynamic environment, in constant change.

The main task is to find and innovate current technologies to pass through the Input – Analysis – Output process, and here SoMeDi will enhance deep syntax structured resources that go beyond the n-gram and bag-of-words paradigm and better capture the complexity of natural language sentences, integrating the target of sentiments, or considering the holder of the opinion via a deeper syntactic and semantic representation or inference system.

[110] Calculating The Real ROI From Lead Nurturing, DemandGen Whitepaper

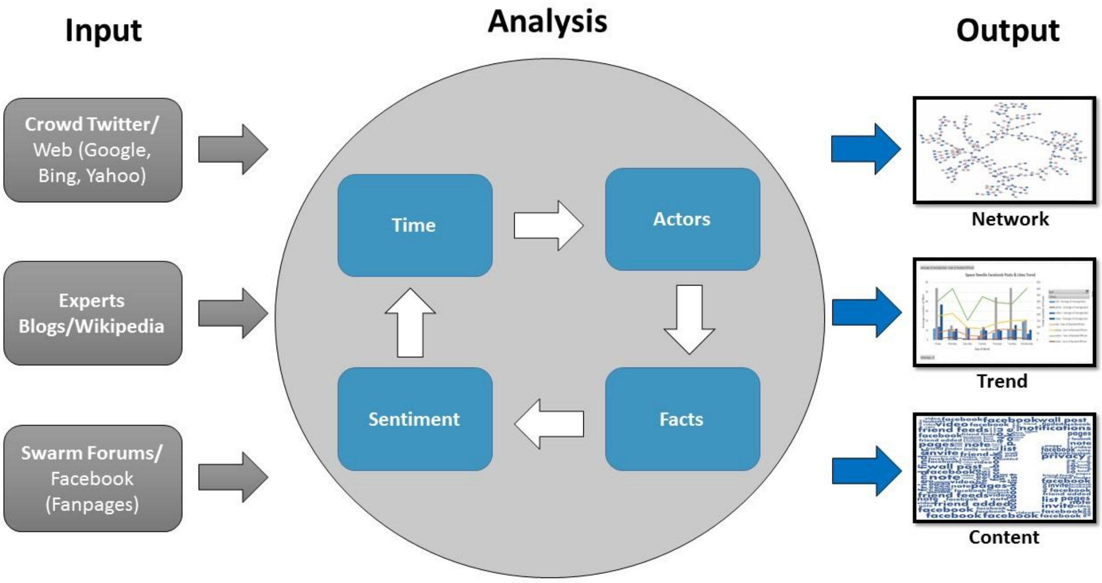


FIGURE 12 - SENTIMENT ANALYSIS PROCESS

Also, other key areas to develop regard the exploitation of advanced weakly supervised training approaches that go beyond clustering and corpus-based semi-supervised models to further exploit both large unstructured documents and contextual information as well as the time and evolution factor.

SoMeDi will go beyond the-state-of-the-art in this field through the implementation of sentiment analysis technologies based on classification models under the Deep Learning paradigm (one step forward of Machine Learning) that will allow faster and more efficient response. By using recursive neural network deep learning algorithms can better understand sentiment than traditional methods. Traditional methods isolate words into features and apply different feature selection and dimensionality reductions techniques to pick out the most 'important' word in the given input. Deep learning allows algorithms to understand sentence structure and semantics. The model is built as a representation of the entire sentence based on how the words are arranged and interact with each other.

With the proliferation of images and videos posted online (e.g., on YouTube, Facebook, Twitter) for product reviews, movie reviews, political views, and more, affective computing research has increasingly evolved from conventional unimodal analysis to more complex forms of multimodal analysis. Existing literature lack of frameworks with multimodal affect analysis, multimodality is defined by the presence of more than one modality or channel, e.g., visual, audio, text, gestures, and eye gage. In SoMeDi we will not only work with textual features, instead we will process also multimedia information, such as, images and videos. The fusion of different sources imposes a big challenge and will outperform the result of the sentiment analysis algorithms.

In addition to this, as social media platforms become the primary medium used by people to express their opinions and feelings about a multitude of topics that pop up daily on news media, the integration of static analysis of the emotion in social structure and complex event-aware machine learning approaches will be reinforce through SoMeDi. SoMeDi will treat them as a first result which may evolve with time and will be transmitted through influencers in the net. This has to be taken into account as the importance of influencer marketing is evident in the brand’s performance as it creates an emotional bond between consumers and brands.

Moreover, normally, a deeper analysis of syntax structured resources and natural language sentences which include sentiments and opinions, cannot be trivially enumerated or captured using predefined lexical patterns. SoMeDi will treat to optimize this process through a deeper syntactic and semantic representation beyond the n-gram and bag-of-words paradigm by extracting sentiment expressions for a given target from a corpus of unlabeled and unstructured documents (social media opinions).

In consequence, SoMeDi will deal with approaches that try to exploit large unstructured documents considering both contextual information and time and evolution factors by means of deep analysis of pieces of information including opinions and sentiments.

### Dashboards for social media

We are going to give an insight into techniques used in this project. First of all, we are going to explain Polymer web components, the technology in charge of the visualisation's structure. Secondly, the technology that has made possible interactive widget making. Finally, the technology we have used to store all the data.

- Polymer 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 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.
- D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using standard web technologies HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.
- Elasticsearch 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.

### Large Scale Data collection

In the area of large-scale distributed data collection from Social Networks and Medias  specific approaches for multi-site, multi-language and multi-modal crawling strategies for social media outlets will be developed.

The reuse of available existing language and semantics descriptions (e.g., localized versions of WordNet and SentiWordNet) will be leveraged to provide semi-automatically interoperable data models for language, sentiment and reputation that can be reused across domains and even language with ease.

A global ontology or semantic data model of the workflows to follow in the process will be generated to ensure that the process is kept interoperable across different applications and may be connected with third party modules and applications as well as extended to other locales such as different languages and social media. The crawling process will target text and also multimedia descriptors (e.g., metadata and analysis for images) and include open designs to enable deep analysis of media.

As we already mentioned, the crawling process will also need to be efficiently distributed to deal with the large quantities of available data, which will be done using Big Data strategies and technologies (e.g., Apache Spark for job distribution, Apache Solr or Elasticsearch for massive data indexing) and stored using high performance and scalable NoSQL engines such as MongoDB to store this heterogeneous data.

Finally the data crawling and indexing platform need to be flexible, configurable and adaptable to changes that may appear in the social Networks API, which can limit the information extraction capabilities of the SoMeDi crawling module. The crawler should be configurable in order to be able to use multiple social network accounts that will collaborate in a data extraction campaign, or other configurable strategies that may be required in case of appearing limitations in the social Network APIs.

# 5. Industry-wide needs and current pain points

The objective of this section is to identify the industry current needs that have guided the consortium to develop a project like SOMEDI. The needs and pain points will be categorized based on the use cases identified in the project (further described in SOMEDI Preliminary Use Cases) as the domains and the motivations could be totally different although the solutions and technologies used to solved them are the same.

## Marketing Industry needs

Nowadays, marketing industry tends to create useless content following the idea of getting something cheaply and easily. In 2008, Havas revealed on the Meaningful Brands survey that "60% of all content produced by brands is poor, irrelevant or fails to deliver", which means that most brands focus on quantity instead of quality. Actually, online marketing has become stronger over the time based overall on the usage of social media with the objective of getting more website traffic, Twitter followers and Facebook 'likes'. This kind of marketing allows brands to focus on direct-response marketing on a cheaper way, but this behavior has become a problem regarding to the expected impact. The report also found that most people would not care if 74% of brands disappeared, so creating quality marketing recommendations will be the main differentiating point of SoMeDi results.

Modern marketing industry is defined as customer-centric which means that the trend is to create personalized contents for each customer, by identifying patterns when he operates and purchases. The problem is that if advertisements and content are not relevant, the strategy morphs into personalized spamming, so the content fails to deliver. Contents by themselves are ineffective due to the absence of any marketing strategy and the fact that many actual marketers are inadequately skilled. By the integration of SoMeDi platform to the customer's business, accurate data will be provided to assist in the elaboration of strategies and marketing plans aiming to optimize and generate effective contents.

Publicity articles usually counts with limited number of words. Same in radio and television advertising that use to be limited by time, so every single word and second must be relevant. This limitation is less restrictive in online marketing since websites and media networks can store high amount of information. The existence of the limitation gives importance to the content, getting reduced in online marketing. In short, SoMeDi will measure brands social impact providing truthful information that will improve customer's content quality and marketing plan performance by creating reports focused on client's needs and likes.

The digital intelligent data tool outputs will support customer's marketing plan by: the definition of optimal target market that the customer should aim to, elaborating product's related recommendations and profiling users of interest; so simulations can be performed aiming to improve their marketing plan efficiency by the time. The above described problems happens mostly because of lack of efficient research due to the poor results achieved by conventional methods for collecting customer's opinion, like surveys taking place on annoying situations, the usual e-commerce mailing similar to spam or the intrusive online advertising including web announcements that prevent seeing the visited website. The social opinion database created by SoMeDi will provide an adequate source of information for reaching the desired impact.

## Recruitment/Education industry needs

According to a survey conducted by ANOSR (The Romanian National Alliance of Student Organizations) in 2011 on a sample of 20,000 students from 24 universities, analyzing the situation of student services in terms of student satisfaction regarding the CCC (Career Counseling Centers) efficiency, revealed the most unsatisfactory situation: in almost all universities, the percentage of students satisfied with their work was below 20%. This has led to distrust of these centers, and their effectiveness and thus to the non-use of the services they provide for students.[111]

These programs are quite detailed and complex and could provide a very useful information base. So, when facing the reality, studies as the one mentioned before show that there is a strong need for better solution to employ these programs.

Even if there is a European context that recommends them, and recently a legislative framework has been created, that regulates their existence and functionality in all Romanian universities, CCC guidance services are still at an early stage in terms of efficiency in relation to the objectives they were created. Sadly, there are no studies and data on the impact they have had on reducing university dropout, increasing access to higher education, facilitating the transition from secondary to higher education or from education to the labor market, but we understand the reality by taking into account the number of graduates that work in their field of education.

Unfortunately, this context is due to the fact that in general, the budgets allocated to these centers are quite small, and there are many situations where they need to look for self-financing sources to support their work. Some of them are trying to fill these gaps with the voluntary work invested by academics and university students working in the centers, but it is obvious that without enough staff with occupation in the field, the quality of the services offered and organized activities is rather poorly compared to the real needs of career counseling and guidance.

The effects of these issues that lead to a high risk of university dropout could be mitigated by functional CCC that deliver quality services, free of charge for students.

However, without a proper prioritization of their activity correlated with good strategies that actively engage universities, and without real investments by the state towards these centers budgets, these centers will remain entities that deal superficially with the problems for which they were set up and fail to produce real and impacting effects.

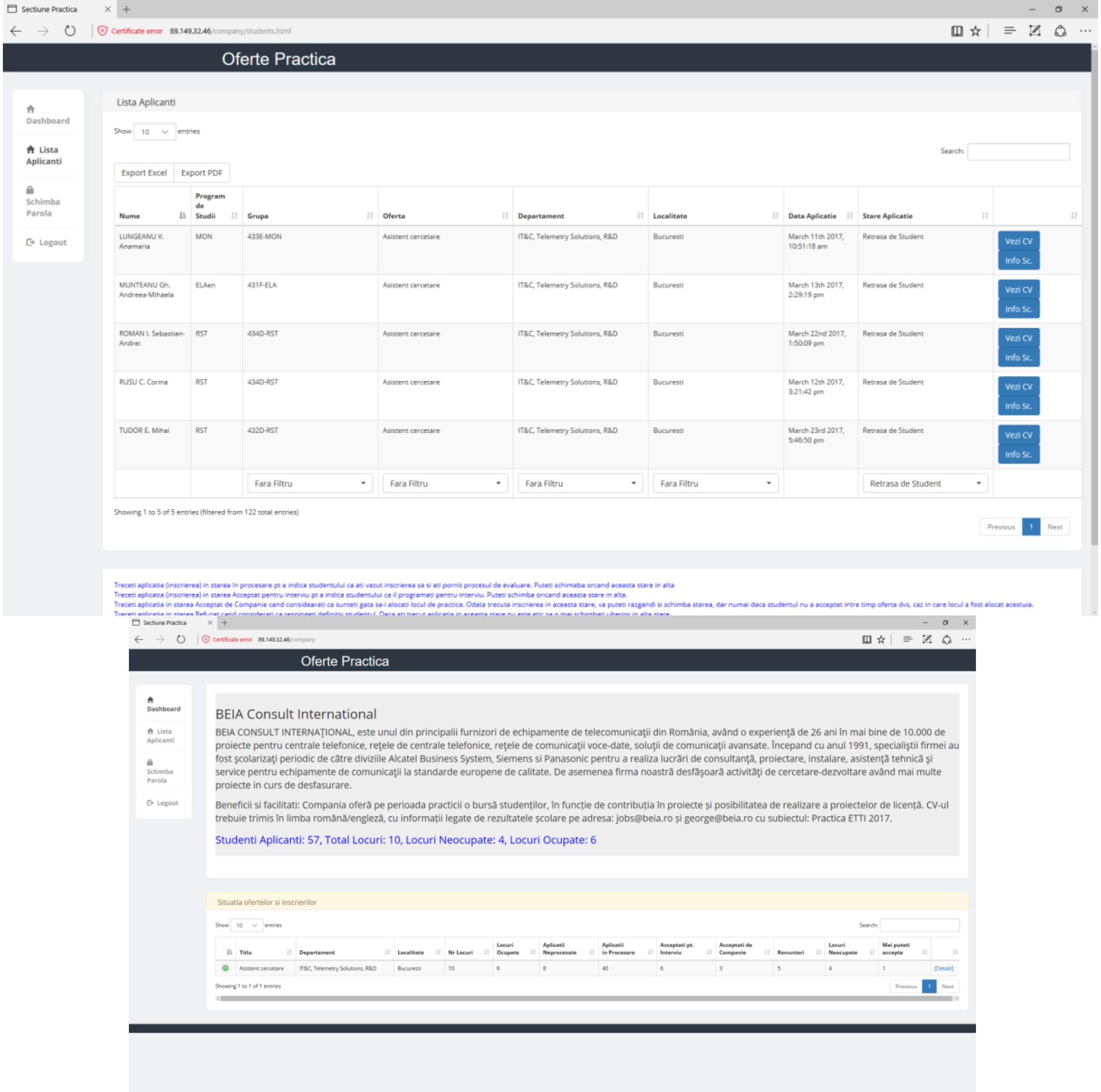
Industry needs for internship/apprenticeship programs in Romania and Europe have resulted in several initiatives, as below:

- Get in & GROW http://www.startinternship.ro/despre START Internship Romanian Program is an educational partnership initiated by a broad group of business associations, embassies, multinational companies and supported by the Government, universities and student associations. The main purpose of the program is to facilitate the preparation of young people as future employees.
- Go2intership https://go2internship.com/?gclid=CMXF28nP-NICFWgW0wodRvIBjg Go2internship.com is the platform where you can find a starting point on the way to the job you want. Go2internship aims to help students and young people by providing them a wide range of internships that meet various working areas.
- Hipo.ro http://www.hipo.ro/Interships Hipo gives users access to a rich collection of resources, content and online career tools including: jobs, career advice, career articles. Find a unique selection of jobs for professionals and internships for students from top employers.
- Internshipul.ro http://internshipul.ro/ Internship or Intership? This is the platform that allows students to apply to find different sort of opportunities regarding internships and also read blog articles that are related to this subject and participate to events.
- Intership.gov.ro http://internship.gov.ro/ Official Internship program of the Romanian Government will give young people the opportunity to familiarize with the working methods

[111] http://www.anosr.ro/wp-content/uploads/2014/04/Serviciile-de-consiliere-si-orientare-in-cariera-Perspectiva-studentilor.pdf)

of the central government. Beyond the daily work in the central administration, trainees will have the opportunity to participate in discussions and roundtables in the presence of senior officials of the Romanian state, as well as trainings and thematic workshops.

- Practica-ta.ro http://www.practica-ta.ro/ The creation of the first traineeships and internships platform facilitates dialogue honestly, directly and professionally between the three actors involved in job training programs: candidates, companies and universities. Students benefited from individual counseling in career, in terms of identifying strengths and career path best personality and working style of each one.
- Student in Romania http://www.studentinromania.ro/stagii-de-practica/ The initiative brings a new approach to the integration of undergraduate and graduate students in the labor market, including consideration of bringing to the forefront and addressing issues such as cultural integration, communication and relationship between people coming from different backgrounds.
- AIESEC http://www.aiesecbucharest.ro/ is the world's largest non-profit, independent and international student-run organization that brings together young people that are concerned about global internships, leadership development skills and volunteer exchange experiences with the same goal of making a positive impact on society.
- Programedeintership.ro http://programedeinternship.ro/ and inPractica.ro http://www.inpractica.ro/ are other platforms for internship and apprenticeship.
- Stagiipebune https://stagiipebune.ro/ project started by University Politehnica of Bucharest, Automation and Computer Science Faculty (UPB ACS) in 2005 for their students, has over 50 companies, 1000 jobs, over 16000 students participated from over 40 faculties.
- Some universities/faculties have their own matchmaking platforms for internship, for example UPB ETTI (Electronics, Telecommunication and Information Technology Faculty) http://www.electronica.pub.ro/index.php/studenti/practica/descrierea-aplicatiei-pentru-practica :



Lista Aplicanti

Show 10 entries

Search:

| Nume | Program de Studii | Grupa | Oferta | Departament | Localitate | Data Aplicatie | Stare Aplicatie | |
|---|---|---|---|---|---|---|---|---|
| LUNGEANU V. Anamaria | MON | 433E-MON | Asistent cercetare | IT&C, Telemetry Solutions, R&D | Bucuresti | March 11th 2017, 10:51:18 am | Retrasa de Student | Vezi CV<br>Info Sc. |
| MUNTEANU Gh. Andreea-Mihaela | ELAen | 431F-ELA | Asistent cercetare | IT&C, Telemetry Solutions, R&D | Bucuresti | March 13th 2017, 2:29:19 pm | Retrasa de Student | Vezi CV<br>Info Sc. |
| ROMAN I. Sebastian-Andrei | RST | 434D-RST | Asistent cercetare | IT&C, Telemetry Solutions, R&D | Bucuresti | March 22nd 2017, 1:50:09 pm | Retrasa de Student | Vezi CV<br>Info Sc. |
| RUSU C. Corina | RST | 434D-RST | Asistent cercetare | IT&C, Telemetry Solutions, R&D | Bucuresti | March 12th 2017, 3:21:42 pm | Retrasa de Student | Vezi CV<br>Info Sc. |
| TUDOR E. Mihai | RST | 432D-RST | Asistent cercetare | IT&C, Telemetry Solutions, R&D | Bucuresti | March 23rd 2017, 5:46:50 pm | Retrasa de Student | Vezi CV<br>Info Sc. |
| | | Fara Filtru | Fara Filtru | Fara Filtru | Fara Filtru | | Retrasa de Student | |

Showing 1 to 5 of 5 entries (filtered from 122 total entries)

Treceti aplicatia (inscrierea) in starea In procesare pt a indica studentului ca ati vazut inscrierea sa si ati pornit procesul de evaluare. Puteti schimaba orcand aceasta stare in alta
Treceti aplicatia (inscrierea) in starea Acceptat pentru interviu pt a indica studentului ca il programati pentru interviu. Puteti schimba oricand aceasta stare in alta.
Treceti aplicatia in starea Acceptat de Companie cand considearati ca sunteti gata sa-i alocati locul de practica. Odata trecuta inscrierea in aceasta stare, va puteti razgandi si schimba starea, dar numai daca studentul nu a acceptat intre timp oferta dvs, caz in care locul a fost alocat acestuia.
Treceti aplicatia in starea Refuzat cand considerati ca respingeti definitiv studentul. Daca ati trecut aplicatia in aceasta stare nu este etic sa o mai schimbati ulterior in alta stare.

BEIA Consult International

BEIA CONSULT INTERNAŢIONAL, este unul din principalii furnizori de echipamente de telecomunicaţii din România, având o experienţă de 26 ani în mai bine de 10.000 de proiecte pentru centrale telefonice, reţele de centrale telefonice, reţele de comunicaţii voce-date, soluţii de comunicaţii avansate. Începand cu anul 1991, specialiştii firmei au fost şcolarizaţi periodic de către diviziile Alcatel Business System, Siemens si Panasonic pentru a realiza lucrări de consultanţă, proiectare, instalare, asistenţă tehnică şi service pentru echipamente de comunicaţii la standarde europene de calitate. De asemenea firma noastră desfăşoară activităţi de cercetare-dezvoltare având mai multe proiecte in curs de desfasurare.

Beneficii si facilitati: Compania oferă pe perioada practicii o bursă studenților, în funcție de contribuția în proiecte și posibilitatea de realizare a proiectelor de licență. CV-ul trebuie trimis în limba română/engleză, cu informații legate de rezultatele școlare pe adresa: jobs@beia.ro și george@beia.ro cu subiectul: Practica ETTI 2017.

Studenti Aplicanti: 57, Total Locuri: 10, Locuri Neocupate: 4, Locuri Ocupate: 6

Situatia ofertelor si inscrierilor

Show 10 entries

Search:

| | Titlu | Departament | Localitate | Nr Locuri | Locuri Ocupate | Aplicatii Neprocesate | Aplicatii in Procesare | Acceptati pt. Interviu | Acceptati de Companie | Renuntari | Locuri Neocupate | Mai puteti accepta | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ● | Asistent cercetare | IT&C, Telemetry Solutions, R&D | Bucuresti | 10 | 6 | 8 | 40 | 6 | 3 | 5 | 4 | 1 | [Detalii] |

Showing 1 to 1 of 1 entries

FIGURE 13 - MATCHMAKING PLATFORM FOR INTERNSHIP

Also, other universities/faculties advertise a weblink with possible internship programs, for example ASE FABBV (Academy of Economic Sciences) : http://fabbv.ase.ro/practica

There are also several International internships initiatives:

- GrowingLeaders https://growingleaders.com/ Growing Leaders offers events, tools and comprehensive programs that promote healthy character and leadership development in the next generation. They also offer a variety of resources for the parents, teachers, coaches and mentors who shape their lives.
- ErasmusIntern https://esn.org/ErasmusIntern?gclid=CMCB06_S-NICFdQK0wodXyUNNg Erasmus Student Network (ESN) is a non-profit international student organisation. Their mission is to represent international students, thus provide opportunities for cultural understanding and self-development under the principle of Students Helping Students. ErasmusIntern.org provides an integrated market place that aims at bringing together traineeship providers and students seeking a training opportunity abroad.
- GoOverSeas https://www.gooverseas.com/internships-abroad Is a company of travelers who have studied, worked, taught, lived and volunteered overseas, and they assume that people have more information about the TV they would buy from BestBuy rather than when

they choose an internship program abroad, this is why they come with information right from the source and don't just come up with an endless list of internship programs that may or may not work out for you. Their mission is to "empower more people to spend meaningful time overseas."

- International Internships https://www.international-internships.com/ is a platform with customized internship placements, extraordinary customer service, and a great value. They provide an on-site orientation and on-site support for all of their programs and also they can arrange for students housing as well.
- GoAbroad https://www.goabroad.com/intern-abroad was first conceptualized for students with a desire to travel abroad and companies offering international programs. So their mission has transformed into something much greater than building a bridge between travelers and organizations: they've developed and evolved over the past two decades to meet the ever-changing needs of travelers, positioning ourselves as the resource for meaningful travel around the world.
- Global Experiences https://www.globalexperiences.com/ Global Experiences has been providing customized international internship programs to university students and graduates since 2001. We specialize in providing each intern with individualized attention to create fulfilling and career-enhancing international experiences in nine cities across the globe, including Barcelona, Florence, Milan, Dublin, London, Paris, New York, Washington, D.C., Stockholm, and Sydney.

# 6. SOMEDI PRELIMINARY USE CASES

SoMeDi proposes 2 thematic use cases, which all aim to illustrate different aspects of technological innovation, their use as well as novelties in terms of products/services. All the use cases exploit various technologies explained above.

## Social Media for Marketing purposes

The main objective of this use case is to improve the marketing performance of companies in two different sectors: e-commerce and restaurants with social media presence including promotions and marketing campaigns. By means of DID tool provided by SoMeDi, monitorisation on forums and social networks in order to check the impact of their marketing campaigns will be implemented to generate recommendations and automatized changes which can be used to increase sales volume and business impact.

- e-commerce

The main characteristic of these companies is that their business logic is based on the virtual environment.

The store transactions and trade has been transformed due to the worldwide expansion of the internet. The commodity for customers to find and compare searched products and the emerging markets has made e-commerce acquire great impact on worldwide economy. The physical stores have been forced to enter the online market for remaining competitive on their respective sectors, which makes e-commerce sector the ideal target market for improving their logic business according to their social impact on the internet.

E-commerce business logic is based on their online sales, so they depend on their social reputation. The objective is to integrate e-commerce system with the SOMEDI platform aiming to work together to increase the company sales and customer's opinion. The integrated system will support the customer's platform as well as the establish campaigns and promotions, providing social impact information the whole time.

- Restaurants

The main characteristic of these companies is that they focus their marketing strategies in two main approaches: word of mouth and social media. Currently restaurants highly depend on their social reputation, so they make use of social networks as Facebook or Twitter to disseminate new promotions, special events, etc. and attract new clients so their sales could be also increased. The objective is to integrate their social media strategy through the SOMEDI platform aiming to work together to increase the company incomes and customer's opinion. The integrated system will provide feedback on the campaigns and promotions impact, as well as impact of specific locals reputation, comments and opinions through opinion mining, etc.. In summary SoMeDi tool will provide reports based on social impact information helping the business to modify their marketing campaign to align it with clients needs, understand their brand impact and performance and in a whole help them to increase their impact and incomes.

This Use Case will be based on four axes: 1) Competitor Analysis based on Social Media, 2) Brand monitoring (analyse the reputation of the brand), 3) Event detection with sentiment detection, 4) marketing campaigns track and recommendations. Goal 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

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

- Online sales increase (e-commerce)
- Brand reputation increase
- Successful marketing campaign translated in more clients.
- Better knowledge about the clients' satisfaction and actions to be taken.
- Social activity

Partners involved in this use case are Innovati, Hi-Iberia, Taiger, and Alsus.

### Social media driven platform for recruiting purpose

The main objective of this use case is to find new ways to improve the employment rate of young adults using social media information. The focus will not be on recruiting process but on facilitating access of students to internship and to apprenticeship programs.

According to "Using social media in the recruitment process", a whitepaper by Robert Walters, has established the existence of distinct personal and professional social media sites, the whitepaper surveyed quizzed respondents, from UK,  on where they would typically look to advertise or search for a new role.

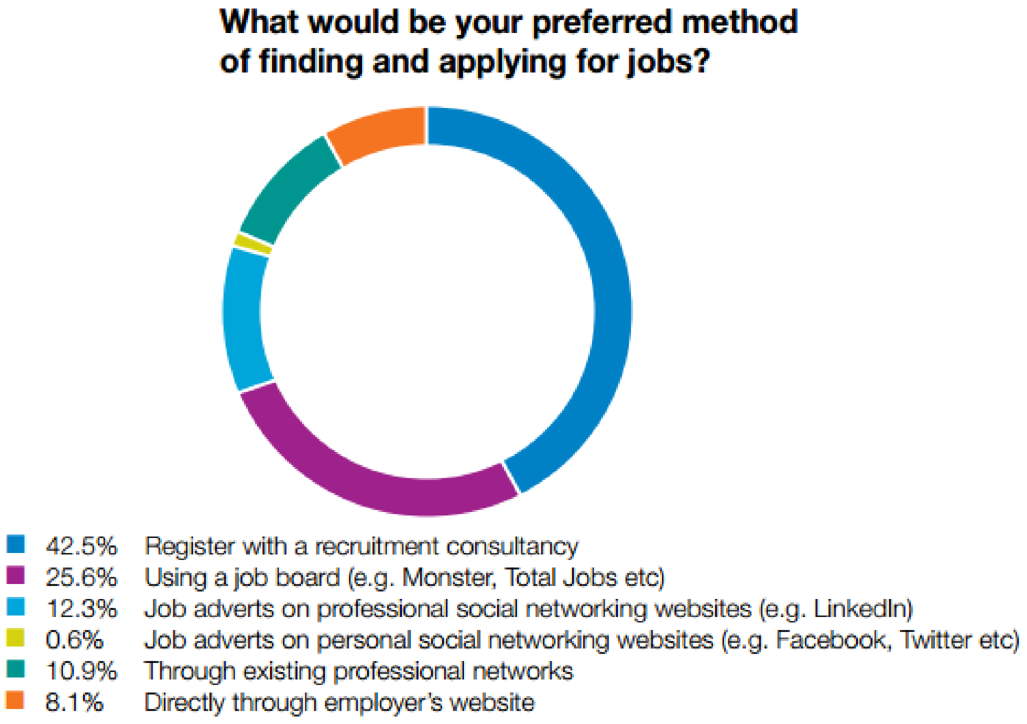


FIGURE 14 - PREFERRED METHOD OF APPLYING FOR JOBS - CHART

Methods outside of the social media space are the preferred option for the majority of job seekers, with 43% turning to a recruitment consultancy to secure their next move and 26% first looking at online jobs boards. Just over 10% of job seekers turn first to existing, 'offline' contacts such as friends or colleagues, while 8% search for adverts posted to company websites.

On the other side, recruitment preferences among hiring managers are similarly conservative, with 49% citing a recruitment consultancy as their most favoured option. Only 13% prefer to use professional networking sites such as LinkedIn to advertise vacancies. Adding job advertisements to the company's website is preferred by 18% of employers, with 12% first turning to jobs boards and 11% to offline networks and word of mouth.

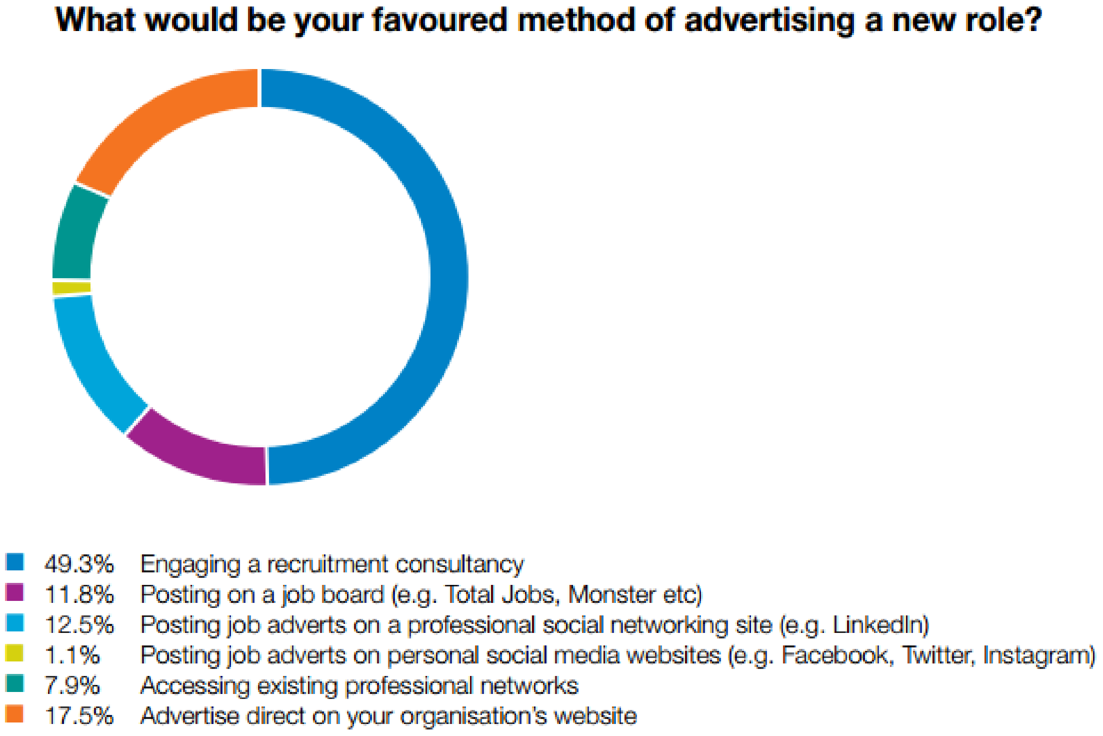


FIGURE 15 - PREFERRED METHOD OF ADVERTISING A JOB OPENING

Using SoMeDi's DID tool, the platform will generate:

- personalized recommendations: based on their education, interests, location etc. the users of the platform could receive personal recommendations in the form of career path advice; events like meetups, conferences and workshops; key companies/organizations;

- positive and negative opinions regarding companies that offer internship and/or apprenticeship programs;

- Social presence for the companies that offer internship;

- Internship position awareness in social media;

- Company profiling.

Also, considering the employers' appetite to verify the job candidate social media background, the SoMeDi DID tool will assist candidates in consolidating their social profiles to highlight their professional skills.

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

- Increase students' access to internship and apprenticeship programs
- Better collaboration between universities and companies regarding internship and apprenticeship programs
- Increase employment rate (students that follow internship and apprenticeship programs have a real chance to get employed within the same companies that provided these programs)
- In depth knowledge of the students regarding possible career paths according to their education level and profile

For the involved companies, users of this platform, use of  SoMeDi's DID tool will bring additional help when evaluating candidates.

Main partners involved in this use case are SIVECO Romania and Beia.

# 7. CONCLUSIONS

Through this document we have provided a vision of the SoMeDi context starting for the Manifesto, where the consortium goal regarding final DID Tool has been approached, with an overview of the initial functionalities envisaged for the tool.

Once the final objective has been defined the technologies needed have been analyzed starting from their current state of the art and analyzing the progress beyond the state of the art that the project needs to achieve in order to fulfill SoMeDi goal.

Once the SoMeDi vision and technologies have been analyzed, we identified the market needs that have motivated the consortium to start this project. For the needs identification is important to take into consideration that different domains (like the ones in the use cases) will have different needs and motivations although the solution could be the same, as same technologies are the solution to the problem.

From the different use cases just different kind of recommendations will be provided; but SoMeDi DID tool objective is to be a personalized tool that will allow the user to select input sources, target objectives, etc. So it could be used to provide added value to the data and user interaction in different domains just through the customization of the platform.

# 8. ANNEX

## Social media tools

- Crystal

Crystal[112] is an extension of Google Chrome browser that helps to personalize communication with others, giving advice about communication preferences receptor, based on data taken from Linkedin Profiles. The solution uses public data taken from Linkedin to detect the type of personality test based on the characteristics of DISC (Dominance, Influence, steadiness, Compliance).

- Live stream

In August 2015, Facebook began offering VIP users the ability to broadcast live video content using the Facebook Mentions[113]. Steam sites appear in the News Feed section and users can post real-time comments. This functionality has been launched to compete with dedicated applications such as Meerkat and Periscope. In January of 2016, this functionality became available to the general public, and in August of the same year Facebook launched a live streaming API that allows other devices (cameras, professional and drones) to integrate this functionality.

Figure 1 shows the Facebook functionality on an Apple device.

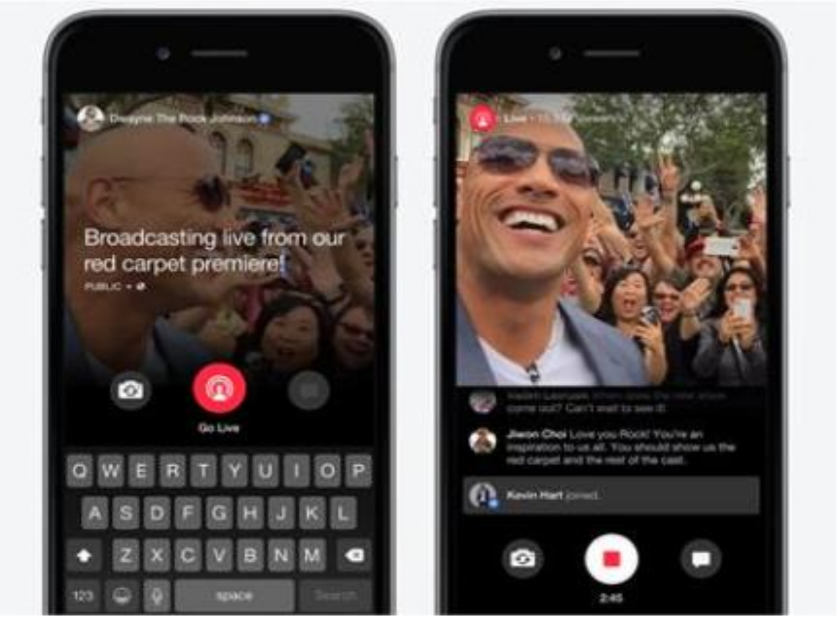

Dwayne The Rock Johnson
Broadcasting live from our red carpet premiere!
PUBLIC
Go Live
Q W E R T Y U I O P
A S D F G H J K L
Z X C V B N M
123
space
Search
Jiwon Choi Love you Rock! You're an inspiration to us all. You should show us the red carpet and the rest of the cast.
Kevin Hart joined.
2:45

FIGURE 16 - FACEBOOK LIVESTREAM ON IPHONE

[112] https://www.crystalknows.com/

[113] https://live.fb.com/about/

- Younited[114]

Younited is a secure cloud for storing your photos, videos and documents. Once the files are loaded into the cloud, they can be accessed from a computer, phone or tablet. Also, Younited offers the possibility of content distribution with other persons, turning it into a private, well secured collaborative space. In addition, the Facebook, Picasa and Dropbox accounts can connect anytime to the cloud. Once connected, all photos from these accounts will be synchronized in the Younited cloud. This service provided by F-Secure is hosted in Finland, offering increased security due to strict laws and regulations regarding data and personal files. Also, this service is designed and optimized by security specialists, protecting each user's daily content against data loss, malware, unauthorized access and other inconveniences.

- Welodias

Welodias115 is innovation management service offered as cloud service to collect ideas, feedback and comments with streamlined funnel. It enlarges expert network by offering single point service for all experts from other industries, research institutes and investment banks to join to innovate. Welodias enlarges also knowledge network by offering interface to several knowledge sources. It can be connected to external data bases as patent and research databases and even to social listening.

- OneRiot

OneRiot[116] is the largest real-time web search engine today. It aggregates information about popular topics from Twitter and other social media sites. OneRiot consolidates all these articles into one headline that you can click that leads you to a list of articles associated with a topic. This is a great way to see multiple articles on one topic all at once and in real time.

Instead of having to search different news sites looking for a hot topic, users can access OneRiot's aggregated real-time information on the topic. Also, OneRiot does a good job of blending hot topics from different spheres. It has a balanced mix of entertainment news and hard news, but seems to concentrate more on entertainment news.

- Collecta

Collecta[117] has more of a mainstream U.S. and international news focus, and does not offer the entertainment and lifestyle content that OneRiot integrates into its results. As a result, Collecta is a great source for people who want important news stories and like to view several sources to understand the breadth of angles on a topic.

Also, Collecta does a great job of gathering recent social media updates from several sites such as WordPress, Digg and niche blogs; whereas OneRiot seems to pull most of its content from Twitter.

- LinkedIn

LinkedIn is by far the #1 spot for job seekers, those currently employed, marketers who are looking to build lists and salespeople who are seeking out new clients. With 35 million users, including recruiters and job seekers, LinkedIn is quite a hot spot[118].

[114] https://younited.com/en/home

[115] http://www.inno-w.com/welodias-idea-management

[116] http://www.davidgeer.com/IEEEComputer_201003.pdf

[117] http://www.sciencedirect.com/science/article/pii/S0306457311000082

[118] http://dl.acm.org/citation.cfm?id=1531689

The problem is that most job seekers don't optimize their profile, cultivate their network, join and participate in groups, use applications and exchange endorsements. It's also recommended to use a distinct URL (linkedin.com/in/yourfullname) and an avatar that is consistent with the picture on other used social sites.

- Plaxo with simply hired

The real value in Plaxo[119] is the address book that keeps track of all of the contact information, including a Yahoo! Map indicating where the contacts live. Plaxo, which is owned by Comcast, is also integrated with Simply Hired, which is a job aggregator that searches thousands of job sites and companies and aggregates them in a single location. After building a Plaxo profile, it can be used as part of the recruitment process when applying for jobs with Simply Hired for success.

- Twitter with URL Blog or LinkedIn

Twitter breaks down communication barriers and allows talking directly to hiring managers, without having to submit a resume immediately to a machine[120]. Although Twitter is probably one of the best networking tools on the planet, it needs to be supplemented with a blog or LinkedIn profile

- Jobster

Jobster[121] is a powerful platform for networking with employers who are offering jobs, when searching. Allows uploading CV's, even embed a video resume, showcase links to a personal website, which are all unique differentiators.

- Facebook

Facebook can be used to get jobs. There are two main ways of acquiring a job through Facebook. The first is to go to the Facebook marketplace, which lists job openings or other opportunities in ones network. Aside from jobs, there are "items wanted" and a "for sale" listing. When searching for jobs, allows to see who listed the item and then message the person to show the interest. [122]

- Craiglist

Craigslist[123] is an extremely valuable job search tool. Most of the positions on Craigslist are for consultants (design/programming help) and at small to midsize companies that are hiring.

- MyWorkster

MyWorkster[124] focuses on exclusive networks for colleges, allowing students and alumni to connect for exclusive career opportunities. This social network allows to create a professional profile and network with potential employers. MyWorkster also has job listings, which are provided by Indeed, a job search engine and aggregator, which is very similar to Simply Hired.

- VisualCV

[119] https://www.plaxo.com/about/plaxo
[120] http://mashable.com/2013/02/09/twitter-job-search
[121] http://www.jobster.com
[122] https://books.google.it/books?hl=it&lr=&id=rm34TRqg3vkC&oi=fnd&pg=PP1&dq=MyWorkster&ots=XiwR1oHq5c&sig=NitOx0EffecM01qQnakuYQeoGxU&redir_esc=y#v=onepage&q=facebook&f=false
[123] http://www.websitetology.com/wp-content/uploads/2014/01/13.02.Craigslist.pdf
[124] http://myworkster.com/

VisualCV[125] understands the importance of personal branding in a job search. Instead of a traditional resume, allows to create a personal branded webpage, where can be added videos, audio files, images, graphs, charts, work samples, presentations and references.

After  the VisualCV is created, it can be displayed publicly or privately,  it can be emailed it to a recruiter, saved as a PDF or just forward the URL.

- JobFox

JobFox[126], like online dating, tries to pair the candidate to the job that best fits him. Their differentiator is their "Mutual Suitability SystemTM" that enables them to match the job applicant wants and needs to those of employers to find the best relationship. The system learns about the candidate skills, experiences, and goals and then presents the adequate jobs.

- Ecademy

Ecademy[127], like LinkedIn, is a prime source for professional networking. Although, there isn't a job search area on the site, 80% of jobs are from networking and this place is dedicated to it.

- Breezy HR

Breezy HR[128] is a dead simple, uniquely visual recruiting tool for SMBs. Import candidates from around the web with a single click using the Google Chrome extension. Post to more than 2,000 job boards and manage the candidate pipeline using a Trello-inspired workspace on any device.

- Job Adder

JobAdder[129] manages and organises the recruitment process for anyone who hires people, offering simplicity, mobility and superior support. JobAdder's comprehensive, cloud based recruitment management platform is affordable and used by thousands of recruitment professionals every day.

- The Ladders

With more than 6 million members, TheLadders[130] is the premier online job-matching service committed to finding the right person for the right job. In this regard, more than 43,000 recruiters use TheLadders periodically in order to find suitable candidates. TheLadders allows viewing potential employees available online that are potentially suitable and complete access to their profile, which includes information on how to contact and desired income.

- Hiring Solved

HiringSolved[131] is "Google for talent" making it easy to find people to hire. Also, it can find and combine data from many sources including resumes, web sites, social profiles, forum posts, and contact information to create a HiringSolved profile. Their mission is to make it possible to find everyone and make the information easy to search in a fast, easy to use search interface.

- Connectifier

[125] https://www.visualcv.com/

[126] https://jobfox.topresume.com/

[127] http://www.emeraldinsight.com/doi/abs/10.1108/02756660910987581

[128] https://breezy.hr/partners?gspid=semanticlabs

[129] https://jobadder.com/

[130] https://www.theladders.com/

[131] http://www.prweb.com/releases/2016/10/prweb13728444.htm

Connectifier[132] is changing the way recruiting is done today. Their engineering team from Google, Microsoft Research, Berkeley National Lab, Carnegie Mellon, and Stanford built large scale data analysis software that processes professional profiles from all around the web. By analyzing code repos, blogs, and social profiles Connectifier helps recruiters and hiring managers connect more efficiently with the candidates best for their openings. Connectifier was recently acquired by LinkedIn.

- Data.com

Salesforce Data.com[133] delivers contacts and company profile information, sourced from Data.com Connect and Dun & Bradstreet (D&B), right inside Sales Cloud. It's a liaison with decision makers faster, and easily plan territories with the latest, most accurate data.

- SocialMention

SocialMention[134] is a search engine and real-time analytics for social media which allows searching for terms in a wide range of media types and blogs, websites booking, events, images, news, videos and audio and other networks. The solution enables a wide range of search locations and quantify results in values osif real interest such as:

- Power -probability in which the keyword is mentioned in the social environment.
- Distance: measuring the influence of a term.

- BuzzSumo

BuzzSumo[135] focuses on the most distributed postings on social networks and through the platform allows to:

- Determine the most distributed posts by domain;
- The analysis topics, titles and formats of content work;
- Find valuable content according to your preferences;
- Discover anonymous postings and interview opportunities;
- The analysis of popular content of the competition by searching over the site's domain

- SearchWiki (SocialSearch.com)

The aim of SearchWiki (or Swiki)[136] is to link search queries to relevant results through improving the general web searches by community sites activation that can be initiated from any platform.

- Google Analytics, Yoast SEO

Google Analytics[137] is a tracking service for website's traffic. The service can be integrated on different platforms like WordPress and Drupal through a special module created for accelerating the association of a website with a Google Analytics account.

The plugins installed in CMS (Content Management System) have a basic interface where it can be set the main website's tracking filters. The Google Analytics account dashboard is complex, but comes with advanced settings for segmenting and monitoring the website traffic.

Analysing data obtained from Google Analytics is a key activity for establishing online marketing campaigns. Google Analytics accelerates the traffic analysis through an intuitive interface and the

[132] https://www.connectifier.com/
[133] https://www.data.com/
[134] http://link.springer.com/article/10.1057/fsm.2011.19
[135] http://buzzsumo.com/
[136] https://en.wikipedia.org/wiki/Google_SearchWiki
[137] https://yoast.com/tag/google-analytics/

tools included. The marketing campaigns, both online and offline influence the website traffic and also the conversions.

- Smart Recruiters

SmartRecruiters[138] transforms recruiting for leading companies in today's Talent Economy and the candidates they seek to hire. Atlassian, Ancestry.com, Marc Jacobs, Skechers, Equinox and IBM are among the tech, retail, hospitality and entertainment leaders using the SmartRecruiters Talent Acquisition Platform to run recruiting like a sales and marketing machine. Its modern platform delivers an unparalleled mobile and social candidate experience, engages hiring managers along the way, integrates quickly and seamlessly with HRIS, and supercharges recruiter productivity.

- Stack Overflow

Stack Overflow[139] is a question and answer website for professional and enthusiastic programmers. It is a fast-growing network of over 100 question and answer sites on diverse topics from software programming to cooking, photography, and gaming.

Stack Overflow receives more than 26 million unique visitors every month and helps developers find answers to any programming question. It also hosts a hiring platform, Stack Overflow Careers which helps its users connect with top employers all around the world to find new programming opportunities.

- GitHub

GitHub[140] is a web-based Git repository hosting service offering distributed revision control and source code management functionality of Git as well as adding its own features. Over 4 million people use it to share code. It is the place to share code with friends, co-workers, classmates, and complete strangers.

The collaborative features of the GitHub platform, its desktop apps, and GitHub Enterprise make it easier for individuals and teams to write better code.

- TalentStream Recruit

TalentStream Recruit[141] is an intuitive system for tracking applications and its aim is to attract talent quality in an effective way and the operability of internal recruitment processes. The software makes a transition of the candidate through a recruit process from the moment when he posts an application.

The company uses on average 12 different useful tools for recruiting and managing candidates.
Characteristics:
- Configurable ATS (Applicant Tracking System) Workflow;
- Optimized Career Site;
- Unique search process by combining past records, recent applicants and members with a high level of experience into one single way of finding existing candidates from the database;
- Broadbean Job Distribution which permits the transmission of an individual application to all online job platforms;
- Automated CRM (Customer Relationship Management);
- Real-time reports of applicant performance.

[138] https://www.smartrecruiters.com/
[139] http://stackoverflow.com/
[140] https://jobs.github.com/
[141] https://hiring.careerbuilder.com/talentstream/recruitment-software/recruit

- Handshake

Handshake[142] application - tightens the academic environment with labor market, the first job being just one click away. Offering high quality networking services, opportunities for both students and employers.

The application provides access to over 100,000 employers in all fields; most universities confirming once with the transition in the online environment, the employment opportunities have tripled after the first 6 months of platform utilization. Even from the planning stage, the application was designed to be easily customized and used. Students, employers and university staff have access to personalized content within the app using a browser or mobile device. But the most important detail is evident in increasing the engagement of students in the first year (an average growth of 50%).

The benefits of using the platform:

- Performance evaluation: students complete picture of their status, their needs and future destinations -jobs or internships.
- Event Management: digital invitations, advanced tools like check in, social media sharing functions can easily assess the success of an event.
- Fairs for career planning: from processing payments to recording employers, you can easily organize such events.
- Job postings: recrutors can easily post with an intuitive menu, thus connecting with an extensive network of academic centers.
- Planning: students can plan, view and manage courses, meetings easily from anywhere.
- Contact management.
- Advanced reporting features: with the help of sophisticated reporting tools you can easily answer any questions.
- Modern design: the application is developed by the same technologies that were created with your favorite web apps, keeping a familiar look.

- 48ers

48ers[143] has a more streamlined interface than either Collecta or OneRiot, but it also integrates multiple data sources and offers the ability to filter results based on source.

The best way to use this site for online marketing is to analyze a larger number of social media results on a particular topic and discover brand mentions or industry discussions. Results are being pulled by the second and thus have a lot more content to work with compared to the suggested social media results from OneRiot and Collecta.

[142] https://www.joinhandshake.com/
[143] https://www.48ers.com