Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A BIG DATA ANALYTICS FRAMEWORK FOR COMPETITIVE INTELLIGENCE SYSTEMS

[1]BOUKTAIB ADIL, [2]FENAN ABDELHADI

[1]Abdelmalek Essaadi university, faculty of sciences and techniques, LIST Department of Computer Science, Tangier, Morocco

[2]Abdelmalek Essaadi university, faculty of sciences and techniques, LIST Department of Computer Science, Tangier, Morocco

E-mail: [1]bouktaib.adil1@gmail.com, [2]afennan@gmail.com

ABSTRACT

Nowadays, companies are facing a lot of challenges due to the volume and velocity of data available online, the nature of this data which comes in different formats structured, semi-structured or unstructured, forces the adoption of new tools and techniques to process and transform data to knowledge, competitive intelligence systems aims at setting-up tools and software to handle this stream of data from data collection, data analysis, data visualization to the results dissemination for stakeholders to enhance the decision making process of companies. In this paper, we present a big data analytics layer/framework for competitive intelligence systems and we implement it in the case of XEW 2.0 system relying on Apache Spark capabilities and big data analytics technologies, we validate the proposed framework with a case study about Research in Morocco in order to achieve a technological surveillance, the framework shows promising results in providing analysts with a toolbox to extract strategic information.

Keywords: *Big Data*, *Competitive Intelligence, Big Data Analytics, Apache Spark, Data Mining*

1. INTRODUCTION

Data management is crucial to modern companies, big data era comes with new challenges [1] that requires new techniques and tools to handle them, from the volume of data that increases every day with the trend of people and organizations to use new communication technologies, to the velocity of data which impose companies to think about handling real time streams of data, and ending by the variety of data which means that the information we seek is not stored in one place but can be found partially or fully in multiple sources. Tacit knowledge can be extracted from data using data mining techniques [2] and text mining techniques [3] to extract knowledge from textual data.

Competitive intelligence systems are a trend in modern companies, many organizations are trying to develop and implement systems to enhance the decision making process, by automating the process of data collection and gathering from multiple public sources to respect the legal deontology of intelligence [4], and analyzing data with the use of advanced data mining algorithms and machine learning paradigms [5], ending by data visualization and knowledge extraction from graphs and visual diagrams to ease the process of interpretation.

Strategic information management requires tools of automatic data gathering, and the adoption of surveillance strategies, many CI strategies [6] were proposed to monitor and enhance the position of companies and organizations through the analysis of many environment factors. SWOT [7] and PEST models are some popular examples of strategies [8] that were proposed to conduct an analysis of the environment factors surrounding the company like political, economic, social and technological factors that must be monitored continuously by an updated data collection approach.

Big data analytics gained a lot of attention in the computer science community due to the success of its applications in many fields [9], companies are investing to gain the capabilities of this technology [10], and enhance their arsenal, the increasing volume of data and the variety of sources needs new approaches and technologies in order to process and extract hidden information from the raw heterogeneous nature of data online [11].

Big data strategic information [12] extraction require new software and systems to support unstructured data that comes in large volumes and raw format that needs transformation and homogenization from different sources, one of the systems that were proposed we found the XEW model [13] which describes and decompose the system to multiple layers, Data collection layer combining multiple agents for data scrapping and crawling , Data warehousing service to store data in a data lake architecture , Data analytics relying on Weka and R for data analysis and ending by Data visualization to communicate knowledge to stakeholders.

2. COMPETITIVE INTELLIGENCE AND BIG DATA ANALYTICS

2.1 Competitive intelligence

Competitive intelligence is an iterative activity and process of transforming data to intelligence [14], a knowledge that is profitable to organizations and valuable to companies, it's the process of translating data to information to tacit knowledge so it can be disseminated to decision makers in order to execute a strategic response to any kind of change in the environment. Competitive intelligence varies from country to country [15], the culture of intelligence is different in every continent, looking at the history of CI it started in the military field and it has some non-legal aspects and it was considered as espionage [16], nowadays this process is applied on an economic level where the companies try to legally monitor their environment and the movements of their competitors to enhance their position or guarantee their dominance in the market. CI cycle is defined [17] and there is a consensus about its steps (Fig .2) that starts from the definition of strategic business need, then the determination of potential data sources that may contain the targeted information, with the aim of data analysis and processing in order to extract hidden information and latent signs from the raw data, then the visualization of the processed data to facilitate the interpretation of the knowledge and the valorization of collected information. The dissemination of intelligence analysis result to stakeholders gives strategic advantages to the decision making process [18].

The challenges facing competitive intelligence are numerous and increasing in the modern age, due to the evolution of communication and information technologies and the changes it made on a lot of, cultural, technological and social aspects surrounding the environment of companies [9]. In order to face those challenges CI tools must be able to handle large volumes of public data available online, and they must enhance their data analysis capabilities in order to analyze the data that comes in unstructured, heterogeneous and raw ambiguous nature or form, that needs to be cleaned and prepared for accurate analysis and robust knowledge extraction.

The automation of this process of intelligence gathering is beneficial and important for the growth of companies and organizations, and it will optimize the time and the cost of data collection from formal and informal sources, most of formal data is available online and can be scraped and crawled using web mining techniques [19], data about culture, social behaviors and tendencies, or technological changes and advancements in a specific domain are all published and made public online.

Many attempts were done to define the competitive intelligence process, but most failed in finding a unique definition, due to the differences in everyone perspective and objective from applying CI, so many economists proposed many CI models and strategies [8] in order to limit the scope of this field.

Competitive intelligence and information systems goes along together [20], a combination or a hybrid perspective of those two models can be beneficial and crucial to the competitiveness of companies and organizations, in order to take advantage of internal and external data in order to enhance their strategic decision making and enhance the quality of decisions by relying on accurate data and advanced analytics, that helps in extracting latent and hidden knowledge from unstructured data mainly using data mining techniques [21]. Business intelligence are a set of tools and software that analyze the scope of the organization while competitive intelligence aims at monitoring the environment of the organization in order to determine and enhance the position of the organization in the market.

Competitive intelligence goes along data, where there is data there is intelligence to be harnessed and collected and tailored to create knowledge, big data comes with challenges that need innovative solutions, new technologies, robust

algorithms, and advanced architecture to support the volume and velocity of data coming every day and minute with incremental volume and capacity

So what are the benefits of using big data analytics in competitive intelligence systems? Which tools are suited for this type of data? What are the algorithms that can be applied in this field to analyze the data? And how can we visualize analysis result and disseminate it to stakeholders? How CIS is designed and implemented? Which Model to follow?

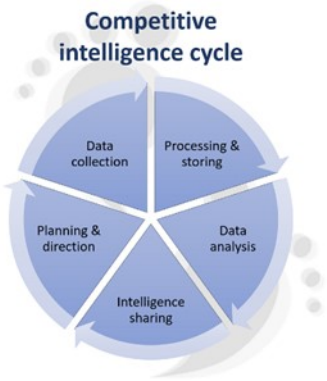


Figure 2: Competitive intelligence steps

2.2 Big Data Analytics

The history of data analysis, data processing was done on simple machine using simple techniques and algorithms [22] that converge to the required solution in the optimal time and using the available resources without overconsuming and no shortage of resources, in case of need of more resources, algorithms were enhanced to adapt to the new requirements, NP problems[23] though still give a headache to researchers and analysts due to the complexity they present and the creative approaches it needs to approximate the targeted result. With the explosion of data and information online, those techniques and classical tools and software that was used in solving those problems are not enough to handle the huge volume and velocity of new data coming in every day and increasing every second.

Big data analytics is created from a need of enhancing the capabilities of large volumes processing and fast velocity data digesting [24], the first technologies to appear to tackle those problems we find GDFS google distributed file system [25] that comes to solve the problem of dealing with huge volumes of data, which needs a clustering approach to distribute data into multiple nodes in order to enhance the storage capability. The appearance of Hadoop file system and Map Reduce [26] did revolutionize big data processing field, and brought technical solution and architecture and infrastructure to handle big data analysis and unstructured data processing relying on distributed file storage and distributed horizontal processing. Cloud computing came to handle the infrastructure of big data analytics and gave companies access to data centers to perform their analysis on the cloud taking advantage of all machine and nodes available on a cluster to execute tasks and conduct advanced analytics on data.

Apache spark is an enhancement of Hadoop HDFS, which propose fast lightning speed of data analysis and a comparative of this technology with Hadoop proved to be more beneficial [27] and gives promising result when it comes to processing big data.

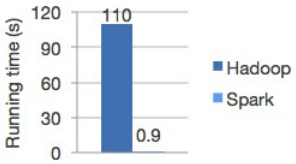


Figure 3: Apache spark vs Hadoop running time

Big data analytics applications were widely used and proved to be efficient and fruitful when applied on many fields and domains to solve a variety of problems, smart city big data analytics [28], internet of things uses it to analyze the evolving volume and velocity of data, so data in economy is a domain where big data analytics have an increasing demand, beside business intelligence that analyze internal data and tries to perform analysis on business keys and performance indicators, competitive intelligence on the other hand relies on the analysis of unstructured data using data mining, text mining, predictive analytics and web mining to analyze the nature of the data surrounding an organization. In the next section we will show the benefits of using big data analytics in competitive intelligence systems, those advantages encouraged the proposed framework to use big data analytics technologies and architecture to exploit the benefits of advanced algorithms and approaches that can be applied in competitive intelligence analysis.

2.3 Advantages of big data analytics in competitive intelligence

Big data analytics can provide businesses with multiple capabilities [29] and tools that allow them to analyze their environment by achieving the following goals:

- ✓ Increase visibility by making relevant data more accessible in data storage facilities.
- ✓ Facilitates performance improvement and variability exposure by collecting accurate performance data.
- ✓ Complements the decision making with automated algorithms by revealing valuable insights.
- ✓ Generate new business models, products and services.
- ✓ Knowledge creation, new management strategies implementation.
- ✓ Innovate new product and service development ideas through weak signal detection.
- ✓ Enable a dynamic market analysis.
- ✓ Accompany managers in the decision making process through customized dashboards.

3. COMPETITIVE INTELLIGENCE SYSTEMS

3.1 MEDESIIE

The MEDESIIE (MEthode de DEfinition de Systèmes d' Information en Intelligence Economique) project aims at developeing a method of strategic information monitoring system for competitive intelligence [30], the model studies the strategic requirements of companies, it aims at describing the company by its model, its strategy, its environement, competitive intelligence requirements, and its products/services.

This model is based on 4 concepts, the first one is Way of thinking which focuses on companies' strategy and decision making process which view the company as an organization with tacit capabilities, Way of modeling that represents a new complexity to the design and conception of Information System with the introduction of a Meta Model composed of a various of ontologies around the major players in Competitive intelligence (Environment, Company, Decision Maker), Way of Organizing following the design by prototype to take in consideration the information needs on the environment, Way of supporting that aims at guiding the designer and providing necessary tools and software to implement the targeted CI Model.

This model remains a high abstracted model for designing competitive intelligence systems, by adapting the system information design paradigm to competitive intelligence, in order to create a product to solve the informational need of companies, and transform a decision problem into information search problem.

3.2 XEW Model

Amine el Haddadi et al [13] proposed an innovative model to design a modular architecture for competitive intelligence systems baptized XEW 2.0 as in Fig 4 by combining multiple techniques in Information Research, Extraction, Big Data Mining and Big Data Visualization, the system was inspired from different previous models and took advantage of each one to propose a model that respond to the modern companies need in the era of big data, the proposed system contains multiple layers:

- ✓ XEW Sourcing Service (XEW-SS) is a service layer that allows searching, collecting, and processing data from mutiple sources at the same time.
- ✓ XEW Data Warehousing Service (XEWDWS) is a service layer that allows the storage of heterogeneous data, after homogenization step, in a multidimensional form for further analysis.
- ✓ XEW Big Data Analytics Service (XEWBDAS) is a service layer that allows the multidimensional analysis of the corpus using data mining algorithms relying on R and Weka.
- ✓ XEW Big Data Visualization Service (XEWBDVS) is a service layer that allows the communication of extracted knowledge using an interactive and collaborative dashboard.

The first step of this model is the definition of the inforamtion problem, and the detemrination of possible data sources whether they are formal or informal, then the storage and processing of hetergonsues data in ordre to store it in multidimensional form, then analysis are performed relying on big data minig techniques that handles the unstructured nature of data. once the analysis is done, results are disseminated using data visualisation techniques in order to ease the interperetation and the understanfing of the strategic information extracted from the corpus.

In this paper we enhance the XEW data analytics layer by proposing a novel framework to

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

analyze big data using Apache Spark as a technology to conduct big data analytics [27], machine learning, data minng and text minning algortihms, and we show how the use of the proposed layer gives promising results and allows the application of a variety of algorithms and processes in order to overcome the challenges of big data and enhance the decision making process [31].

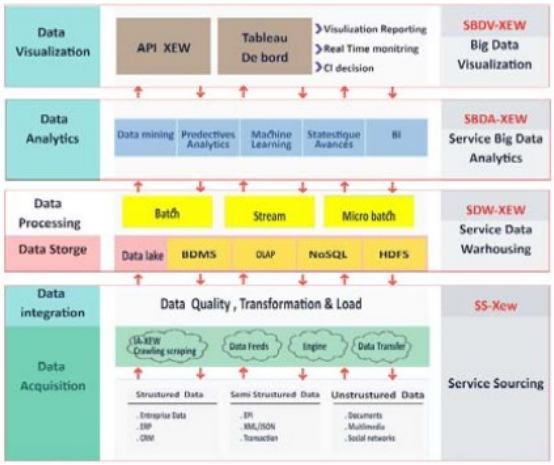


Figure 4: XEW 2.0 architecture.

4. CI STRATEGIES

4.1 PEST Analysis

Pest model focus on the monitoring of poltitical, economic, social and technological factors surrounding the environement of the company or organization [32]. Most of this factors are availble online due to the data explosion that came with the spread of internet, people share all ascpects of opinions and information, the analysis of this data can give the company a competitive advantage in knowing its surrounding, with a minimum need of competitive intelligence personnel, social media data for instance can give geo-oriented analysis in order to analyze the culture, politics and trends in people by filtereing on geo fileds, which can help companies in relying on accurate real data to extract the featuers and charatectsics of any market they are intersted in, scientists and reasearchers are publishing scientific document and material every day in scientific databases to communicate their advances and findings, which can be exploited to conduct a technological survaillance in a specific field.

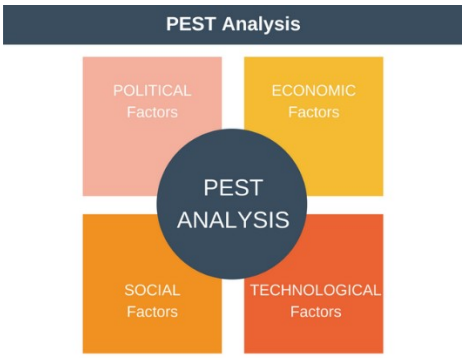


Figure 5: PEST Analysis.

4.2 SWOT Analysis

SWOT matrix analysis is one of the popular strategies of competitive intelligence [7], the aim of this model is to analyze weakness and strengths from internal data to enhance the operational system of the company, and threats and opportunities from monitoring external factors such as the entering of new competitors, new products, or new collaborations that may threatens the sustainability of the business and the stability of the company if they don't plan a strategic response to tackle the problem. The main focus of competitive intelligence is the analysis of external data which contains factors of the environment to be monitored and analyzed, through the use of web mining methods.

Figure 6: SWOT matrix.

5. RELATED WORK

Many researchers proposed CI systems that respect the steps in the process of CI and responded to the analytical needs of companies. JMP [33], is a powerful commercial tool for competitor market analysis its power resides in the use of a variety of data sources and the possibility

to analyze them in one software to give quick results about a specific topic, but they cannot handle huge datasets and the analysis cannot be done on a large scale which is a limitation in modern business's needs.

Vicki L. Sauter [34], is one the early researchers who proposed a web mining method of designing competitive intelligence systems, and illustrates the application of such method in a case study about tertiary hospital, by using data mining and artificial intelligence tools to analyze qualitative data collected from the web, in order to facilitate the process of decision making.

Xew 2.0 [13] is a promising competitive intelligence system it is designed for big data and it can analyze data from multiples sources and store them in a data lake environment which is a perfect choice for storing raw data by adopting the schema-on-read paradigm to allow further analysis on the same data without losing any part of it. XEW 2.0 proposes a scalable and modular architecture through its use of many micro services, from the sourcing service to the big data visualization service. But its analytics component doesn't support distributed technologies and rely only on R and Weka that have limitations in dealing with huge datasets that are high dimensional in nature.

Jesus Silva. et al [35] proposed a custom system of competitive intelligence and technology surveillance which relies on web mining techniques and machine learning for better collection of data, but it does not address the data analytics part, and doesn't propose a solution to handle huge amounts of data.

Miguel Ángel Ospina Usaquén. Et al [36] designed a competitive intelligence system using business intelligence to analyze data coming from meat sector companies, the proposed system relies heavily on Power BI, a Microsoft product that allow data analytics and visualization on a desktop software, but it has some limitations when it comes to handling big data and running distributed computations which is a must specifically in text analytics.

Van der Berg et.al [37] provides an example of application of the CI process on sport performance analysis, in order to collect and analyze surrounding environmental and competitor information for a better decision making, the study aims at giving the sport coaches an information management process to support them in their choices and strategic decisions relying on a systematic analysis of competitors.

Choi.J et al [38] proposed a social data mining approach to conduct competitive intelligence on a specific market, by monitoring product opportunities stated in opinions mined from the customer reviews on social platforms, and by enabling a real-time system to extract customers' needs in an evolving environment, where the time of access to information is crucial to the survival and growth of companies offering competitive products in the market.

Lutz, C.J et al [39] proposed open-source data based competitive intelligence for stakeholder's industry analysis, and an automatic approach of CI cycle in order to face the evolving market structure and factors and support daily strategic decision making relying on unstructured and qualitative data analysis.

Sahin, M. Et al [40] proposed an implementation of competitive intelligence cycle in Turkish airplane industry in order to monitor environment changes that can affect the business like oil prices and political stabilities, the authors developed a model to suggest improvements in the thinking process of the company on different levels by using a two-step cluster analysis to uncover hidden clusters that contain strategic information.

Villanueva, M et al [41], proposed competitive intelligence for managing knowledge and innovation in Argentina territory, by monitoring technological surveillance extracted from patents scientific publications and identifying leading player in the market by analyzing actors network, the proposed model uses data mining and text mining tools and software in collected data from scientific databases.

Köseoglu, M.A et al [42], proposed a competitive intelligence model and intelligence analysis of big data using text mining techniques with combination of network analysis the study showed promising result by applying it in the hotel industry by analyzing customer online reviews.

Yafooz, W.M.S et al [43] showed the importance of using big data analytics and text mining and machine learning as a crucial tool to discover knowledge and extract meaningful insights from big unstructured data that is available and

stored online in platforms such as Facebook, twitter.

Patel, J [44] proposed a big data lake model to support the storage of evolving volumes of enterprise data available online, his work relies on the adoption of distributed environment to solve the scalability issue of large volumes of data and their analysis to generate insights and enhance the data-based dictions of stakeholders.

After reviewing the existing solutions and architectures, and the study of work [43] that inspired us to use the social media platform Twitter as a data source which is considered as an assessment of a company's business environment, and due to the lack of use and adoption of big data analytics technologies in previous works. We found a need for a new approach in big data analytics component in most of the available competitive intelligence systems.

Compared to the previous works our proposed framework will introduce Apache Spark as an emerging big data analytics technology. The novelty of this framework resides in the distributed approach for analyzing the corpus on a cluster which will enhance the computational time for extracting hidden information in data, and allow the execution of iterative models that needs to write and read from hard disk space in every iteration, RDDs items of Apache Spark will solve this issue by using data on memory for fast access.

We propose in this article a framework that allows distributed analysis and parallel computing of machine learning, data mining and text mining algorithms to be run on an Apache Spark based environment in order to analyze big data collected in the context of a competitive intelligence project.

6. PROPOSED ARCHITECTURE: XEW ANALYTICS

In light of the findings and the literature review done on competitive intelligence systems and big data analytics platforms, we propose a new big data analytics framework for competitive intelligence, which contains 4 components shown in Fig. 1: Data collection, Data Storage, Data Analysis, And Data visualization, we present the details of each component in the following sections.

The main contribution of this article is proposing the technological environment to conduct competitive intelligence and respond to information needs of an organization, we enrich the framework by implementing a several data analysis techniques using apache spark to give analysts a toolbox to extract valuable knowledge, then we validate our framework by a case study on the synergy of research in morocco.

6.1 Data Collection

6.1.1 Smart query generation engine

Data collection must respect the recall and precision [45] measures to enhance the quality of data to be analyzed and avoid the paradigm of garbage in garbage out which comes from analyzing unrelated data, which may give false results and inaccurate analysis, when it comes to intelligence we must define a process to help us filter the data and keep only the data that relate to our subject with the aim of keeping recall to 100% and precision to a value close to 100%.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

6.1.2 Semantic WEB and Ontologies

We define domain ontology to support data collection process, an expert can introduce the ontology of the domain to be extracted [46], using synonyms of all concepts related to the targeted domain, which helps the query engine to create custom search equations to query from the scientific databases available online.

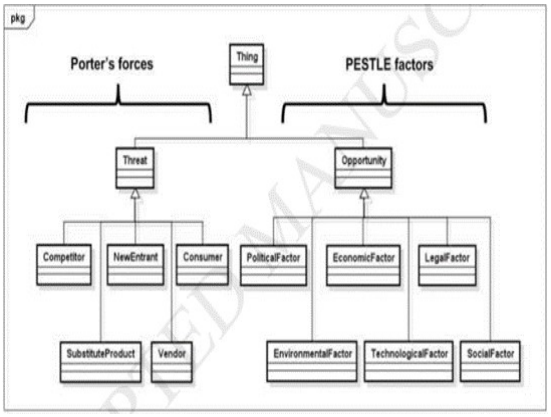


Figure 7: Domain Ontology for competitive intelligence.

By introducing the search term query, the inference engine works on finding synonyms and related terms to generate search queries to be executed on the database search engine, with the aim of filtering all relevant documents. Which will help us in getting accurate analysis result and valuable knowledge to support the strategic planning and enhance the decision making of organizations.

The collection process is based on scripts developed using Scrapy [47], a python framework that enables developers to use a modular approach in crawling and scraping various websites in HTML and extract content from various page tags, to be stored and analyzed accordingly by the framework.

We can scrape data from patent databases as well as scientific journals, news websites, forums and blogs and store them in the platform, by combining this sources we can get rich and complete insights about specific topics introduced by users.

We use Apache Flume [48] as a streaming technology to get data from Twitter Streaming API in real-time, we create agents for each topic of interest which is presented by a set of keywords. Then we store the data in the storage layer and we passed it at the same time to spark streaming to be analyzed in real-time by configuring Flume agent sinks.

6.2 Data Storage

6.2.1 Data Warehouse

CI related data comes in unstructured form, which requires innovative data storage and a homogenization step to ingest data into one format to be stored in a Datawaerehouse where the schema of data is mandatory and the handling of raw data is not possible. The scraped data from the web sources needs a cleaning step, then all the different extracted fields must be homogenized to form a unified dataset and schema to perform analysis and keep the strategic information system available for future strategy audit in order to learn from the mistakes and valid decisions made using the proposed system.

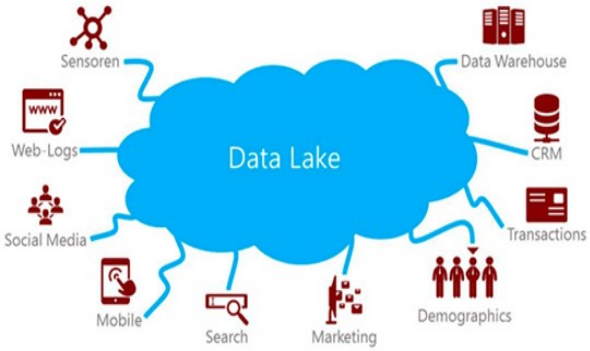


Figure 8: Data Lake Model.

Data lakes [49] design is most suited in this kind of unstructured data, due to its use of the paradigm schema-on-read which allows data to be stored in raw formats in a single system, this raw data can be reused for many tasks such as data analytics and machine learning or for reporting and visualization due to the flexibility they offer and the information they keep, in contrast of the data warehouse which removes noise data that can hold implicit information that could benefit the company in the future.

No-SQL Databases, characteristics and technical features allow storage of document oriented data [50], to support raw data with no prior schema, thus in our architecture we adopt the data lake architecture to store retrieved web documents in their raw format, and then define schema when needed in analysis.

All data is stored in its raw format in our data storage component which is composed of MongoDB and Apache Hadoop, both solutions support storing data in a distributed manner, we choose those technologies for the following advantages:

- ✓ Handling huge volumes of data.
- ✓ Storing semi-structured and unstructured data
- ✓ Scalability and fault tolerance

In our architecture we prepare for an implementation of a lambda architecture Data Lake, we adopt the approach of schema-on-read to transform and format the raw data to the model we are interested in during the analysis, and keep the original in its raw formats for future analysis in order to avoid the loss of potentially valuable information.

6.3 Data Analysis

After collecting data and storing it in its raw format in the data lake, we implement the data analysis layer using Apache Spark, where we create a distributed computing cluster, in order to execute machine learning algorithms in a parallel paradigm for faster results relying on the Resilient Distributed Datasets RDD technology proposed by Apache Spark for in-memory processing as solution to big data analytics and processing of large volumes in optimal time in the next sections we present the definition of each component of Apache Spark and we present the implemented functionalities in our framework for competitive intelligence analysis.

6.3.1 Apache Spark

Apache Spark is a novel distributed computing system, that proved a high time optimization and low execution time compared to other big data technologies like apache Hadoop or apache Flink [51], Spark is based on the Map reduce Paradigm but with another higher Abstract level to ease the programing task for developers by defining an API that allows multiples manipulations and functions, with no need to worry about the details of map reduce tasks execution, due to the adoption of DAG task design many algorithms can be executed in parallel and results of each task can be aggregated according to a user defined function, for better results and advanced analytics, which explain our choice of this big data analytics technology as the main processing layer in our framework.

6.3.2 RDD

Resilient distributed datasets [52] is a fundamental data structure of apache spark, they are immutable objects that lives in the cluster in different distributed nodes, when algorithms are executed they use each RDD in every Node to collect data and aggregate the results of execution of each task in one Node, most of the time in master node, RDDs are loaded in memory and they are easily accessible for Apache spark API in order to execute iterative algorithms with no need to reload data every time from disk space, when dealing with large volumes the data is partitioned across the cluster in order to be able to process each partition on its own to optimize execution time and perform accurate analysis on big data.

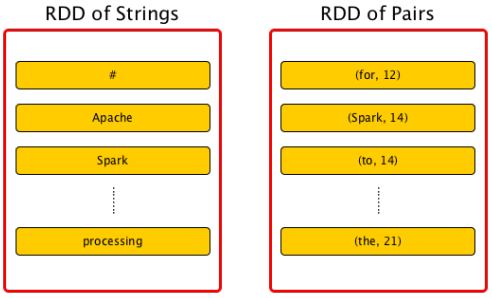


Figure 9: Example of RDDs.

6.3.3 Spark Mlib

Spark Mlib provides developers and analysts, a variety of Machine learning APIs in order to perform different analysis on the corpus, from basic statistics like correlation analysis or chi-square test, to feature extractors like TF-IDF and Word2Vec [53] to advanced classification and clustering algorithms like Decision Trees, Gradient-boosted tree, and K-means classification. When dealing with competitive intelligence data it is advantageous to have multiples tools and techniques in hand to extract knowledge and valuable insights from the corpus of collected external data.

6.3.4 Spark GraphX

For a thorough analysis of competitive intelligence, it is important to do analysis on collaboration network representing the actors in the environment, or semantic network containing tacit knowledge and implicit innovations and weak signals to be monitored and extracted doing advanced graph analysis, which is offered in Apache Spark GraphX API that allows the creation of large graphs across multiple nodes in a cluster.

In case of authors network analysis [54], we define edges as authors and the collaboration relation between them as vertices. We define G a graph of n vertices and m edges, the graph is distributed on a k nodes of the cluster, GraphX API allows multiple algorithms and graph analysis, like Page Rank, extracting connected components and Triangle counting, in our framework we implement PageRank in order to define the most important authors/actors in our collaboration network

6.3.5 Spark Streaming

Spark provides stream RDDs that allow handling of real time data, in order to keep the analysis corpus updated and ready to execute incremental analysis. The API allow the manipulation of RDDs sequences. One of the problems that this library tries to solve is the fault tolerance that can occur while dealing and processing of streams of data, loss of data can affect the analysis and gives inaccurate analysis and misconduct the strategic real time responses of companies to a change in the environment, whether after a wrong interpretation of people's reaction or after a product announcement in social media for example.

Spark Streaming [55] can be integrated with many inter-system messaging technologies like apache Kafka or apache flume, which can guarantee the communication of large volume of stream data without loss of data.

6.3.6 Implemented Functionalities

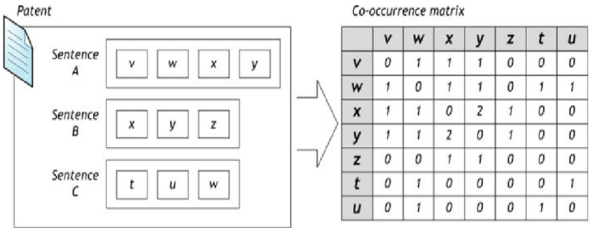
After the presentation of the technology stack that will be used in data analysis in this section we will present some of implemented analysis we can conduct on collected corpus to extract implicit knowledge and hidden facts for competitive intelligence, most of analysis will be using textual content of collected web documents after the cleaning and preprocessing step, the following analysis contains natural language processing, text mining, and machine learning algorithms, each technique can give a different analysis and a new insight, when applying multiple analysis on the same data we can enrich our knowledge about the targeted domain or fields we want to monitor and analyze to gain a competitive advantage.

6.3.6.1 Matrix Creation

The analysis we will conduct needs matrix creation and manipulation to execute multiples algorithms, RDDs allow the creation of distributed matrices using the map reduce paradigm, we can create voluminous matrices. And by following the model of competitive intelligence proposed by Ilham Galmalah [56], we implement the model of multidimensional data in apache spark in order to process the created matrices and manipulate them to extract different types of information and knowledge.

6.3.6.1.1 Co-Occurrence

The advantages of using apache spark capabilities is the ability to create large distributed matrices on Apache Spark cluster, and run different analysis on them, RDDs allow iterative algorithms to reuse the same data in memory with no need to reload data every time which optimize the time of analysis execution, especially near real time analysis where the time is a crucial factor in deciding the efficiency of the system.



	v	w	x	y	z	t	u
v	0	1	1	1	0	0	0
w	1	0	1	1	0	1	1
x	1	1	0	2	1	0	0
y	1	1	2	0	1	0	0
z	0	0	1	1	0	0	0
t	0	1	0	0	0	0	1
u	0	1	0	0	0	1	0

Figure 10: Co-occurrence Matrix example.

The co-occurrence matrix Fig 10 can be created by crossing two variables of multiple values, the cell crossing a variable 1 with variable 2 contains the number of occurrence of the two items. In text mining we generally cross the text variable with another variable or the text itself to create semantic network or the evolution network of a domain in time

6.3.6.1.2 Contingency

Contingency matrix is the same as the co-occurrence matrix but the cell crossing two items contain 1 if there is an occurrence or 0 if they don't cross in the corpus, this matrix can be used in correspondence analysis and factor analysis that tries to explain the relationships between two variables.

6.3.6.1.3 Presence-Absence

The presence-absence matrix crosses two variables, and the value of the cell contains a binary value 1 or 0 indicating the presence or the absence of any occurrence of the two items, generally this matrix gives the adjacency matrix and can be considered as a graph, and network analysis can be applied to be applied on the data, when crossing author with authors we can plot the graph of authors network and study the collaboration of authors, when crossing the terms we can plot the semantic network of the corpus and do analysis on the resulted graph.

6.3.6.1.4 PCA

Principal component analysis [57] is a method of dimensionality reduction, that aims at reducing the variables of the corpus into a smaller space, in order to visualize and facilitate the understanding and the analysis of the data by removing noise and outliers. Principal component analysis relies on an algebraic operation that reduces the dimensions of the original matrix by performing a decomposition on a matrix M.

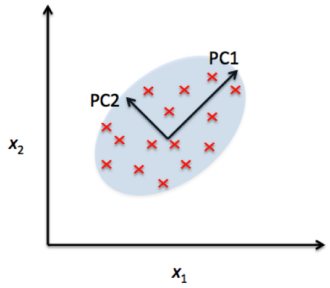


Figure 11: Principal Content Analysis dimensions.

6.3.6.2 Latent Semantic Analysis

Once we create the matrices, we can perform multiple analysis, one of the most popular analysis that we can conduct on textual data is Latent Semantic Analysis, in competitive intelligence, analysts are interested in understanding the latent topics hidden in text data contained in the collected corpus in order to extract insights about a specific field or technology.

LSA [58] implementation in Apache Spark is performed relying on SVD decomposition, an algebra technique that decompose an initial matrix to three matrices in order to reduce the dimensionality of original matrix by removing noisy data. We start by creating a term-document matrix by crossing all terms with all documents and counting the occurrence of each term in each document, this matrix is high in dimensions by nature because of the big number of documents n and terms m, after the execution of LSA we end up with three matrices U, V, and Σ according to the following equation (1), the resulting matrix U is an n by k matrix containing terms and their related concepts, Σ is a k by k matrix containing the importance value for each concept, and V is a k by m matrix expressing concepts in regard to documents. In this analysis we are interested in the extraction of latent topics that may contain hidden information in a specific domain which can be a sign of innovation or a threat to the company product.

6.3.6.3 Latent Dirichlet Allocation (LDA)

LDA [59] latent Dirichlet allocation is another rich algorithm proposed by the spark Machine learning library to perform semantic clustering of the corpus, this algorithm can be applied in our case alongside the LSA algorithm to extract hidden and latent topics in order to monitor threat and opportunities. The algorithm is based on the assumption that documents are a set of topics and each topic is defined by a set of terms, so it starts assignment of each term to a specific topic accordingly. LDA is based on a distribution of dirichlet. This algorithm was applied successfully in wide range of text classification and topic modeling problems, hence the choice of LDA in Competitive intelligence analysis which will enforce and enhance the quality and accuracy of environment monitoring and knowledge discovery.

The Algorithm takes as input a set of text documents and k the number of topics to extract, and by the analysis we can get k topics defined by a set of words, those topics are latent in nature and by further interpretation, analysts may derive valuable insights from the hidden topics that cannot be seen and monitored easily even by experts due to the large volume of data, and processing time needed by experts to read and understand the details of each domain and discover possible new innovations, threats or opportunities.

6.3.6.4 Near-real-time analysis

We will follow a hybrid approach to combine data from static web documents that update weekly or monthly, especially in the scientific databases where advances in research take longer time to be published, so we want real time data in order to combine it with previous analysis for a thorough perspective and complete insights that can help companies and strategists to make fine decisions in an optimal time in order to gain a competitive advantage over the other companies.

So in order to achieve that we collect streams of data coming from twitter using the Twitter Stream API [60], by defining a set of keywords and hashtags that we want to monitor and analyze, we move the data to our system using Apache Flume, a system messaging system that allows fault-tolerant mechanism of communication data, to the processing engine of our framework

that relies on apache spark streaming to perform incremental analysis on the data frame extracted from the social database.

RDDs in spark streaming are added in a data frame that holds information in real time and increment it by appending new data in order to keep the dataset updated, once the data is appended in a time window, analysis are conducted on the corpus of twitter data in order to get the latest insights about the market, and communicate the findings through a dashboard using Apache Kafka as a real time messaging system support.

6.3.6.5 N-Gram for weak signals detection

One of the benefits of using big data analytics in competitive intelligence is the ability of analyzing large volumes of documents and extract implicit knowledge from unstructured textual data. As an example of this analysis we can see that weak signal detection gains a lot of attention because it allows analysts to be proactive and predict the future events and evolution of the market, besides giving them the tools to discover opportunities and threats that can come in form of small changes in the semantic space of a specific domain.

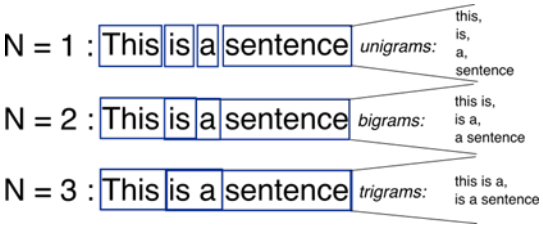


Figure 12: N-Grams Illustration.

By using Apache Spark Machine learning library, we can run the n-gram model on the corpus and create the vocabulary that will help in semantic analysis and tacit knowledge extraction, especially in scientific fields where most of the terms are complex and contain more than one word, the execution time on the cluster allows the use of this model on a large corpus of documents. After the extraction of n-grams and the establishment of domain vocabulary further analysis can be conducted on the resulted dataset like classification or clustering in order to extract latent topics of emerging terms that may hold strategic information in form of weak signals.

6.3.6.6 Graph Creation

6.3.6.6.1 Semantic Network/ Actors Network

Using the capabilities of Apache GraphX, we can create graphs from the corpus using authors field to create authors network, or using the textual field to create semantic networks from the terms used in a specific domain, to conduct a thorough analysis on the evolution of domains and the collaboration between its authors and actors. GraphX allows the creation of a distributed graph on the cluster to optimize the processing and execution time of graph algorithms.

6.3.6.6.2 Semantic Network Analysis

In order to perform analysis on the created graphs we plot the graphs using Gephi or in a dashboard to visualize the networks, and the results of some algorithms after execution of the specified algorithm PageRank for instance.

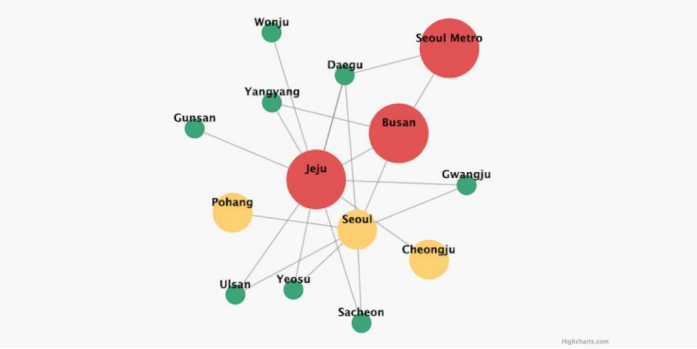


Figure 13: Graph example of network analysis.

6.3.6.6.2.1 PageRank

As an example of analysis we can conduct on a graph we find page rank, which assigns to each node a value of importance in the network, it was originally used to rank web pages in google search engine but it gained more attention when applied in real world and social networks, in the case of our framework we will apply it in authors network to extract important authors and actors of a domain of research.

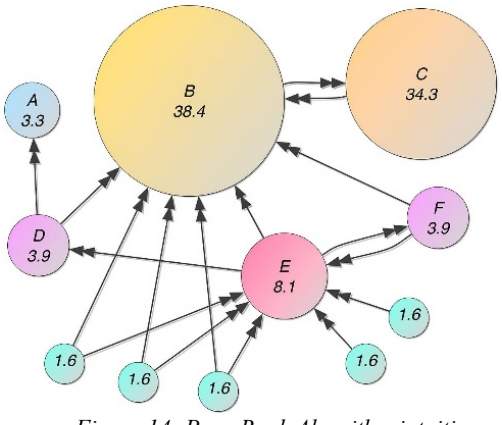


Figure 14: Page Rank Algorithm intuition.

We use Spark GraphX [61] to create graphs and represent our data in networks to simplify interpretation and explanation of its structure and connections, graph computations supported by Spark are powerful and give fast results, we can create graphs from the corpus by using the author field for the creation of authors network which enables us to analyze teams and the relations between research labs and understand the interactions between them. Or the text fields to create a semantic network to show hidden connections between concepts and terms that are not very evident for the users

6.4 Data Visualization

We show all the results of analysis in a dashboard through a web application, that consumes a Spring Boot Rest API connected to the data analytics layer, in order to be delivered and shared with the right stakeholders to interpret them and make decisions based on hidden information mined from all the various raw data collected and stored in the previous layers, and presented finally by multiples charts and graphs. The dashboard provides the following views:

- ✓ Charts for basic statistics showing a summary of data and fields of data, like histograms and evolution charts of number of document per year or per country.
- ✓ Maps to represent geographical data contained in the corpus, many documents contain information about the location of publication, twitter data also may contain geodata about users in order to track the opinion and reaction of clients by location, country or city.
- ✓ Graphs to represent collaboration network from documents authors fields and semantic network from textual field of document

7. CASE STUDY: TECHNOLOGICAL SURVEILLANCE OF "BIOMEDICAL REASEARCH IN MOROCCO"

In this case study, we collected data from various scientific papers and patents and social media. For the first use case we collect data from PubMed Scientific database for research in biomedical field, we extracted 10,797 articles with details of authors, abstract, journal and publication date:

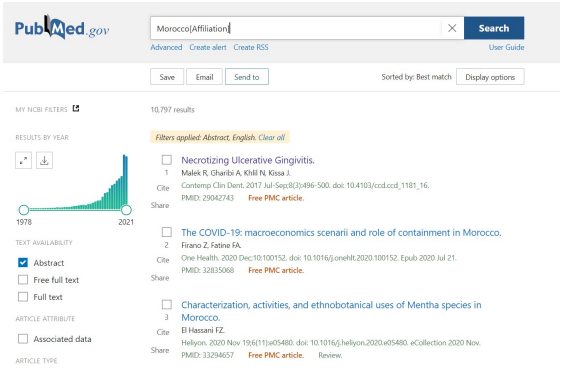


Figure 15: Search Query for Research in morocco.

Figure 16: Example of an Article details to be analyzed.

After collecting and storing the articles in "data lake" we start using the framework to apply the proposed functionalities and conduct a competitive intelligence analysis on the corpus to study the collaborations in Moroccan research and the clusters of interest topics in the field as well as semantic network of concepts and ideas that are present in the domain.

Each article is sored as JSON object in MongoDb, the collected fields will be used in various machine learning and text mining algorithms for analysis purposes, the document-oriented approach will allow us to change the structure of the object according to business need.

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

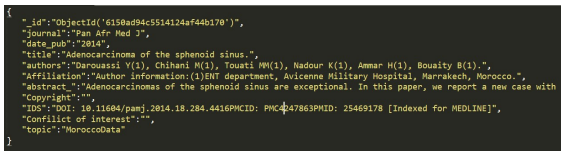


Figure 17: Contingency Matrix plot.

7.1 Matrix Creation

7.1.1 Presence Absence Matrix:

In order to visualize the authors connections and collaborators on the field we create the absence-presence matrix in a distributed manner in spark cluster and we plot the graph to study the situation of research in biomedical researcher's work.

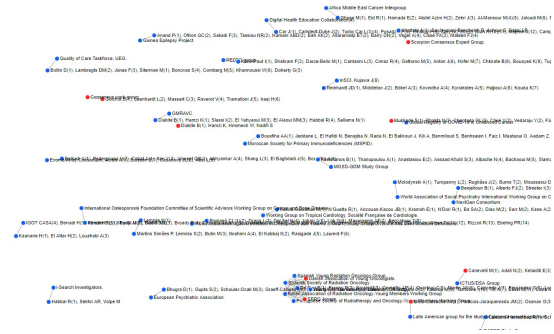


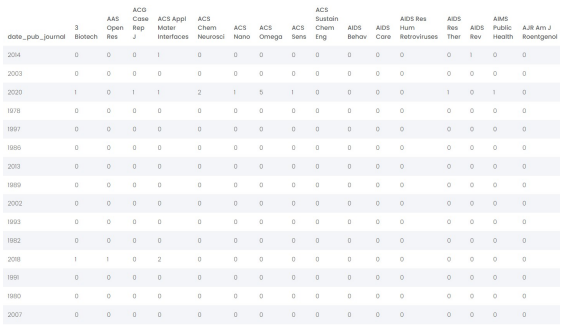
Figure 18: Presence-Absence Matrix plot.

From the graph generated using presence-absence matrix we can run the PageRank algorithm to detect the important nodes of the collaborators.

Author	PageRank Score
Diakite B(1), Hamzi K, Hmimech W, Nadifi S	1.000999
Consensus work-group.	1
Nieto-Calvache AJ(1), Palacios-Jaraquemada JM(2), Osanan G(3), Cortes-Charry R(4), Aryananda RA(5), Bangal VB(6), Staoui A(7)(8), Abbas AM(9), Akaba GO(10), Joshua ZN(11), Vergara Galliadi JM(12), Nieto-Calvache AS(13), Sanin-Blair JE(14), Burgos-Luna JM(1)	1
Goichot B(1), Loenhardt L(2), Massart C(3), Raverot V(4), Tramalloni J(5), Iraqi H(6)	1
Canevelli M(1), Adali N(2), Kelaiditi E(3), Cantet C(4), Ousset PJ(3), Cesari M(4)	1

7.1.2 Contingency Matrix:

We launch the script for creating contingency matrix to visualize a part of it and give analysts a small vision of the generated matrix, this matrix will be used later in machine learning algorithms to cross two variables, in the example below we cross publication date and journal, the visualization of this data will give information about the evolution of activity of each journal and will help us spot the most active one.



date_pub_journal	3 Biotech	AAS Open Res	ACS Case Rep J	ACS Appl Mater Interfaces	ACS Chem Neurosci	ACS Nano	ACS Omega	ACS Sens	ACS Sustain Chem Eng	AIDS Behav	AIDS Care	AIDS Res Hum Retroviruses	AIDS Res Ther	AIDS Rev	AIMS Public Health	AJR Am J Roentgenol
2014	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2020	1	0	1	1	2	1	5	1	0	0	0	0	1	0	1	0
1978	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1997	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1986	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1989	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2002	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1993	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1982	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2018	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0
1991	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 19: Contingency Matrix plot.

7.1.3 Co-Occurrence Matrix:

Through the dashboard we can launch the process for creating co-occurrence matrix which starts the crossing of text content word by word and count the number of occurrence of each pair, we can then sort by frequencies to get the semantic clusters, in our case we have multiple subjects we get a cluster for each domain.

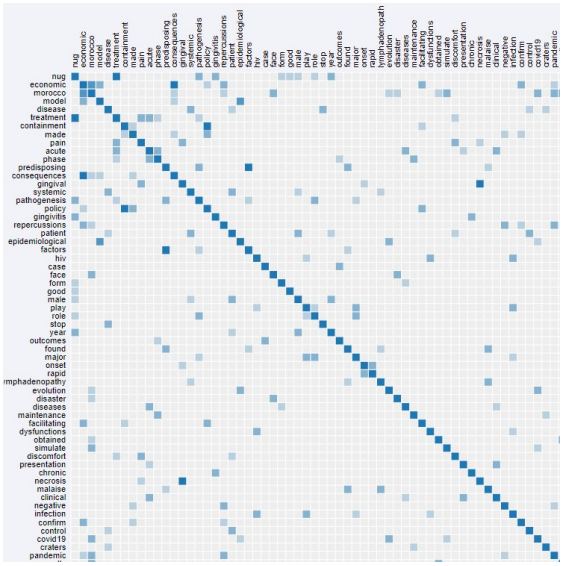


Figure 20: Co-occurrence Matrix plot.

7.2 Basic Statistics

The analysis of basic statics in the collected corpus allow us to visualize bar charts containing valuable information for stakeholders:

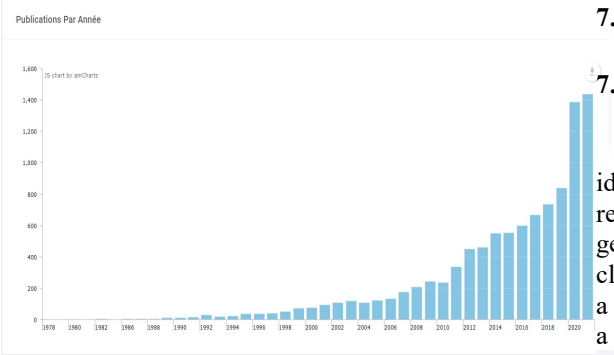


Figure 21: Basic statistics number of publication by year.

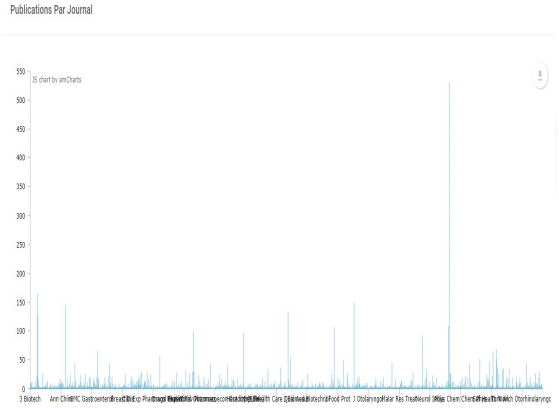


Figure 22: Basic statistics number of publication by journal.

7.3 Principal component analysis:

We apply PCA on the created matrix earlier crossing journal and publication date, this dimensionality reduction technique is crucial for many machine learning algorithms and it will ease access to information regarding the space of data and expose hidden insights.

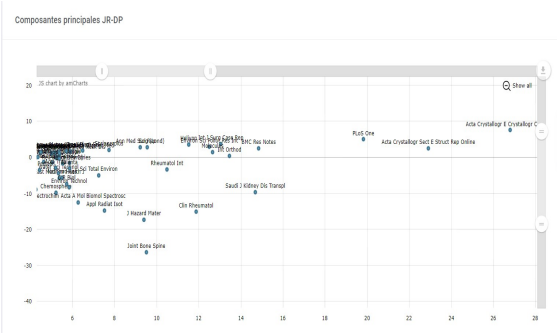


Figure 23: Basic statistics number of publication by journal.

7.4 Semantic Analysis

7.4.1 LDA

Latent dirichelet allocation will help us identify the main topics of the collected corpus in research in biomedical, 5 sets of topics are generated with 10 words defining the semantic cluster of each domain, this algorithm give analysts a thorough general insight about main ideas around a large number of documents.

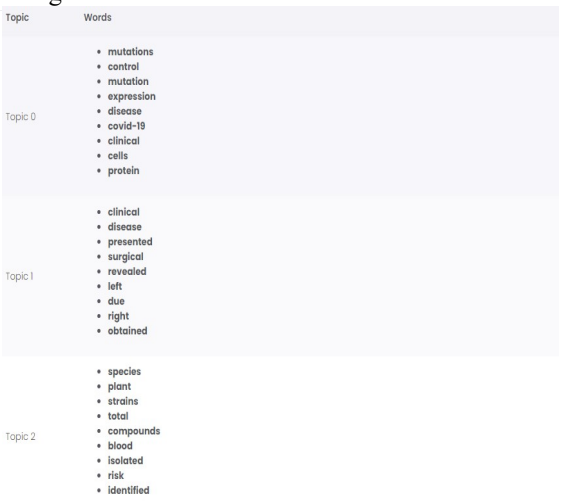


Figure 24: Topics generated by LDA.

7.5 N-grams

From counting number of occurrence of each word we can plot the n-grams (Fig 25,26) and from the most frequent combinations we can spot the interest of Moroccan research in biomedicine, and by sorting the less frequent sets of words we can identify some potential weak signal or opportunities for future research direction.

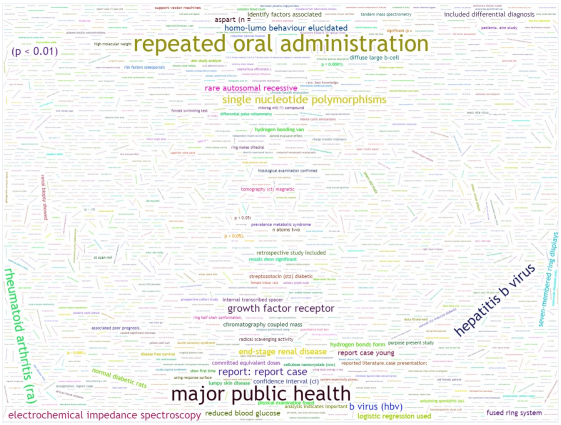


Figure 25: 3-gram example.

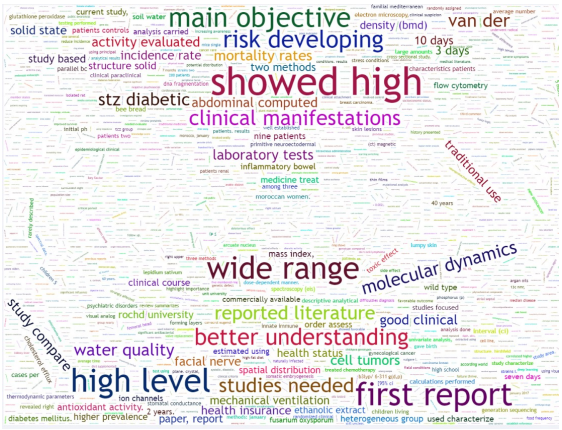


Figure 26: bi-gram example.

7.6 Classification and Clustering

Using the richness of apache spark library we created a dashboard for training clustering and classification models for analysts to run on the corpus for prediction purposes on specific data, which will allow thorough analysis of data and a deep tool for extracting knowledge.

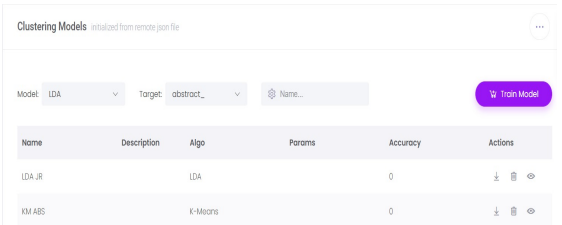


Figure 27: configuring models to train on a specific data field.

8. CONCLUSION AND FUTURE WORK

The increasing demand of automatic decision making systems in companies, guide researchers and practitioners to innovative competitive intelligence systems that can handle the new challenges of big data, in our paper we proposed a big data analytics framework that can serve as analytics layer in competitive intelligence systems relying on Apache Spark as a main technology to support the computational power required for CI analysis, The use of this technology gives the practitioner a wide range of tools and techniques ready to use in hand, in order to conduct multiple analysis, from data collection, textual data analysis using natural language processing, graph networks analysis using the capabilities of GraphX, and streaming data analysis to make near real time based decisions. In addition to the power of Apache Spark we propose an API to format the analysis results and direct them to a dashboard for detailed visualization in order to ease the extraction of valuable insights from the corpus.

The framework did respond to the objective, as we developed in this paper a framework relying on Apache Spark to provide analysts with a toolbox to extract strategic information. The case study above shows the results of our proposed framework and its utility in giving insights to stakeholders. We think that companies can rely on such systems due to the evolution in Artificial Intelligence and Deep Learning models that can assist stakeholders in the decision making process, but there will be always a need for an expert to monitor and tweak the models for a better accuracy of analysis.

In future work we plan on applying deep learning networks on our corpus in order to get more insights from it, and take advantage of the capabilities and features of deep learning models that proved to be successful in a variety of applications, and use a hybrid framework that uses Apache spark alongside DeepLearning4j to perform a thorough analysis and give analysts all the tools in Artificial intelligence field to extract knowledge from raw data.

And after our evaluation of the problem statement and the attempts of proposing a competitive intelligence system analytics layer, we found a lack of evaluation process of such system to validate the efficiency and the accuracy of results and decisions given by such systems, as future research area we can open the question of designing a model of evaluation of competitive intelligence systems and how to automate it and make it part of the system in order to guarantee a continuous evolution of the system by increasing its accuracy.

The framework needs more algorithms for data analysis that can be implemented and integrated in a future work thanks to the modular architecture of our proposed framework, as an example of algorithms to be added we can mention the graph embedding for patent network analysis, and deep learning models for semantic clustering of textual corpus. We can note also the lack of an accurate method for data collection in CI to enhance the quality of the corpus to analyze.

REFRENCES:

[1] Elisabetta Raguseo, Big data technologies: An empirical investigation on their adoption, benefits and risks for companies, International Journal of Information Management. 38, 1 (2018), 187-195.

[2] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," 2017 International Conference on Signal Processing and Communication (ICSPC), 2017, pp. 264-268, doi: 10.1109/CSPC.2017.8305851.

[3] Salloum S.A., Al-Emran M., Monem A.A., Shaalan K. (2018) Using Text Mining Techniques for Extracting Information from Research Articles. In: Shaalan K., Hassanien A., Tolba F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence, vol 740. Springer, Cham. https://doi.org/10.1007/978-3-319-67056-0_18

[4] Răzvan Grigorescu, COLLECTING INFORMATION FROM HUMAN SOURCES FOR COMPETITIVE INTELLIGENCE, Journal: Romanian Intelligence Studies Review 24,2020, 118-139

[5] Fred P. Hoffman, Shelly L. Freyn, The future of competitive intelligence in an AI-enabled world, International Journal of Value Chain Management, 2019, https://doi.org/10.1504/IJVCM.2019.103268

[6] JF Prescott, SH Miller, Proven strategies in competitive intelligence: lessons from the trenches, Book, 2002,1-100

[7] L. Sun, "Knowledge Element-Based Competitive Intelligence Analytics Serving for SWOT Situation Assessment," 2015 8th International Symposium on Computational Intelligence and Design (ISCID), 2015, pp. 576-579, doi: 10.1109/ISCID.2015.147.

[8] Rene Pellissier Tshilidzi E. Nenzhelele, Towards a universal competitive intelligence process model : original research, South African Journal of Information ManagementVol. 15, No. 2, 2013

[9] Vassakis K., Petrakis E., Kopanakis I. (2018) Big Data Analytics: Applications, Prospects and Challenges. In: Skourletopoulos G., Mastorakis G., Mavromoustakis C., Dobre C., Pallis E. (eds) Mobile Big Data. Lecture Notes on Data Engineering and Communications Technologies, vol 10. Springer, Cham. https://doi.org/10.1007/978-3-319-67925-9_1

[10] M. Iqbal, S. H. A. Kazmi, A. Manzoor, A. R. Soomrani, S. H. Butt and K. A. Shaikh, "A study of big data for business growth in SMEs: Opportunities & challenges," 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, pp. 1-7, doi: 10.1109/ICOMET.2018.8346368.

[11] Adnan, K., Akbar, R. An analytical study of information extraction from unstructured and multidimensional big data. J Big Data 6, 91 (2019). https://doi.org/10.1186/s40537-019-0254-8

[12] M. . RTAL and M. . HANOUNE, “Strategic Information Systems and Artificial Intelligence in Business”, IJITAS, vol. 3, no. 2, pp. 78-83, Apr. 2021.

[13] A. El Haddadi, A. El Haddadi, A. Fennan and B. Dousset, "XEW 2.0: Establishment of a new competitive intelligence system for big data analytics”", Journal of Theoretical and Applied Information Technology, vol. 96, no. 16, August 2018.

[14] Clifton L. Smith, David J. Brooks, Chapter 8 - Knowledge Management, Security Science, 2013, Pages 177-198

[15] Tej Adidam, P., Gajre, S. and Kejriwal, S. (2009), "Cross-cultural competitive intelligence strategies", Marketing Intelligence & Planning, Vol. 27 No. 5, pp. 666-680. https://doi.org/10.1108/02634500910977881

[16] Andrew Crane, In the company of spies: When competitive intelligence gathering becomes industrial espionage, Business Horizons Volume 48, Issue 3, 2005, Pages 233-240

[17] Pierrette Bergeron,Christine A. Hiller, Competitive intelligence, Information Science and Technology, 2005, Pages 353-390

[18] Chawinga WD, Chipeta GT. A synergy of knowledge management and competitive intelligence: A key for competitive advantage in small and medium business enterprises. Business Information Review. 2017;34(1):25-36. doi:10.1177/0266382116689171

[19] Anand V. Saurkar, Kedar G. Pathare, Shweta A. Gode, An Overview On Web Scraping Techniques and Tools, International Journal on Future Revolution in Computer Science & Communication Engineering, Vol. 4 No. 4 (2018): April (2018) Issue

[20] Phathutshedzo Nemutanzhela. Tiko Iyamu, A Framework for Enhancing the Information Systems Innovation: Using Competitive Intelligence, The Electronic Journal of Information Systems Evaluation, Vol. 14 No. 2 (2011)

[21] Yafooz W.M.S., Bakar Z.B.A., Fahad S.K.A., M Mithun A. (2020) Business Intelligence Through Big Data Analytics, Data Mining and Machine Learning. In: Sharma N., Chakrabarti A., Balas V. (eds) Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing, vol 1016. Springer, Singapore. https://doi.org/10.1007/978-981-13-9364-8_17

[22] Mannila H. (1996) Methods and problems in data mining. In: Afrati F., Kolaitis P. (eds) Database Theory — ICDT '97. ICDT 1997. Lecture Notes in Computer Science, vol 1186. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-62222-5_35

[23] M. R. Garey, D. S. Johnson, and L. Stockmeyer. 1974. Some simplified NP-complete problems. In Proceedings of the sixth annual ACM symposium on Theory of computing (STOC '74). Association for Computing Machinery, New York, NY, USA, 47–63. DOI:https://doi.org/10.1145/800119.803884

[24] Patrick Mikalef, Maria Boura, George Lekakos, John Krogstie, Big data analytics and firm performance: Findings from a mixed-method approach, Journal of Business Research, 2019, Pages 261-276

[25] Shivam Raj1, Padukere Tejas Upadhya2, Suyash Pathak, A Study of Distributed File Systems

[26] A. B. Patel, M. Birla and U. Nair, "Addressing big data problem using Hadoop and Map Reduce," 2012 Nirma University International Conference on Engineering (NUiCONE), 2012, pp. 1-5, doi: 10.1109/NUICONE.2012.6493198.

[27] Salloum, S., Dautov, R., Chen, X. et al. Big data analytics on Apache Spark. Int J Data Sci Anal 1, 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9

[28] Murad Khan.Muhammad Babar.Syed Hassan. Ahmed Sayed.Chhattan Shah.Kijun Han, Smart city designing and planning based on big data analytics, Sustainable Cities and Society, Volume 35, November 2017, Pages 271-279

[29] Jayanthi Ranjan, Cyril Foropon, Big Data Analytics in Building the Competitive Intelligence of Organizations, International Journal of Information Management, Volume 56, 2021, 102231

[30] Maryse Salles, Philippe Clermont, Bernard Dousset, MEDESIIE: une méthode de conception de systèmes d'intelligence économique, Integrated Design and Manufacturing in Mechanical Engineering, May 2000.

[31] Georg Meyer, Gediminas Adomavicius, Paul E. Johnson, Mohamed Elidrisi, William A. Rush, JoAnn M. Sperl-Hillen, Patrick J. O'Connor, A Machine Learning Approach to Improving Dynamic Decision Making. Information Systems Research 25(2):239-263. https://doi.org/10.1287/isre.2014.0513

[32] Harvey Carruthers, Using PEST analysis to improve business performance, In Practice, Volume31, Issue1, January 2009, Pages 37-39

[33] Bradley Jones,John Sall, JMP statistical discovery software, Volume3, Issue3 May/June 2011 Pages 188-194

[34] J. Wisnowski, F. Castillo, A. Karl, H. Rushing, "Harness the power of JMP®: big data and Social Media for Competitor Analytics," JMP Discovery Conference 2015.

[35] Jesus Silvaa,Lucelys del Carmen Vidal Pacheco, Kevin ParraNegrete, Johana CómbitaNiño, Omar Bonerge Pineda Lezama, NoelVarela, "Design and Development of a Custom System of Technology Surveillance and Competitive Intelligence in SMEs", https://doi.org/10.1016/j.procs.2019.04.177.

[36] Usaquén M.Á.O., García V.H.M., Otálora J.E. (2019) Design of a Competitive Intelligence System for the Meat Sector in Colombia Using Business Intelligence. In: Uden L., Ting IH., Corchado J. (eds) Knowledge Management in Organizations. KMO 2019. Communications in Computer and Information Science, vol 1027. Springer, Cham.

[37] Liandi van den Berg, Ben Coetzee, and Martie Mearns. 2020. Establishing competitive intelligence process elements in sport performance analysis and coaching: A comparative systematic literature review. Int. J. Inf. Manag. 52, C (Jun 2020). DOI:https://doi.org/10.1016/j.ijinfomgt.2020.102071

[38] Jaewoong Choi, Janghyeok Yoon, Jaemin Chung, Byoung-Youl Coh, Jae-Min Lee, Social media analytics and business intelligence research: A systematic review, Information Processing & Management, Volume 57, Issue 6, 2020

[39] Lutz, C.J. and Bodendorf, F. (2020), "Analyzing industry stakeholders using open-source competitive intelligence – a case study in the automotive supply industry", Journal of Enterprise Information Management, Vol. 33

No. 3, pp. 579-599. https://doi.org/10.1108/JEIM-08-2019-0234

[40] Sahin, M., Bisson, C. A Competitive Intelligence Practices Typology in an Airline Company in Turkey. J Knowl Econ 12, 899–922 (2021). https://doi.org/10.1007/s13132-020-00647-z

[41] TODO

[42] Köseoglu, M. A., Mehraliyev, F., Altin, M., & Okumus, F. (Accepted/In press). Competitor intelligence and analysis (CIA) model and online reviews: integrating big data text mining with network analysis for strategic analysis. Tourism Review. https://doi.org/10.1108/TR-10-2019-0406.

[43] Yafooz W.M.S., Bakar Z.B.A., Fahad S.K.A., M Mithun A. (2020) Business Intelligence Through Big Data Analytics, Data Mining and Machine Learning. In: Sharma N., Chakrabarti A., Balas V. (eds) Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing, vol 1016. Springer, Singapore. https://doi.org/10.1007/978-981-13-9364-8_17.

[44] J. Patel, "An Effective and Scalable Data Modeling for Enterprise Big Data Platform," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 2691-2697, doi: 10.1109/BigData47090.2019.9005614.

[45] Monika Arora Uma Kanjilal Dinesh Varshney, Evaluation of information retrieval: precision and recall, International Journal of Indian Culture and Business Management, Print ISSN: 1753-0806 Online ISSN: 1753-0814

[46] Zheng, Jie, Omar S. Harb and Christian J. Stoeckert. “Ontology Driven Data Collection for EuPathDB.” ICBO (2011).

[47] D. Myers and J. W. McGuffee, "Choosing scrapy", Journal of Computing Sciences in Colleges, vol. 31, no. 1, pp. 83-89, 2015.

[48] Vohra D. (2016) Apache Flume. In: Practical Hadoop Ecosystem. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2199-0_6

[49] Pwint Phyu Khine, Zhao Shun Wang, Data lake: a new ideology in big data era, ITM Web Conf. 17 03025 (2018), DOI: 10.1051/itmconf/20181703025

[50] Ali Davoudian, Liu Chen, and Mengchi Liu. 2018. A Survey on NoSQL Stores. ACM Comput. Surv. 51, 2, Article 40 (June 2018), 43 pages. DOI:https://doi.org/10.1145/3158661

[51] E. Shaikh, I. Mohiuddin, Y. Alufaisan and I. Nahvi, "Apache Spark: A Big Data Processing Engine," 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), 2019, pp. 1-6, doi: 10.1109/MENACOMM46666.2019.8988541.

[52] Luu H. (2018) Resilient Distributed Datasets. In: Beginning Apache Spark 2. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-3579-9_3.

[53] Martin Grohe. 2020. Word2vec, node2vec, graph2vec, X2vec: Towards a Theory of Vector Embeddings of Structured Data. In Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS'20). Association for Computing Machinery, New York, NY, USA, 1–16. DOI:https://doi.org/10.1145/3375395.3387641

[54] I. Sorić, D. Dinjar, M. Štajcer and D. Oreščanin, "Efficient social network analysis in big data architectures," 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2017, pp. 1397-1400, doi: 10.23919/MIPRO.2017.7973640.

[55] Michael Armbrust, Tathagata Das, Joseph Torres, Burak Yavuz, Shixiong Zhu, Reynold Xin, Ali Ghodsi, Ion Stoica, and Matei Zaharia. 2018. Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark. In Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18). Association for Computing Machinery, New York, NY, USA, 601–613. DOI:https://doi.org/10.1145/3183713.3190664

[56] Ilhème Ghalamallah, Proposition d'un modèle d'analyse exploratoire multidimensionnelle dans un contexte d'intelligence économique, Thèse de doctorat dirigée par Dousset, Bernard Informatique Toulouse 3 2009.

[57] Wall M.E., Rechtsteiner A., Rocha L.M. (2003) Singular Value Decomposition and Principal Component Analysis. In: Berrar D.P., Dubitzky W., Granzow M. (eds) A Practical Approach to Microarray Data Analysis. Springer, Boston, MA. https://doi.org/10.1007/0-306-47815-3_5

[58] Thomas K Landauer, Peter W. Foltz & Darrell Laham (1998) An introduction to latent semantic analysis, Discourse Processes, 25:2-3, 259-284, DOI: 10.1080/01638539809545028

[59] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, null (3/1/2003), 993–1022.

[60] Michal Podhoranyi and Lukas Vojacek. 2019. Social Media Data Processing Infrastructure by Using Apache Spark Big Data Platform: Twitter

Data Analysis. In Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things (CCIOT 2019). Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi.org/10.1145/3361821.3361825

[61] Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, and Ion Stoica. 2013. GraphX: a resilient distributed graph system on Spark. In First International Workshop on Graph Data Management Experiences and Systems (GRADES '13). Association for Computing Machinery, New York, NY, USA, Article 2, 1–6. DOI:https://doi.org/10.1145/2484425.2484427

Journal of Theoretical and Applied Information Technology
15th January 2022. Vol.100. No 1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

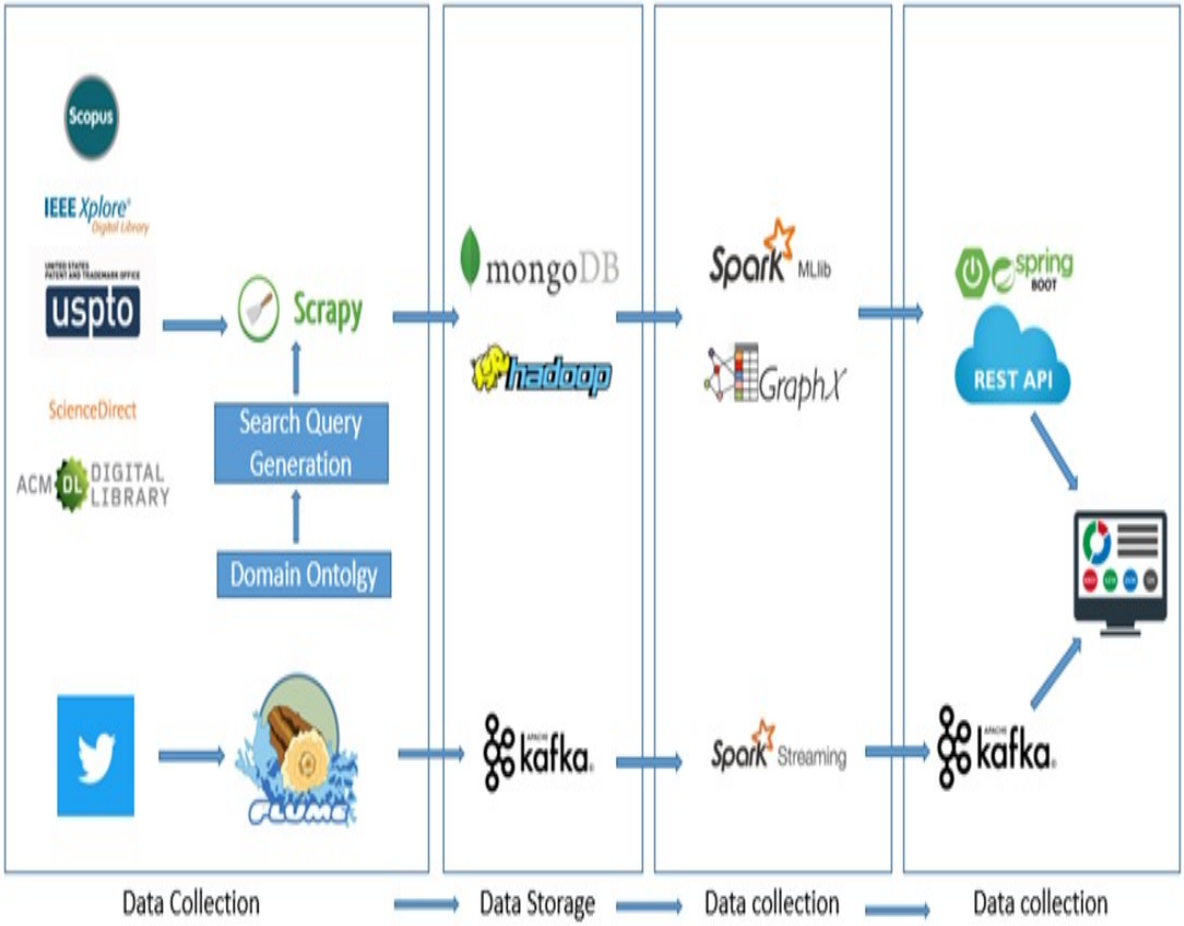


Figure 1: Proposed Framework Architecture.