UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

ESCOLA DE INFORMÁTICA APLICADA

Context-Aware Process Mining:
Using External Contextual Web Scraped Data to Enrich the Event Log

Fernando Cardoso Durier da Silva

Advisor

Kate Cerqueira Revoredo

Flávia Maria Santoro

RIO DE JANEIRO, RJ – BRAZIL

DECEMBER 2017

Cardoso Durier da Silva, Fernando
CC268 Context-Aware Process Mining: Using External Contextual Web Scraped Data to Enrich the Event Log
/ Fernando Cardoso Durier da Silva. -- Rio de Janeiro, 2017.
54

Orientadora: Kate Cerqueira Revoredo.
Coorientadora: Flávia Maria Santoro. Trabalho de Conclusão de Curso (Graduação) - Universidade Federal do Estado do Rio de Janeiro, Graduação em Sistemas de Informação, 2017.

1. Business Process Management. 2. Process Monitoring. 3. Sentiment Analysis. 4. Big Data. 5. Mineração de Dados. I. Cerqueira Revoredo, Kate, orient. II. Maria Santoro, Flávia, coorient. III. Título.

Context-Aware Process Mining:
Using External Contextual Web Scraped Data to Enrich the Event Log

Fernando Cardoso Durier da Silva

Undergraduation project presented to the Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) in order to obtain the title of Systems of Information Bachelor.

Aproved by:

_____

Flávia Maria Santoro (UNIRIO)

_____

Kate Cerqueira Revoredo (UNIRIO)

_____

Fernanda Araújo Baião (UNIRIO)

RIO DE JANEIRO, RJ – BRAZIL.
DECEMBER 2017

Acknowledgements

I would like to thank God for granting me such opportunities, paths and good conditions.

I would like to thank my parents Hilda Cardoso Durier da Silva and Fernando Durier Sarmento da Silva for always being supportive, lovely and kind to me, when making me feel better, always saying not to give up, taking care of me and showing how is to be a great parent. Without them, I would not be writing this text.

I would like to thank my grandparents, Alzira Pinto Cardoso and José Cardoso da Motta, as well that have always been supportive stimulating my intellect and personal growth. My grandmother was always gentle, and sweet with me during all my graduation period, she was and always will be very special to me. My grandfather passed way in 2008, but when alive always valued me for intellect and character and I still can feel his support from wherever he is.

I would like to thank my advisors, as they were always by my side, studying, showing me new aspects of the research subject, guiding me through my academic ways and solving doubts.

I would like to thank my teachers at UNIRIO's Applied Informatics Department, as they were exceptional doing their honored job, always being present, supportive and kind to me. It is thanks to them as well that I am ready for this new step in my life.

I would like to thank the director, coordinator and secretary of UNIRIO's Applied Informatics Department for their efficiency and impeccable service always providing us with information, needed documentation and being always ready to help.

I would like to thank IBM, as it had me there for my internship and allowed to apply all my knowledge acquired during all this four years and develop even more my skills.

I would like to thank Kate Cerqueira Revoredo and Flávia Maria Santoro for awake and stimulate my research desires, during our scientific scholarship period. And I would like to thank Bruno Lima Cardoso for believing in my capabilities and my research desire, when he invited me to work with him at IBM Research Brazil, and for doing all he could do to get me there and hone my research and development skills.

Abstract

Predictive monitoring of business process is one of the main challenges in process mining. It makes possible to perform tasks such as performance check, compliance, prediction and improvement by analyzing and discovering patterns in the event logs generated from Information Systems. However, usually only information on the activities being executed are considered for these analyses and sometimes this information is not sufficient to describe the context in which the process instance is running. On the other hand, contextual information has been shown to bring benefits in many areas of Business Process Management, since it provides extra knowledge related to goal, organization and environment of the business process. Motivated by that, in this work we propose an enrichment/enhancement of the event log with the sentiment about news that where brought up in the same day the process instance was running. We then analyze the impact of this context to the process monitoring. To evaluate our proposal a main dataset was used, the Business Process Intelligence Challenge 2013, BPIC2013, that contains the IT Incidents Support Event Log given by Volvo, and a set of contextual data, the News Web Scraped from the New York Times. Then we will enrich the BPIC2013 dataset with the contextual one, using the country as a joint attribute, so we can see if and how the process was affected by the context expressed by its country's socioeconomic context.

Keywords: Business Process Management, Data Mining, Sentiment Analysis, Process Mining, Context Mining.

Index

Table Index

Image Index

1 Introduction

1.1 Motivation

Predictive monitoring of business process is one of the main challenges in process mining. It makes possible tasks such as performance check, compliance, prediction and improvement by analyzing and discovering patterns in the event logs generated from Information Systems. However, usually only information on the activities being executed is considered for this analysis[29][30][31][32].

We argue that, this information is not sufficient to describe the context in which the processes instance is running, due to the fact that some of those internal deviances can be caused by events that are external to the organization hosting the process or may be caused by altered psychological state of peopleware responsible for the business process execution.

On the other hand, external contextual information has been shown to bring benefits in many areas[20] of Business Process Management, since it provides extra knowledge related to goal, organization and environment of the business process[19].

Motivated by that, in this work, we propose a new approach in observing these external patterns that the systems were not considering during the process execution and establishing relations between what was described and visible to it and what was

unwary and hidden from it, thus making an enrichment/enhancement of the process event log with this external contextual information.

We also argue that sentiment about news that were brought up in the same day that the process instance was running, may influence against the ones who execute the process.

We can imagine that an employee of a multinational enterprise just woke up and got his/her newspaper and read about a problem that is affecting the enterprise matrix and could cause a serious loss in its revenue. Knowing that, the employee might perform its fundamental activities motivated to help against the crisis or he/she might get demotivated and give up to the bad news. So affecting the business process in which he/she is immersed into.

Given that scenario, it could be relevant to consider that news an external contextual element, in fact, to be more precise we need to consider the sentiment of that news an external contextual element. Therefore, we propose to analyze the impact of this external contextual information to the process monitoring.

1.2 Objectives

The goal of this project is to discuss how external context may affect business process' activities flow by correlating the context of the process environment with process instances' results. Focusing into this new approach of using the web news as source of context.

1.3 Text Organization

The present work is structured in chapters and, aside this introduction, it will be developed as follows:

- Chapter II: presents the main concepts about the areas related to this work: Business Process, Business Process Management, Data Mining, Web Scraping, Sentiment Analysis, Process Mining and Context and Context Mining.

- Chapter III: describes how we could acquire and manipulate such massive data, that half is structured and other half semi-structured; which were the needed technologies, programs, data structures and tools;

- Chapter IV: describes the datasets used, from where data they come, which were the relevant aspects of it, why we chose them and visualize how the characteristics of data.

- Chapter V: presents which algorithms, techniques and data structures were used to perform knowledge discovery over those data. We shall see the results of the execution of them and perform analysis over them.

- Chapter VI: provides the conclusions taken from the analysis we did, presents the main difficulties and problems encountered and speculate about future work.

2 Preliminaries

2.1 Business Process Management

A Process is a set of related actions that transform inputs into products or services[1].

There are processes in all areas, however, we are considering business, so, we are considering business processes of an organization or an enterprise, in which feedstock or human resources get in as input, and what we get at the end of the process is a product (a car, a toy, a computer, …) or a service (maintenance, credit concession, ...). Thanks to them it is possible to control the product making or service performance, audit and fasten the day-by-day activities of the organization, predict failures, discover opportunities and much more.

[1] LAUDON, Kenneth C.; LAUDON, Jane P. **Management information system**. Pearson Education India, 2016.

Data related to processes may be represented and stored in such many ways. They can be stored in relational databases, non-relational databases or common data repositories. The representation format may vary as well, since they can be stored like computer files (like XML, JSONs, TXTs, etc.) or like physical formats (like diagrams, set of instructions, etc.).

The main standard notation to represent business processes is the Business Process Modeling and Notation (BPMN). Using this notation, we can represent richer forms of activity composition, message exchange, data objects, concurrency of processes etc.

According to VAN DER AALST et al (2013), "Business Process Management (BPM) includes methods, techniques, and tools to support the design, enactment, management, and analysis of operational business processes".

BPM uses a set of tools to visualize and monitor business processes workflows to perceive bottlenecks, critical events to the processes, check its alignment with the BPMN model, optimization, and discovery of new workflows.

BPM can also be used in other areas than business, like when used in software engineering as service oriented architectures, in order to model sequence of functionalities, and establish communication patterns.

2.2 Data Mining

To help management activities of all kinds it is necessary to accumulate big volumes of information and apply them to generate knowledge, so, finally get the wisdom that is so important to take actions, elaborate new strategies, and much more. Since, nowadays we have an explosion of available data because the users are not only consuming but producing data as well, we can use this in favor of improving our decision making and decreasing uncertainty. So, following this motivation, Data Mining area surged in order to handle that big volume of data and to extract the most information and knowledge as possible.

The data mining process (workflow) consists in gathering the needed data, preprocessing it to cleanse our datasets, transforming the data, mining for patterns, understanding and evaluating the results, and finally generating a new knowledge.

In this work we will use some of its techniques with the main focus on how to process the natural language of the news' text and extracting the sentiment expressed in such news.

2.3 **Web Scraping**

A Web Scraper is a program that will access the webpages in internet and collect data from part(s) of the HTML that are relevant. As said by VARGIU, URRU et al (2012), "Web scraping (also called Web harvesting or Web data extraction) is a software technique aimed at extracting information from websites".

It is possible to perceive the web scraping as the initial step of the data mining process as acquiring data, but also, the post processing of enhance the already obtained dataset in order to make better analysis and pattern discovery. To illustrate better the web scraping process we can follow this description below.

For the making of this project, I developed a set of web scraper/preprocessor scripts to gather the needed info, developed using Node.js[2], a javascript framework designed to develop network applications in an easier and faster way.

In order to access the news webpage and return its HTML data, the npm module called request was necessary. This module was designed to make the HTTP request process easier[3]. But returning the pure HTML data is not enough, as the needed information is hidden within it.

So, in order to extract the news' content it was necessary an HTML content explorer, in this case, the one used was cheerio. Cheerio parses the HTML data, exploring it's CSS, HTML tags and other structures[4].

The last one npm module used to persist the web scraped data in a staging area before the sentiment analysis algorithm could be used, was the npm fs, most known as npm file

[2] Node.js about - https://nodejs.org/en/about/
[3] npm request - https://www.npmjs.com/package/request
[4] npm cheerio - https://www.npmjs.com/package/cheerio

system module. Npm fs was designed to make file operation simpler, leveraging the local environment awareness from the developer to its API[5].

As we could see above, the web scraping process follows the steps of acquiring data, preprocessing (parsing of the webpage) and storing in a database for future analysis.

Although the news dataset is our base to infer context, we needed another technique that would help us analyze the sentimental context from people impression of the news, but, automatically.

2.4 Sentiment Analysis

According to Liu et al (2012), “Sentiment analysis, also called opinion mining, is the field of study that analyze people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes.”.

Benevuto, Ribeiro and Araujo et al (2015) affirmed that "Opinions from social media, if adequately retrieved and analyzed, allow not only to comprehend and explain diverse social complex phenomena, but to predict them".

Pang et al (2008) affirmed that “An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis,

[5] npm file-system - https://www.npmjs.com/package/file-system

which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object."

Benevenuto, Ribeiro and Araújo et al (2017) describe the two main techniques used to extract sentiments in texts, as it follows

Supervised Technique: demands a training step, where a set of texts will be previously classified and used as the reference for next iterations.

This technique contains the following four steps:

1. Tagged data acquisition, in order to perform train (making the algorithm aware of the dataset patterns) and test (checking the accuracy and precision of the algorithm).
2. Feature and Characteristics definition. Allowing distinction between data.
3. Computational model training with a machine learning algorithm (meaning that in this step we prepare the program with training data, in order to analyze different data in future iterations, like humans when study for an exam).
4. Model application (in this step the program will apply its knowledge upon real dataset, much like a human doing an exam).

Non-Supervised Technique: doesn't require model training. In general, are based upon sentiment lexical evaluation involving a polarity calculus of a text, considering the semantic orientation of the words present in it. Normally, utilizes a big corpus of terms,

which each is associated with a sentiment score that has a quantitative and qualitative meaning, in other words, a scale with sentimental degrees.

For this project we chose the Non-Supervised approach, as we had already an adequate library that uses a pre-tagged and precalculated list of words as its source. Basically, it is process of sentiment analysis is, receiving the textual data, analysing each word present in the text and consulting its score on the library's list, calculate the text sentiment and outputting a result of it.

Based on this, by using Sentiment Analysis it is possible to comprehend the social context from the news and explain part of the social context of the population which is referred by it. It is possible to assume that since the business processes need people to work, it is possible that the social context of a given country may affect its population, actors of the process, and finally an actor can affect the process flow depending on how it performs its activities (with more or less care, attention, dedication and/or with or without subjective worries). For example, when a country is passing by a crisis, its population is likely to have its routine altered and eventually it might change the way they handle their work tasks.

2.5 Process Mining

When applying data mining techniques into business process event logs, we are doing process mining. The idea of process mining is to discover, monitor and improve real processes by extracting knowledge from event logs available in today's systems. Process mining includes process discovery, automated or not, conformance checking, social network/organizational mining, automated construction of simulated models, model extension, model repair, case prediction, and history-based recommendations[14].

In our proposal we are focusing on Process Monitoring an area of Process Mining that can be stated by Kung et al (2005)[27] as “Process Monitoring: To check a situation carefully in order to discover something about it. The aspects monitored (i.e. the metrics taken) may change quite often. Normally, data gathered is not stored for long periods of time.”.

2.6 Context

According to KISELEVA et al.(2013), “Context often has a significant impact on the way humans (or machines) act and on how they interpret things; furthermore, a change in context causes a transformation in the experience that is going to be lived.”.

Taking a more aligned approach to our work, DA CUNHA MATTOS, SANTORO and REVOREDO et al.(2014) said that ”Context can be defined as a complex description of the knowledge shared on physical, social, historical and other circumstances where actions or events happen. All this knowledge is not a part of the action or the event, but will constrain the execution of the action or the interpretation of the event.”. In our

approach we considered that this constraint that alter execution and perception of the events would be the socio-economic one, instantiated in the web news, of the countries that host the process' instances during its execution period.

2.7 Context Mining

In this section we discuss about the main approaches of context discovery/mapping and which of those we chose.

According to KISELEVA et al. (2013) there are three general strategies to discover contextual categories.

The first one is discovering from an existing feature set, that consists in mapping and exploring a predefined context known from domain experts.

The second one being discovering from hidden context, that consists in relying upon automatic pattern understanding methods for example clustering, subgroup discovery, mixture models, or more sophisticated techniques.

And the third one, context discovery from external sources that relies on exploring factors like weather, populational indicators and etc

.

In this project we will be working with the second and third strategy since we are resorting to use external sources (web news), and using automatic pattern understanding when analyzing the sentiments of the news and grouping those by timestamp and country location.

2.8 Node.js

JavaScript is a script interpreted language, that has resemblance with C, C++ and Java and have the same capabilities of Object Orienting, and has some structures that are more aligned with web-services, Flanagan et al (2006). This language was chosen due to its proximity with the main domain of origin of our datasets, and also for the ease to use. In order to not work just with the pure JavaScript language and to increase the language already existent potentialities, it was decided to work with Node.js, that is a JavaScript framework that supports package management, server creation, deployment, parallel computation and many more features. As highlighted by JUNIOR et al (2012), "Node.js is a platform which objective is the easy building of fast and scalable network applications. Due to this, it uses a model based on events and non-blocking I/O".

3 Related Works

We consider related works as works that use external contextual elements for process monitoring. This area of research and work, process predictive monitoring, is very recent and has its focus on other fronts like modelling and linking between participants of the process.

In this work[28], the author elaborated a process much similar to the one we propose. The main difference is that its process first steps require a domain specialist to curate the relevant external contextual data and the last steps of it were for recommendation and direct interference to the main flow responding to the deviations brought by external context, its external context focused on unemployment taxes and inflation, while our contextual elements were more generic and abrangent.
In this work we can observe as well that the event log was a real one took from a set of other processes' logs, had a very strong basis on information science and did recommendations after its analysis.

During my researches I found other related works in [29][30][31][32], but, those works focused on internal contextual elements such as succession of events, actors and networks of actors or simply on timestamp. So, they were related to the predictive monitoring. Meanwhile, the work [28] was more related to this project, it had a different focus.

4 The Proposal

4.1 The Proposal

In this work, we investigate how context may affect business process' activities flow by correlating the context of the process environment with process instances' results in a specific scenario. Thus, we defined an experiment to gather evidences and allow some conclusions about the topic. The following steps are proposed for the experiment depicted in Figure 1.

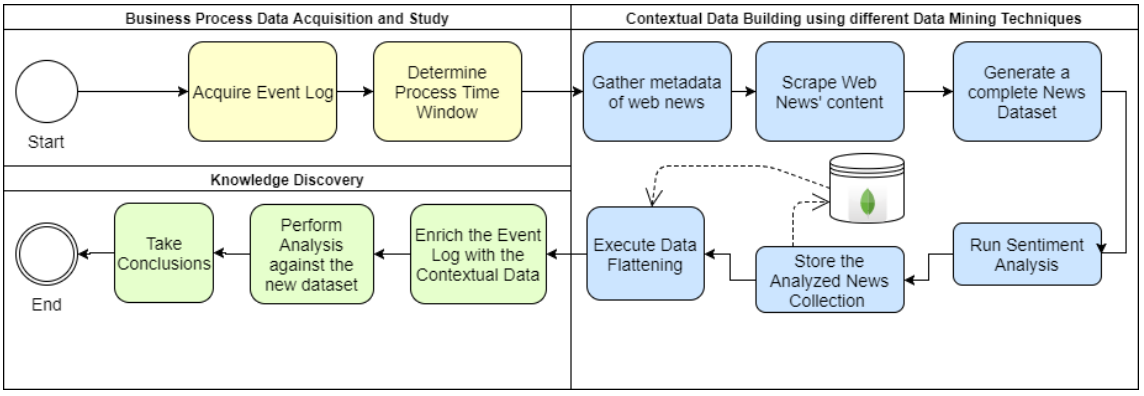

Figure 1. Experiments' steps

First, we needed to define a process event log to compare how the external context may affect that process in that given moment. We chose the Business Process Intelligence Challenge 2013, BPIC2013[6]. It is an IT Incident Resolution Process event log from

[6] BPIC2013 - http://www.win.tue.nl/bpi/doku.php?id=2013:challenge

Volvo, a multinational enterprise that may be influenced by the tendencies and economics of the countries where it has its branch offices located.

After the event log acquirement, we need to define the external source of context to enrich the event log and enhance our analysis. Due to its reliability, impartiality and well-known API, we chose the The New York Times newspaper as our external source, more specifically it's News Archive API as the channel to get the needed web news metadata, and its news' web pages as sources from where we will extract the context.
To extract the context, we use sentiment analysis techniques to extract sentiment of each news. We can use this sentiment impression as a context indicator of how well or how bad was the situation of that given country when the process was executed.

After the datasets are preprocessed and annotated (in the contextual one case), we will proceed to the log enrichment step, where will guarantee data compatibility between the dataset's structures, as we are joining a structured plain set of registries and a semi-structured hierarchical set of documents. We will reach this compatibility through flattening of those hierarchical attributes in the documents from the contextual dataset, then yet using the same dataset we will perform an aggregation operation (average of sentiment scores), grouped by country, timestamp and news category.

Having all these postprocessing we can finally enrich the process event log with the external context data and perform analysis, tests and inferences. To enrich the event log, we needed to aggregate to each event the context of its country during the same timestamp it was running. So, we did a join between the event log dataset and the

contextual data on the country location of the process and the same timestamp from where and when the average sentiment was calculated from. This joining was done by the well-known join from relational databases, in fact it was a left join from the event log dataset to the news summarized dataset. So, we will ensure that all data from the event log appear even if it is null, as its information is more important to us, however, since there was a news for each day of the event log this left join could be seen like a precaution that did not become real, but can be generalized for future works.

Finally, after this data mining task of dataset enrichment, we can evaluate the results through graphic visualization and through algorithm played against the enriched dataset and take conclusions..

In order to execute the experiments, it was necessary to define the adequate environment that could address all requirements of the above-mentioned areas. So in this section, we describe the elements of implementation needed. Here it is possible to see which was the language, framework, data interchange format, database and implementation of the mapped process to fulfill our experiment and more details about the data gathering.

4.1 The Web Framework Node.js

It was the framework used to implement our proposal, due to its ease of use, very effective package manager and its affinity with web applications as mentioned in the second chapter.

Through it we could implement the web scraping functionality to extract the contextual data from its web source, the sentiment analysis step using its well-known module, and to process, preprocess and store our data into our database.

4.1 The JSON format

Although the data used in the experiments is originally represented in HTML, after the web scraper processes data from the web pages, it retrieves the main result in the JSON format, that stands for JavaScript Object Notation. It is a key/value structure much common in web development.

The JSON is a well-known and accepted format. According to BASSET et al(2015), “JSON is a data interchange format that many systems have agreed on using for communicating data“. JSON is based on JavaScript Literals but it is an independent language, and can be used by any language that has an adequate parser to it.

4.1 The Sentiment Analysis module

The library chosen to perform sentiment analysis was the npm sentiment, that is a Node.js[7] module that uses the AFINN-165[8] wordlist and the Emoji Sentiment Ranking[9] to perform sentiment analysis on arbitrary blocks of input text. Basically, the analysis is started when receiving the response from the web scraper. The module analyses the news text and produces as outputs the calculated sentiment score for each news.

Although there are many web services for sentiment analysis, this option was chosen because it is free of charge and no limited number of calls (different from the common web services as they restrict to thousands calls freely and paid calls after), and also its reliable analysis background in AFINN-165 wordlist. As a matter of fact there is a disclaimer footnote in the npm module official page showing its accuracy that varies from 70% up.

4.1 MongoDB

As most of our data is semi structured, and JSONs, the most adequate database system to support it would be a non-relational/non-conventional one, because of its schema flexibility and its exponential growth through the gathering step of the data mining pipeline.

---

[7] NPM module sentiment - https://www.npmjs.com/package/sentiment

[8] AFINN is a list of English words rated for valence with an integer between minus five (negative) and plus five (positive). The words have been manually labeled by Finn Årup Nielsen in 2009-2011. The file is tab-separated.

[9] Novak et al (2015).

Another interesting thing about our data, that is JSON formatted, is that it can be individually stored as file documents, allowing not only their well-known portability, but also allowing many file system resourceful operations to be applied to them such as read, write, append, versioning, replication, among others.

The news data that we took were in fact the news HTML document body content annotated in some way that, besides the textual content of the news, we could also store metadata of it like URL, headline, publication date, news desk (from which sector that news was written), words counted, sentiment analyzed and country that is affected by it.

As our main dataset takes this particular format, so the more adequate choice to store such would be a document-oriented database. Inspired by Lotus Notes, document databases were designed to manage and store documents. These documents are encoded in a standard data exchange format such as XML, JSON or BSON (Binary JSON).

If we would think in a key-value store the value column in document databases contains semi-structured data - specifically attribute name/value pairs. A single column can house hundreds of such attributes, and the number and type of attributes recorded can vary from row to row. Also, unlike simple key-value stores, both keys and values are fully searchable in document databases.[10]

[10] NoSQL Databases definitions and concepts - MONIRUZZAMAN, A. B. M.; HOSSAIN, Syed Akhter. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. **arXiv preprint arXiv:1307.0191**, 2013.

So, looking upon this scenario we chose MongoDB, a well-known and trusted document-oriented database. With it and all its type of database qualities we can store, query and aggregate complex hierarchical relationships, handle gigabytes of data and work with a database that is in sync with the way developers nowadays think (Object-Oriented Thinking)[12].

4.1 The Statistical Analytical Tool

As the data gathered has more than 500.000 rows and 20 columns to process and analyze, it is necessary to use a third-party tool to perform the adequate statistical analysis and uncovering of patterns and rules present in the datasets, so we will be able to generate a new knowledge. Due to its trusted ability to handle such data and perform the needed statistical and KDD operations, we chose R.

R is a language and environment for statistical computing and graphics. It is a GNU project. It provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. One of R's strength is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Also, it is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS[16].

5 Study Case

The dataset explored was the business process event log, the BPI Challenge 2013, BPIC2013, (a challenge of process mining that uses real life event logs as contest material/proof. The participants shall answer a set of questions brought by the event log's owner, using the results got by their application of process mining techniques and tools) proposed dataset. It is a log took from a system that implements the IT Support process of Volvo. Its primary goal is to restore normal service operation (normal service operation is defined within Service Level Agreement -SLA) as quickly as possible and by that ensuring that the best possible levels of service quality and availability are maintained. Services that cannot be handled by the service desk or expert helpdesk should be escalated to Second Line and/or Third Line teams. After solving or implementing a work around the incident is closed.

The attributes of this dataset are:

- **First Line** - is the common name for service desk and expert help desk.
- **Service Desk** - normally the single point of contact to advise, guide and assist in rapid restoration of normal services to its customers and users. It can be divided into Front Desk, Offline Desk and Desk Side Support where Offline Desk can be

both local and global teams.

- **Expert Help Desk** - Expert Help Desk is the entry point for end user support on specific services e.g. application support. Expert Helpdesk is an additional service agreed upon with the customer. It can be divided into: Front Desk and Offline Desk where Offline Desk can be both local and global teams.
- **Second line** - is taking care of incidents that cannot be resolved within First Line. In this case the second line can be divided into Org Line C or a support team working with Org Line A2.
- **Third line** - is the experts in their products area, could be a support team within Org Line C or a support team working with Org Line A2. The third line is taking care of Incidents that cannot be resolved within the Second Line function.
- **Impact** - is a measure of the business critically of an Incident often equal to the extent to which an Incident leads to degradation of agreed service levels. Impact is often measured by the number of people or systems affected. Criteria for assigning Impact level should be set up in the SLA's and here are those descriptions:
    - Major - very high visibility to the customer and/or have a defined major impact on a given service. They can be defined as incidents that may disrupt plant production and result in lost units of production; Incidents that impact a particular system that has been identified as critical or is a new occurrence of a repeated Impact Level High Incident over a short period of time; Incidents that impact site, several dealers, large workgroup or key business process for which the incident may result in: Vehicles/Product cannot be designed; Vehicles/Product cannot be

delivered or ordered; Service to customer cannot be performed; Cash flow is negatively impacted; Customer's revenue or public Image affected. Resuming: Has to be resolved with high attention from all parties involved.

    - High - when the service is regarded as unavailable by the user/customer and affects an important system/service or if access is lost which incapacitates a site or threatens to compromise other sites. Examples of high impact incidents are Server incidents which may result in loss of service to customer; wide/local area network incidents; lost production units due to an Application or Infrastructure incident.
    - Medium - are those ones that have limited visibility to the users and/or have limited impact to a given service such that it may be viewed as operational, but in a degraded mode, by the user. They can be defined as: A small share of the Customer base is affected; An incident which does not negatively impact the Customer ability to meet its service levels; An incident which will not result in Volvo IT not being able to meet their service level commitments; Incidents that occur outside the scope of IT services but with an impact on IT services such as power outage;
    - Low - those Incidents that have low visibility to the Customer and/or have minor impact to a given service and do not limit the User in functionality. Even if it happens, the user can still achieve full functionality and normal performance as long as the circumvention procedure is followed.

- **Urgency** - is about the necessary speed of solving an Incident of a certain impact

to the user. Which the values may be:

    - High - The problem has a very high influence on the Users work at the time of reporting the Incident.
    - Medium - The problem has influence on the Users work at the time of reporting the Incident. Are incidents that cause delays on the process, but, they do have turnarounds.
    - Low - The problem has an low influence on the Users work at the time of reporting the Incident.

- **Priority** - The priority of an Incident is determined by the Impact on the business and the Urgency for which a Work Around or Solution is needed. The Impact level, agreed with the customer and embodied in an SLA, together with the Urgency level, agreed with the User, is translated to a Priority level. The Priority level is for internal use and guides the support personnel in solving Incidents in the correct order of Impact and Urgency for Volvo IT customers. If the Incident is not handled within the specified time frame (according to SLA), the priority of the Incident shall be increased. However an Incident can never rise automatically between the different levels of Impact.

Although Volvo has its own definitions, their logs come from a system called VINST, which generates logs with the following attribute nomenclature and meaning:

- **Problem or SR_number**: unique ticket number for a problem or incident;
- **Problem change Date+time**: moment when the status of the problem changed;
- **Problem Status**: Queued, accepted, completed and closed;
- **Problem sub-status**: Assigned, awaiting assignment, cancelled, closed, in

progress, wait, unmatched(?);

- **Impact**: level of impact the problem creates for the customer (major, high, medium, low);
- **Organization (problem involved org line 3)**: the business area of the user reporting the problem to the helpdesk
- **Function Division**: The IT organization is divided into functions (mostly technology wise);
- **ST** (support team): the actual team that will try to solve the problem;
- **country_code**: the location that takes the ownership of the support team;
- **action_owner**: the person that works in a support team that is working with the incident. An action owner can transfer an incident or problem to another action owner within the same support team, or he can escalate or return a problem to another ST (Support Team);

The contextual dataset used, is a mass of news data from years 2006~2013, the same timebox when the process described above occurred, produced by The New York Times. It was gathered from its News Archive API, that is a RESTful HTTP POST callable endpoint that by passing the private development key, the year and the month, it returns in response a set of metadatas (headline, abstract, author, news desk, publication time etc.) of all news from the given time period asked. The choice of The New York Times was specifically due to its reliable, interesting, and global range news. After the steps of web scraping, sentiment analysis and country detection, we got the following dataset structure to represent the news:

- **_id**: id autogenerated by mongoDB;

- **newsURL**: URL that identify each news. Used during all the data mining pipeline as the unique key id value, due to the own definitions of URL;
- **text**: the news body text web scraped from the URL;
- **sentiment_score**: the sum of the sentiment score of each word from the news text, following the same score presented in AFINN-165 wordlist;
- **sentiment**: label of the sentiment score: indifference (score == 0), unhappiness (-2 < score < 0), sadness (score < -2), happiness (0 < score < 2), joy (2 < score);
- **smiley**: emoticon representation of the labeled sentiment, that can assume the following values: [:´( ],[:-( ],[:-|],[:-)],[:-D]. That would be sadness, unhappiness, indifference, happiness and joy, respectively;
- **index**: is the index of the news object when first gathered from the web scraper;
- **news_metadata**: is a JSON attribute, this is the strict response from the News Archive API {
    - snippet: snippet of the news,
    - lead_paragraph: lead paragraph,
    - abstract: abstract of the news, can be null,
    - print_page: page where the news was when printed,
    - blog:[],
    - source: Which product of the New York Times was the source of that news,
    - multimedia: an array of multimedias present in the news web page following this pattern [{multimedia object1}, …],
    - headline: a headline object that has the main headline and a kicker {main: main headline, text,kicker: kicker},

    - keywords: a set of keywords objects, that may have their contents variable throughout their distribution, so it was really hard to consider them, they may follow this structure [{name:, value:}, {subject:, value:}],
    - pub_date: yyyy-mm-ddThh:mm:ssZ, it is the publication date and time,
    - document_type: if it is an article, a blog post, or any other kind of media,
    - news_desk: news category,
    - section_name: news subcategory,
    - subsection_name: other news subcategory,
    - byline:[{author object}], an array of author objects that consists of authors relevant informations ,
    - type_of_material: if it is a news, or other kinds of communication vehicles,
    - word_count: total count of words in the text,
    - slideshow_credits: credits:
    - };
- **news_country**: is an object that indicates which countries were detected inside the news body {
    - countries_detected:[<countryName1>,<countryName2>,...];
    - country_codes_detected:[<countryCode1>,<countryCode2>,...];
    - };

So in the end of dataset enrichment we ended with a relational set of 520208 lines of event log data enriched by the news dataset (that consisted of 34927 news of the specific categories we wanted to analyze)

6 Results' Analysis

Here we are reaching the final steps of the modelled experiment process in Figure 1, after studying the event log, defining the timebox of our analysis, gathering and processing the contextual data and performing the log enrichment we will start to analyze the information we got so far.

As tasks from the second activity modeled in the knowledge discovery milestone from Figure 1, we performed some visual statistical analysis of the enriched event log as we can see in the graphics presented as follows. Then, we ran knowledge discovery algorithms against both datasets in order to compare how beneficial was to use the context enrichment approach. To focus our analysis, we decided to monitor how the contextual data exerted influence upon the status and sub status changing along the process (being our knowledge discovery task).

6.2 General Statistics and Data Visualization

Through data visualization we want to observe the existence of patterns, disturbances, deviances and causes that we cannot be easily seen through raw algorithmic applications. In this section of the analysis, we focus on how the sentiments extracted from the web news impacted on randomly chosen instances of the process flow and check if our approach is going through the right path.

We begin analyzing the sentiment score distribution through a series of aspects like these ones below. Figure 2 shows how the sentiments have been distributed through the news' categories.

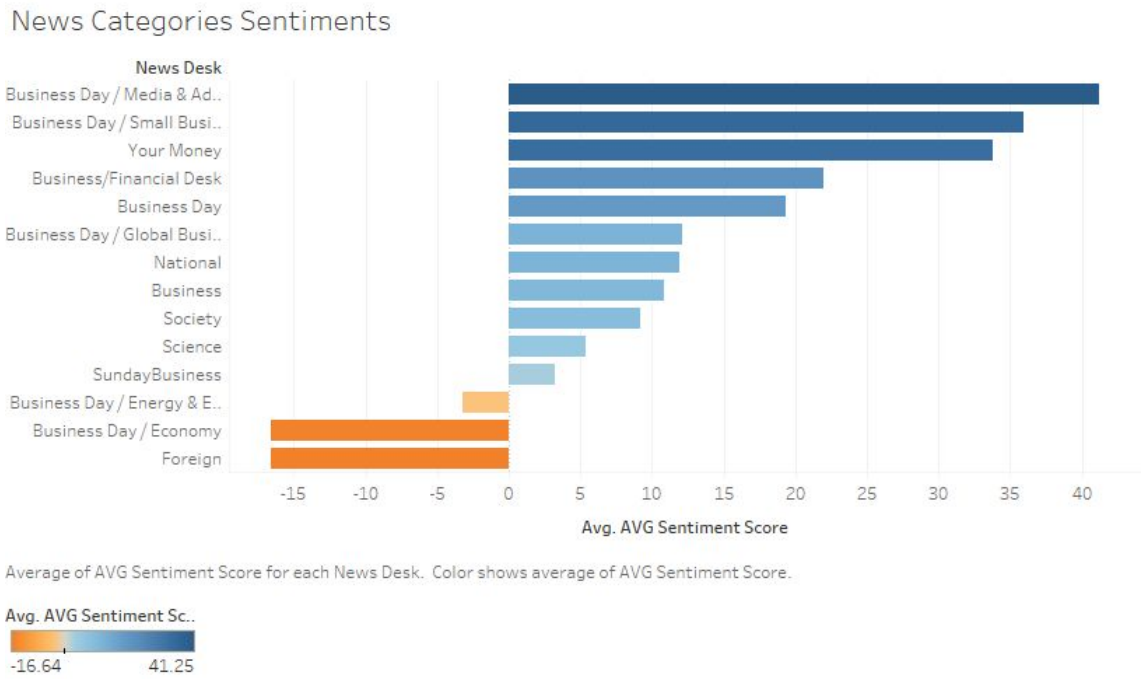

Figure 2: Average Sentiment Score of each news categories during the period of 2010~2012.

The X axis shows the average sentiment score of all 34927, from year 2010 to 2012, news from a given category expressed by Y axis. As mentioned before we are focusing on socioeconomic news, so the chosen categories were all related to Business,

Economy, Society and Global Important news like Scientific ones and Foreign news.

Here we can observe that, in general, the socio-economic situation of all countries which host the process were not that bad, having more positive aspects than negative ones. Even with this impression, it is necessary to look deeper into those statistics to perceive some disturbances.

From this point on, we will dive deep into the correlation between the sentiments and the process' aspects like its execution time and status changes. Figure 3 shows which were the years in which the processes instances ran and its sentiment distribution.

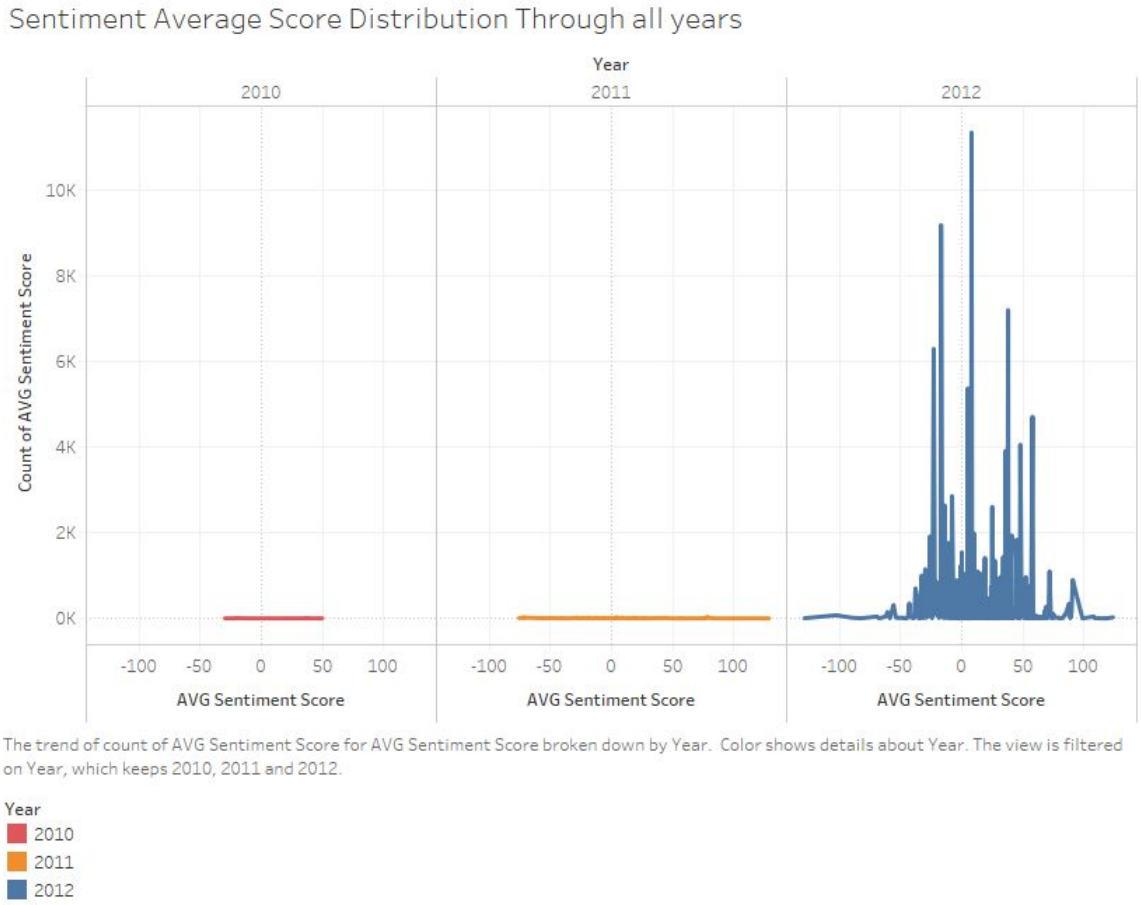

Figure 3: Average Sentiment Score Distribution Through the Years.

By analyzing the graphic in Figure 3, we can observe that 2012 is where most of the process instances were running. However, this disparity between the years is shadowing the last two and shrinking 2012 distribution. To know better the process context in general in those years it may be necessary to split our visualization and explore year by year to get an overview of what we are looking for.

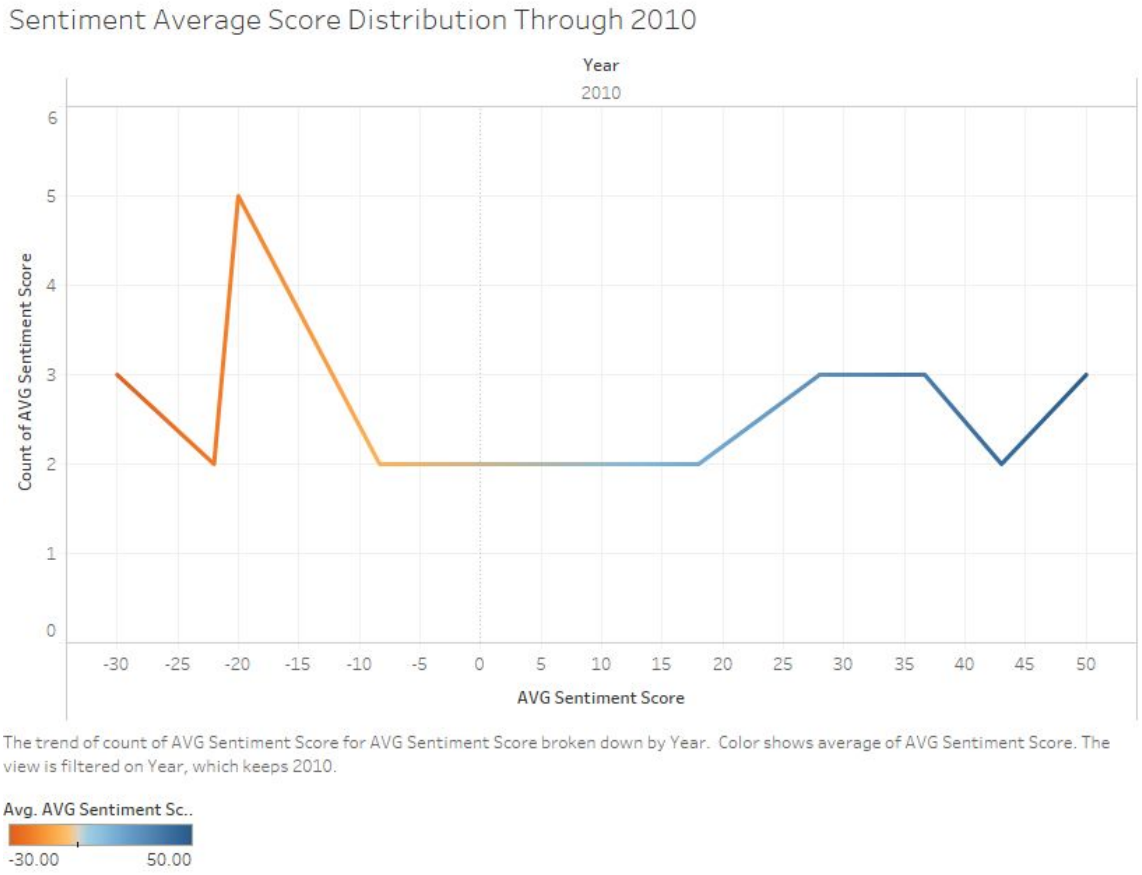

Figure 4: Average Sentiment Score Distribution in 2010.

As we can see in Figure 4, among the 3 years analyzed, 2010 was the year where most people were probably affected by socioeconomic problems, as we can see through that negative expressiveness. Since we are considering a socio-economic context, it may indicate that the countries who host the process' instances passed through a period of

crisis which can be dangerous for the process execution and ultimately success.

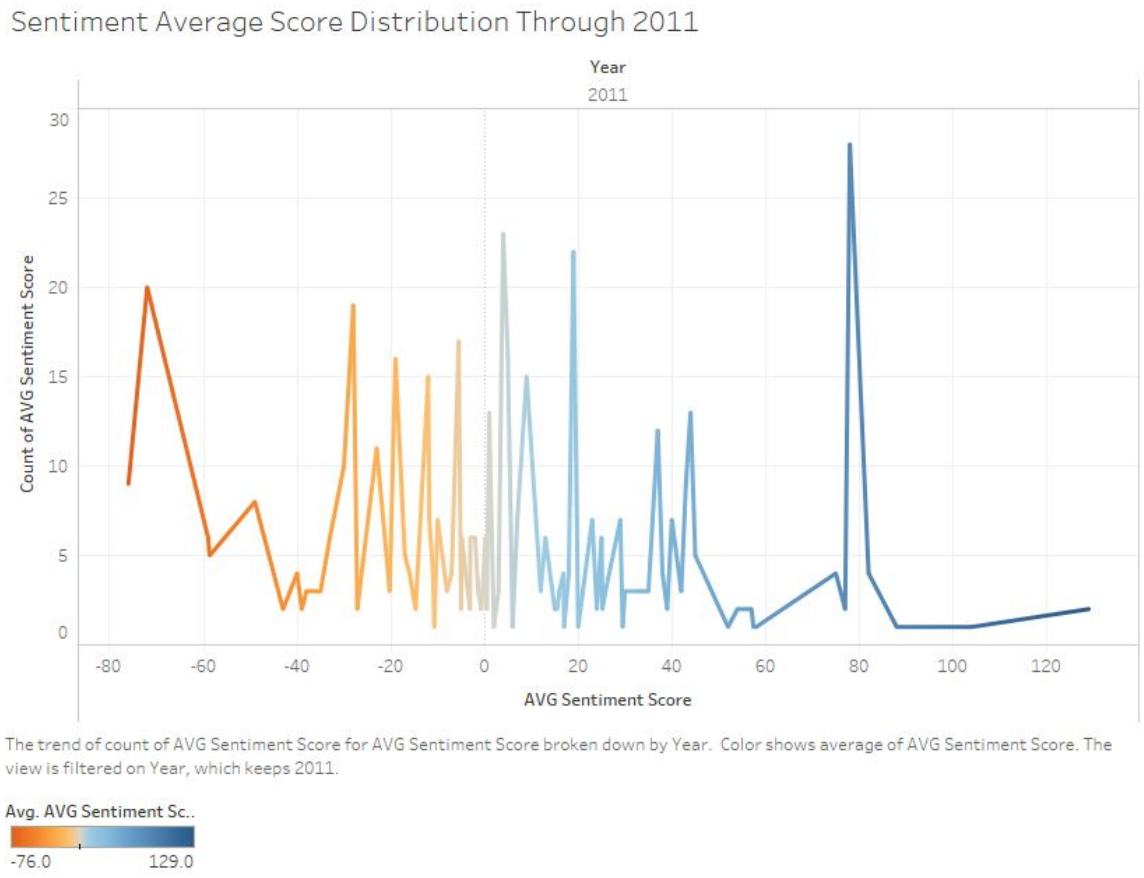

Figure 5: Average Sentiment Score Distribution in 2011

In 2011, we observe some balance between the positive and negative quantities. But this abrupt variation in sentiment distribution might have caused turbulences throughout the execution of the process.

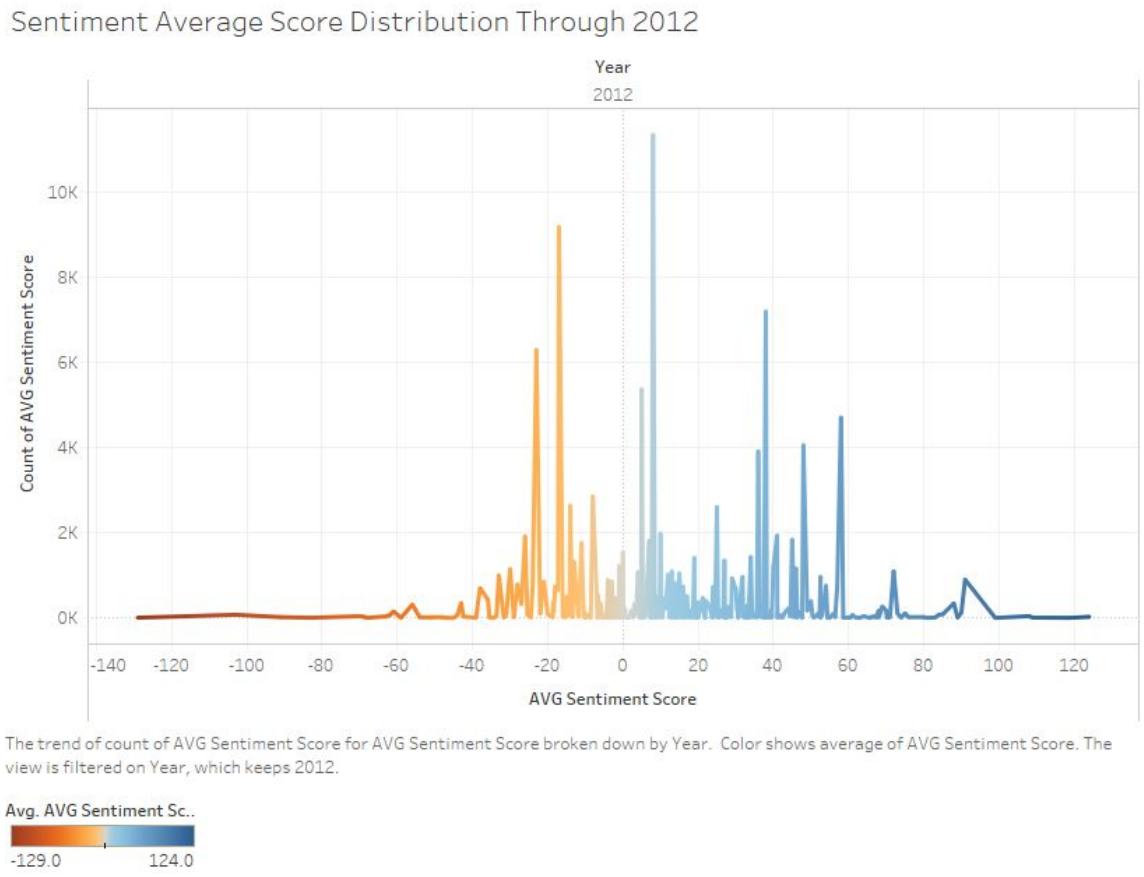

Figure 6: Average Sentiment Score Distribution in 2012

Figure 6 depicts the most expressive year, 2012, when most of the process' instances ran. As, we can see in the graphic shown in Figure 6, this year had a balanced sentiment score distribution, as we can see a format like a normal distribution.

Now we analyze some randomly picked process' instances and how the sentiment variation affected their flow. From studying the process logs and the given documentation of them, we concluded that apparently all process instances end when status is set to complete, but, in some cases the process may be reopened to be solved once more, also through this same documentation we can observe that the dynamic and operational aspect of the process (Change Date Time) flows towards the changing of

this aspects, in process mining terms, the SR_Number defines the cases, the Change Date Time the timestamp lifecycle and the status combined with the sub status define events. Due to this, we are considering as the most susceptible aspects affected by sentiments to be the status and sub status, as they are changed by the work effort of each participant of each process instance.

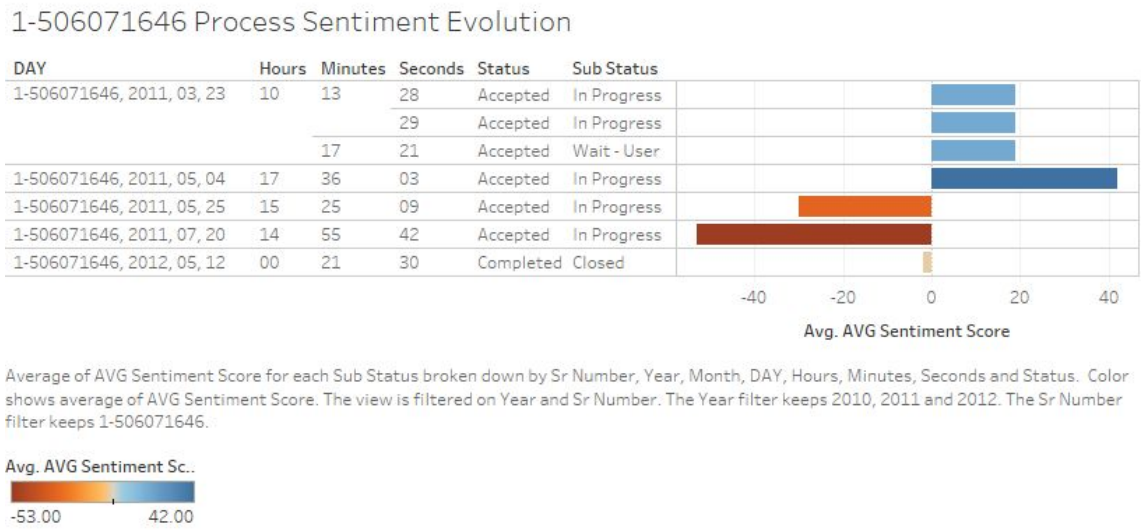

Figure 7: 1-506071646 Process Sentiment Evolution

We analyze the evolution of the instances of the process relating them to the sentiment variation on its socioeconomic context. This socioeconomic context can be perceived by its average score variation, and can be interpreted in that way, the more light blue to blue represents a propitious moment and stability of economy, politics and society in general like when the government is stable, or new industrial technologies are being developed, and as it approaches to orange and red indicates that instability, crisis or unfavorable economic, political or social moments are present, like corruption, changes in labor laws, etc.

In Figure 7, we can see that when the business process context is stable to favorable (shades of blue), that process changed its sub status, indicating fluidity, very fast. But, when the environment began to turn unfavorable the process fluidity slowed down so much that took almost one entire year to it reach its conclusion. Now, considering more examples.

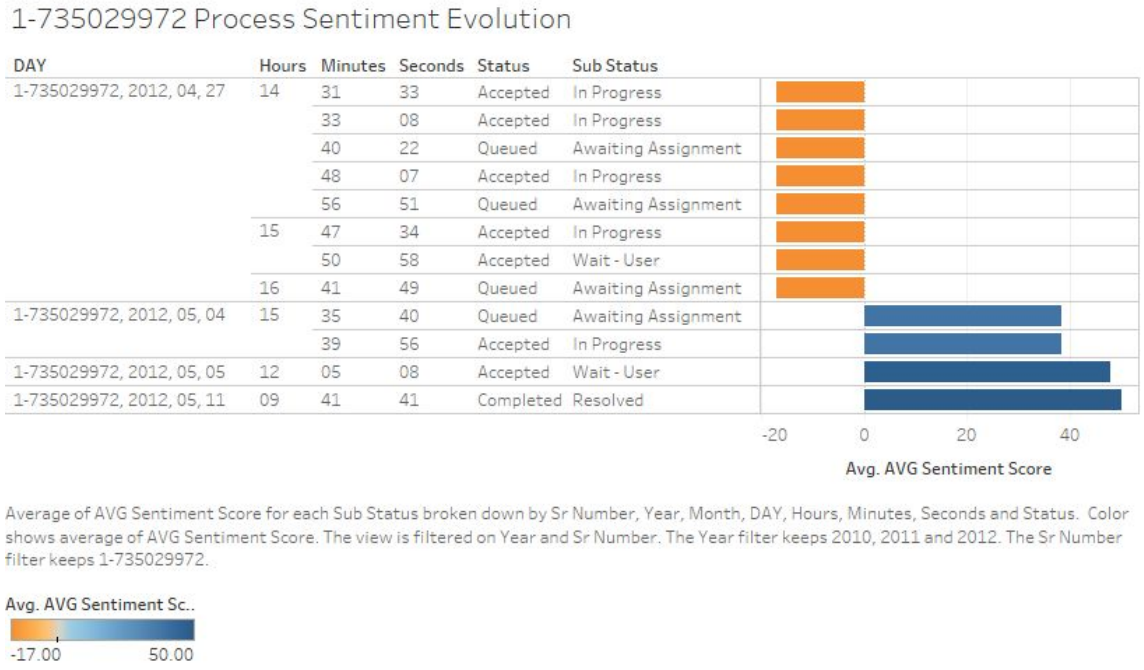

Figure 8: 1-735029972 Process Sentiment Evolution

In Figure 8, we can see that this process instance began in a negative context as the first iterations were being ran in an orange zone of unfavorable situation. It stays that way for an entire week as we can see in the Y axis DAY column that contains the SR_Number, year, month and day. While in this unfavorable context, it progressed while dragging itself through, while waiting for someone to take and resolve it. After

this unfavorable moment, we observe that the context began to change into a more positive vibe, due to it, the process instance sped up a little more, being resolved in just one day, after that the “process waited” for the user to confirm resolution and in the end, it was confirmed as complete.

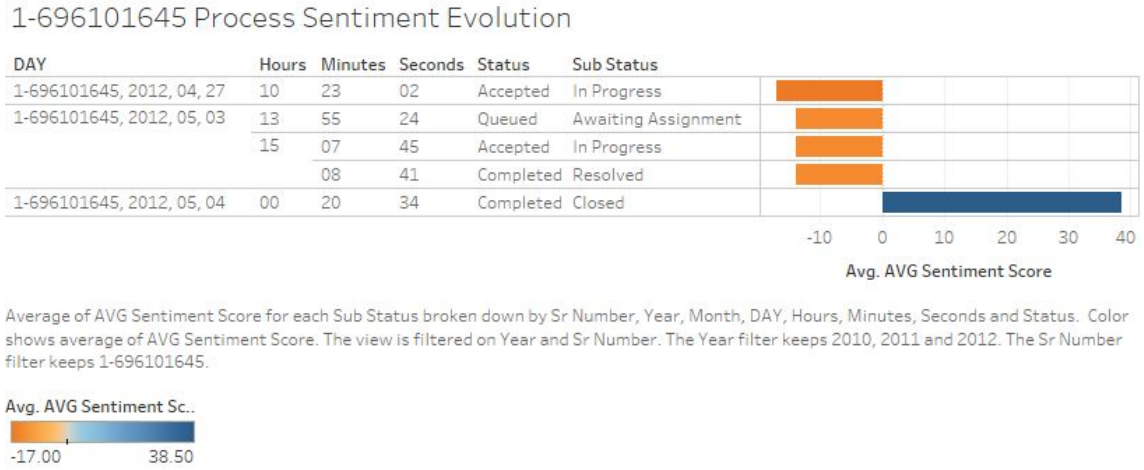

Figure 9: 1-696101645 Process Sentiment Evolution

In Figure 9, we can observe that although it had a small execution path, due to its unfavorable context it took a week to be completed.

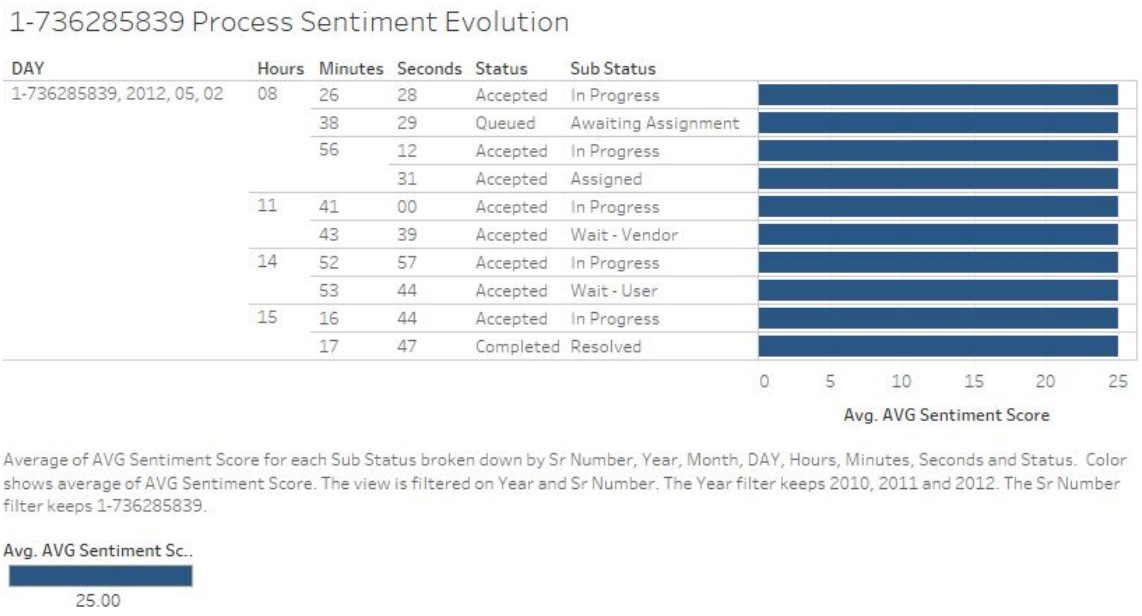

Figure 10: 1-736285839 Process Sentiment Evolution

Figure 10 depicts a very interesting example. Look at how positive, how favorable was the context when this instance was running, its fluidity was high, and the process could be resolved within one day, changing its statuses in a matter of minutes and hours.

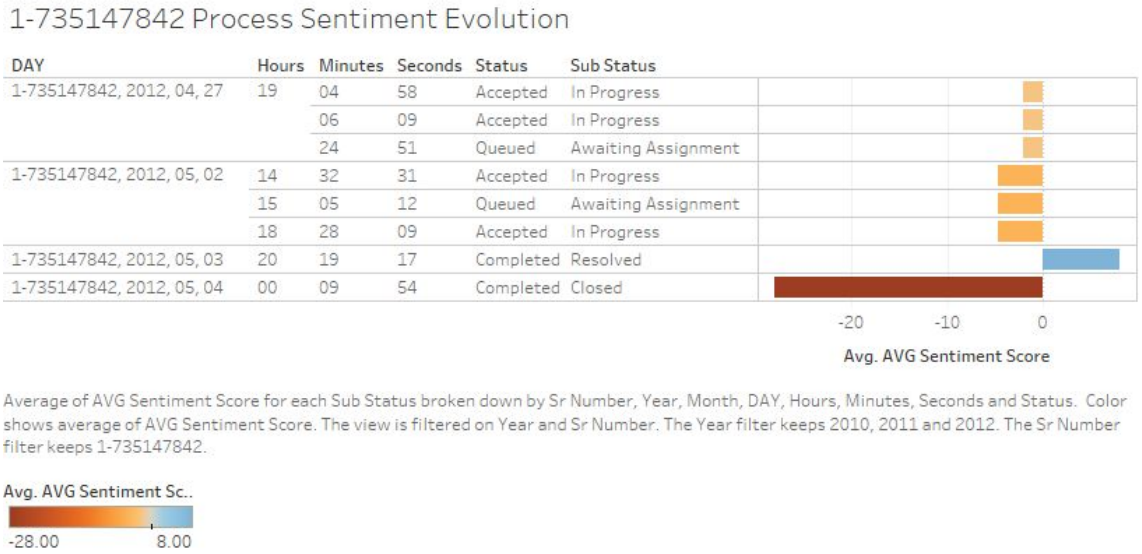

Figure 11: 1-735147842 Process Sentiment Evolution

This instance, in Figure 11, is mostly negative with just one peek of positivity by the final status change of the execution. But we can observe that when there were negativeness present in the environment, the process slowed down and got a week to be completed. And when completed, the environment hosting it will enter a negativeness period, what can be worrying to the next instances which may happen after it.

6.2 **Quantitative Analysis**

Now we reach the final part of our analysis, where we pay more attention to the statistical and quantitative aspect our approach evaluation. We are trying to discuss with this project how important and relevant could be to consider external context when doing process mining. We already have seen that visually there are some linkages between the time spent by a process and its contextual sentiment score, now we will evaluate which were the benefits of aggregating that context to our dataset. Therefore, we applied knowledge discovery in database`s algorithm against the pure event log, and against the enriched log as well in order to compare the results obtained the goal was to enforce what we observed in the section before.

The first algorithm was the Apriori, an algorithm to discover association rules in itemsets[25]. Generating rules like “if itemset contains({x,y}) then itemset (may) contains({z}) as well”.

It is a very popular algorithm to use in process mining, due to it being designed to work with transactional data like our processes. And by applying this algorithm we expect to see rules that originated from context or rules that help us to better understand it.

For this evaluation between the two logs, we defined that the confidence of our rules should be of at least 90% and a support of 30%, that means that the rules discovered by this algorithm had 90% of chance to happen in other instances and 30% of coverage through the datasets.

The results for this running were not that interesting as the enriched dataset produced just one rule more than the original one and both got a little quantity of them, being 4 to the pure one and 5 to the enriched. As we can see the results from the enriched log below in Table 1.

Table 1 presents the rules found in our enriched event log and its aspects of support, confidence, lift and count. As support and confidence are probabilities varying from 0 to 1, count being the number of times the ruleset was perceived in the enriched dataset. And Lift being the measure of predictability for each of those rules. Due to the fact that the final enriched dataset came from a shallow and relatively little process and that its format was not much adequate for this algorithm, may resulted in find less rules than expected.

Table 1: Apriori Algorithm Results over the enriched dataset

#	Rules	support	confidence	lift	count
1	{INVOLVED_ST_FUNCTION_DIV=V3_2} => {INVOLVED_ORG_LINE_3=Org line C}	0.434911487157997	0.984829951812788	1.48588103084111	226244
2	{SUB_STATUS=InProgress} => {STATUS=Accepted}	0.479124656146496	1	1.58096485576397	249244
3	{STATUS=Accepted,INVOLVED_ST_FUNCTION_DIV=V3_2} => {INVOLVED_ORG_LINE_3=Org line C}	0.307419930912118	0.989310238168883	1.49264074859442	159922
4	{SUB_STATUS=InProgress,SUBSECTION=} => {STATUS=Accepted}	0.308369552889523	1	1.58096485576397	160416
5	{SUB_STATUS=InProgress,INVOLVED_ORG_LINE_3=Org line C} => {STATUS=Accepted}	0.339301470376216	1	1.58096485576397	176507

In Blue we perceive the additional rule found by Apriori when analysing the enriched dataset. However, any of those rules helped our analysis. This not so satisfactory result may be originated from dataset limitations and overestimation during algorithm setup.

Table 2: Apriori Algorithm Results over the original dataset

#	Rules	support	confidence	lift	count
1	{INVOLVED_ST_FUNCTION_DIV=V3_2} => {INVOLVED_ORG_LINE_3=Org line C}	0.434911487157997	0.984829951812788	1.485881030841111	226244
2	{SUB_STATUS=InProgress}=>{STATUS=Accepted}	0.479124656146496	1	1.580964855763397	249244
3	{STATUS=Accepted,INVOLVED_ST_FUNCTION_DIV=V3_2}=>{INVOLVED_ORG_LINE_3=Org line C}	0.307419930912118	0.989310238168883	1.492640748594442	15992
4	{SUB_STATUS=InProgress,INVOLVED_ORG_LINE_3=Org line C} => {STATUS=Accepted}	0.339301470376216	1	1.580964855763397	176507

Although, the rules hadn't had the effect we were expecting, we can see much of the org_line attribute appearing. That being a "human" factor is favorable to our analysis, but again, not enough.

The second chosen algorithm was OneR, where we will see how adherent the attributes included when enriching the log are to the classification of the target attribute, Sub Status. To perform the OneR algorithm test and see which were the top 6 most relevant attributes to explain the class of our problem, some preprocessing was needed.

This preprocessing was done to not add bias to OneR when ranking the attributes.
So first we needed to remove the SR_NUMBER, as it is an id and would not add much value to a classification.

Then we removed the timestamp attributes, as it and its similar attributes have their own separated analysis methods. We remove the Status as well, as it is a direct consequence of the sub status and would not make sense to be considered here since OneR would accuse Status as a predictor of Sub Status, but this is already known due to the logs documentation and would pollute the top 6. So, the remainder attributes in the enriched dataset were:

SUB_STATUS (our class attribute);

INVOLVED_ST_FUNCTION_DIV

INVOLVED_ORG_LINE_3

INVOLVED_ST

SR_LATEST_IMPACT

PRODUCT

COUNTRY

OWNER_COUNTRY

OWNER_FIRST_NAME

NEWS_DESK

SECTION

SUBSECTION

AVG_SENTIMENT_SCORE

The OneR algorithm was set with the following parameters:

batchSize = 100, minBucketSize = 6

numDecimalPlaces = 2, Cross validation of 10 folds

It ran 6 times in a row, while removing from each next iteration the attribute chosen before. Remembering that OneR uses just one of the attributes to classify each time it ran. And the results were:

- The first one was OWNER_FIRST_NAME getting 53% of right classifications;
- The second one was OWNER_COUNTRY getting 53% of right classifications;
- The third one was the AVG_SENTIMENT_SCORE getting 49,5% of right classifications;
- The fourth one was the INVOLVED_ST getting 48% of right classifications;
- The fifth one was the PRODUCT getting 48% of right classifications;
- The sixth and last one was the NEWS_DESK getting 48% of right classifications;

That is indeed very favorable to our evaluation as it shows that the most relevant attributes to the success of the process may be the ones linked to human activity in the

process (event owner name, country, involved division) and showed as well that the attributes that came within the contextual aggregation had influence, what corroborates with our thoughts that if something would directly interfere (like being aware of a crisis, changes in labor laws, etc.) with people performing their process' activities, would interfere as well in the process flow, as they would be not in their psychological normal state of mind. Furthermore, we can do one more test.

And the third one chosen was the Decision Table. Set to be ran against the enriched dataset using the attributes determined above by the OneR iterative ranking, as now we are going to evaluate if One R was correct in ranking those attributes. The parameters settings are cross-validation of 10 folds.

The result of the Decision Table test brought a tax of correct classifications of 65%.

Since our dataset is a little generic and shallow in data quantity, we can consider this tax of correct classification good, since it is way above the 50% of randomness and is much similar to the accuracy in [32] as an initial tax that was improved in their machine learning pipeline after the first analysis.

Our approach to use context extracted from a source that directly affects the parts that execute the process seems promising and very likely to produce even better results with real bigger logs and more contextual related sources, e.g. other newspapers.

7 Conclusions

In this chapter we present the conclusions from the entire development and analysis process of this project, as well to expose the difficulties handled through its making and what we intend to be future works.

In general, the results of our analysis were very good to validate the viability of the approach, as the contextual elements aggregate to the process event log were present in the most relevant elements ranking and during the visual analysis of random picked processes contextual sentiment evolution through its execution, and we got an untie decision through applying the decision table which return a tax of 65% of right answers. Probably the Apriori did not had a satisfactory result due to the fact that the dataset was not in the usual format of item sets, but instead in its integrity, something that could not be done within R, at least as far as I tried and because the event log itself was shallow.

1. Difficulties

The main difficulties handled in this project were the fact that most of this project required external learning from what was thought by graduation's main course, to deal with data mining classical problems as data capturing and data pre and postprocessing,

to choose the most adequate technologies as there were many ones to perform the sentiment analysis, but they were paid in its majority and require heavy training, and finally to council the time between the last disciplines of the graduation, the job and research. But, in the end we could make it.

There were some secondary ones involving hardware restrictions, since we were manipulating huge sets of data in the first steps of our process, sometimes it was required to let the machines that were working on them running up all night, and sometimes failures would happen in between this period and need to be restarted again.

2. Future Work

We see as future works the following items: transforming the acquiring, analysis and management process of such massive quantity of data into an online automatic framework in order to scale its use, to evaluate our proposal with other datasets and other contextual sources, to perform more statistical analysis like time series and finally evaluate the accuracy of the sentiment analysis module with other data (social media's, web articles from blogs, etc.).

References

1. LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, v. 5, n. 1, p. 1-167, 2012.
2. BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para Análise de Sentimentos em mídias sociais. 2015.
3. VARGIU, Eloisa; URRU, Mirko. Exploiting web scraping in a collaborative filtering-based approach to web advertising. **Artificial Intelligence Research**, v. 2, n. 1, p. 44, 2012.
4. VAN DER AALST, Wil MP. Business process management: a comprehensive survey. **ISRN Software Engineering**, v. 2013, 2013.
5. CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S.. . Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and data Engineering**, v. 8, n. 6, p. 866-883, 1996.
6. VAN DER AALST, Wil. Process mining: Overview and opportunities. **ACM Transactions on Management Information Systems (TMIS)**, v. 3, n. 2, p. 7, 2012.
7. MINING, What Is Data. Data Mining: Concepts and Techniques. **Morgan Kaufinann**, 2006.
8. BASSETT, Lindsay. **Introduction to JavaScript Object Notation: A to-the-point Guide to JSON**. " O'Reilly Media, Inc.", 2015.
9. JUNIOR, Francisco de Assis Ribeiro. Programação Orientada a Eventos no lado do servidor utilizando Node. js. 2012.
10. FLANAGAN, David. **JavaScript: the definitive guide**. " O'Reilly Media, Inc.", 2006.
11. NOVAK, Petra Kralj et al. Sentiment of emojis. **PloS one**, v. 10, n. 12, p. e0144296, 2015.
12. CHODOROW, Kristina. **MongoDB: The Definitive Guide: Powerful and Scalable Data Storage**. " O'Reilly Media, Inc.", 2013.
13. VAN DER AALST, Wil MP et al. Business process mining: An industrial application. **Information Systems**, v. 32, n. 5, p. 713-732, 2007.
14. VAN DER AALST, Wil et al. Process mining manifesto. In: **International Conference on Business Process Management**. Springer, Berlin, Heidelberg, 2011. p. 169-194.
15. MONIRUZZAMAN, A. B. M.; HOSSAIN, Syed Akhter. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. **arXiv preprint arXiv:1307.0191**, 2013.
16. Team, R. D. C. (2011). R Development Core Team: R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2011.
17. BOSE, RP Jagadeesh Chandra; VAN DER AALST, Wil MP. Context aware trace clustering: Towards improving process mining results. In: **Proceedings of**

the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2009. p. 401-412.
18. LAUDON, Kenneth C.; LAUDON, Jane P. **Management information system**. Pearson Education India, 2016.
19. BOSE, RP Jagadeesh Chandra; VAN DER AALST, Wil MP. Context aware trace clustering: Towards improving process mining results. In: **Proceedings of the 2009 SIAM International Conference on Data Mining**. Society for Industrial and Applied Mathematics, 2009. p. 401-412.
20. ADOMAVICIUS, Gediminas; TUZHILIN, Alexander. Context-aware recommender systems. In: **Recommender systems handbook**. Springer US, 2015. p. 191-226.
21. SANTO CARVALHO, Juliana do Espírito; SANTORO, Flávia Maria; REVOREDO, Kate. A method to infer the need to update situations in business process adaptation. **Computers in Industry**, v. 71, p. 128-143, 2015.
22. DA CUNHA MATTOS, Talita et al. A formal representation for context-aware business processes. **Computers in Industry**, v. 65, n. 8, p. 1193-1214, 2014.
23. LIU, Rey-Long; LU, Yun-Ling. Incremental context mining for adaptive document classification. In: **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2002. p. 599-604.
24. KISELEVA, Julia. Context mining and integration into predictive web analytics. In: **Proceedings of the 22nd International Conference on World Wide Web**. ACM, 2013. p. 383-388.
25. AGRAWAL, Rakesh et al. Fast algorithms for mining association rules. In: **Proc. 20th int. conf. very large data bases, VLDB**. 1994. p. 487-499.
26. PANG, Bo et al. Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval**, v. 2, n. 1–2, p. 1-135, 2008.
27. KUNG, Peter et al. Business process monitoring & measurement in a large bank: challenges and selected approaches. In: **Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on**. IEEE, 2005. p. 955-961.
28. Ramos, E. C., Santoro, F., Baião, F. - BP ECREL: UM MÉTODO DE IDENTIFICAÇÃO DE VARIÁVEIS RELEVANTES DO CONTEXTO EXTERNO ASSOCIADAS AO PROCESSO DE NEGÓCIO. 2011
29. DI FRANCESCOMARINO, Chiara et al. Clustering-based predictive process monitoring. IEEE Transactions on Services Computing, 2016.
30. MAGGI, Fabrizio Maria et al. Predictive monitoring of business processes. In: International Conference on Advanced Information Systems Engineering. Springer, Cham, 2014. p. 457-472.
31. TAX, Niek et al. Predictive business process monitoring with LSTM neural networks. In: International Conference on Advanced Information Systems Engineering. Springer, Cham, 2017. p. 477-492.
32. METZGER, Andreas et al. Comparing and combining predictive business process monitoring techniques. IEEE Transactions on Systems, Man, and Cybernetics: Systems, v. 45, n. 2, p. 276-290, 2015.