Exploring Data Analytics of Data Variety

Tiago Cruz, Jorge Oliveira e Sá and José Luís Pereira

ALGORITMI Research Center, University of Minho, Guimarães, Portugal
a66785@alunos.uminho.pt;{jos,jlmp}@dsi.uminho.pt

Abstract. The Internet allows organizations managers access to large amounts of data, and this data are presented in different formats, i.e., data variety, namely structured, semi-structured and unstructured. Based on the Internet, this data variety is partly derived from social networks, but not only, machines are also capable of sharing information among themselves, or even machines with people. The objective of this paper is to understand how to retrieve information from data analysis with data variety. An experiment was carried out, based on a dataset with two distinct data types, images and comments on cars. Techniques of data analysis were used, namely Natural Language Processing to identify patterns, and Sentimental and Emotional Analysis. The image recognition technique was used to associate a car model with a category. Next, OLAP cubes and their visualization through dashboards were created. This paper concludes that it is possible to extract a set of relevant information, namely identifying which cars people like more/less, among other information.

Keywords: sentimental and emotional analysis, machine learning, data analysis techniques.

1 Introduction

With the increasing use of the internet by people and organizations, the amount of information has grown exponentially. The digital universe presents a diversity of data such as social networking data, RFID sensor data, geospatial data, website data, mobile device data, among others. This data presents a different data formats, i.e., variety, and can be classified as structured, semi-structured or unstructured data.

Structured data has a rigid structure, usually stored in databases. Unstructured data does not have a defined structured, and it can be for example text, videos, audio files, photographs, among others [1]. Semi-structured data has a flexible structure, containing a set of tags or markers, and can be found in CSV, XML, or JSON files [2].

The data analysis can bring interesting information to organizations [3]. Usually, only 5% of the data belongs to structured format, when the remaining 95% are semi-structured or unstructured, and normally these data are not analyzed [1]. This reveals that there is a large amount of data that should be analyzed to understand its potential, which is not the case in most organizations.

The data analysis with data variety can be an important aspect for organizations, as it can offer relevant information to organization managers, facilitating the decision-making process.

There are different data analysis techniques capable of extracting information relevant to decision-making processes. It is considered that structured data analysis techniques are already mature and validated by the scientific community and are widely used by organizations, but this is not true for semi-structured and unstructured data.

The objective of this paper is to understand if it is possible to extract information insights from this data variety, i.e., unstructured and semi-structured data.

This paper is based on an experiment, using datasets composed of semi-structured and unstructured data. Data analysis techniques were applied according to the data formats and the objective is to be able to extract information from these data.

This paper is structured as follows: Section 2 presents the related work, i.e., several data analysis techniques are presented for semi-structured and unstructured data; In section 3 the experience is described, describing in detail the datasets, the tools and techniques used, the architecture of the solution developed and the results of the experiment test; In section 4 the results obtained are discussed through data analysis presented with dashboards; Finally, in section 5, conclusions, limitations and future work are presented.

2 Related Work

One of the concerns of this work was to identify/leverage analytical techniques that can be used in semi-structured and unstructured data. Tables 1 and 2 identify some studies that use data analysis techniques and which data types are used.

Table 1. semi-structured data techniques

Reference	Techniques	Data type
[4],[15],[16]	Natural Language Processing	text
[4]	HTML Trees (DOM)	text
[5],[6]	Link Prediction	Collaboration Networks; Social Networks; Infrastructure Networks; Sports Networks; Biology Networks
[7]	Probabilistic Three-way Entity Model (TEM)	text
[8]	Entity Linking	Entities (text)
[9],[10],[11]	Naive Bayes Classifier	Text and numeric values
[12]	Multilayer Perceptron, (MLP)	Time series
[13]	Entity Recognition	text
[13],[14]	Sentiment Analysis	text

Table 2. unstructured data techniques

Reference	Techniques	Data type
[4],[15],[16]	Natural Language Processing	text
[17],[18]	Topic Modeling Probabilistic	text
[19]	Latent Semantic Visualization	Time-stamps
[20]	First Story Detection	text
[21]	Event Extraction	Entities (text)
[22]	Data Discovery	text
[13]	Entity Recognition	text
[13], [14]	Sentiment Analysis	text
[23]	Image Recognition	image

The techniques described in the tables depends on the characteristics of the data types used in the identified studies.

In the following section, the techniques used are Natural Language Processing, as well as Sentimental and Emotional Analysis for comments about cars and the Image Recognition Technique to analyze car images.

3 Experiment Description

The experiment will use a dataset with two data formats as described in section 3.1. The tools used are presented in section 3.2. In Section 3.3, the solution architecture is described. Finally, the solution is tested with a set of test data and the results obtained are evaluated.

3.1 Datasets Used

The dataset used in this experiment, was used in a study on classification of entities based on opinions and it consists of 3 years of comments on cars, namely 2007, 2008 and 2009 [24]. This dataset has up to 250 cars models per year. The dataset content is divided between the fields: author, date, and the textual opinion about the car. The data format inserted in this dataset, are considered semi-structured.

In addition to the dataset referred above, a dataset containing car images from car models of the year 2008 was included. This dataset is considered unstructured data.

3.2 Technologies and Tools Used

In this section, it will be presented the technologies used in this experiment and the associated techniques if applicable, see Table 3.

3.3 Experiment solution architecture

In Figure 1 the experiment solution architecture is presented. This architecture aims to extract information from the dataset.

As shown in figure 1, the dataset is composed of comments on cars with the semi-structured format and images of cars with the unstructured format.

The image car model used as input data will be classified using Visual Recognition IBM tool. To do this it was necessary to create a custom classifier to be able to recognize which category belongs to each car model.

The Natural Language Understanding tool was used to retrieve the sentimental and emotional analysis from the comments dataset, and to do this it was necessary to use the Watson Knowledge Studio. In the Watson Knowledge Studio tool, it was possible to create a machine-learning model capable of identify a set of patterns relevant to this experiment.

Table 3. Tools used in the experiment

Tool	Description	Techniques (if applicable)
Natural Language Understanding IBM	The purpose was to retrieve a set of information: entities, keywords, relations, and sentimental and emotional analysis; of the commentaries alluding to a car model.	Natural Language Processing; Emotional Analysis; Sentimental Analysis
Watson Knowledge Studio IBM	In the experiment, this tool was used as a supplement to the Natural Language Understanding IBM tool, and its objective was to recognize a set of standards through a machine learning model.	
Visual Recognition IBM	The Visual Recognition IBM was used to associate an image with a car category, and this was possible using a custom classifier for the categories that were intended to be recognized.	Image Recognition
Postman	The Postman was used to perform POST requests for the Visual Recognition (to classify a car model into a category), and the Natural Language Understanding (for information extraction about comments associated with a car). The information obtained was saved in JSON format.	
SQL Server	The version used was SQL Server Enterprise Edition 2014 and served as a data repository, where the internship database was stored and the On-Line Analytical Processing (OLAP) cubes.	
SSIS	The SQL Server Integration Services (SSIS) are used to develop some Extraction, Transform and Loading (ETL) functionality.	
SSAS	The SQL Server Analysis Services (SSAS) are used to create OLAP cubes.	
Talend	The Talend was used because SSIS did not have the functionality of handling data in JSON format. Therefore, this tool was used to perform ETL tasks, namely transforming JSON format data and inserting into a database.	

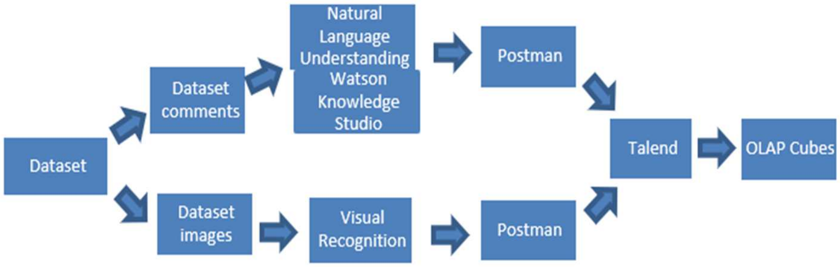


Fig. 1. Experiment solution architecture

As seen in figure 1, the data obtained from the Natural Language Understanding and Visual Recognition tools, are presented in the Postman tool. This tool saves the files in JSON format. To insert these data formats into SQL Server, it was necessary to use the Talend tool, where ETL tasks were performed, and then insert the data into a staging database in the SQL Server tool.

To create the OLAP cubes, it was necessary to use SSIS to define an ETL process to perform a set of transformations on the data, to ensure that they were ready to be inserted into the OLAP cubes. In SSAS, three OLAP cubes were created as entities, relations and keywords studied different and unrelated information.

3.4 Experiment solution testing

To be able to evaluate the possibility to retrieve information using the experiment, an excerpt of comments and the corresponding images are retrieved from the dataset to understand if the architecture proposed can:

- classify the car image and associate it with a car category; and
- find a set of relations, keywords, entities, sentiments and emotions in the excerpt of comments.

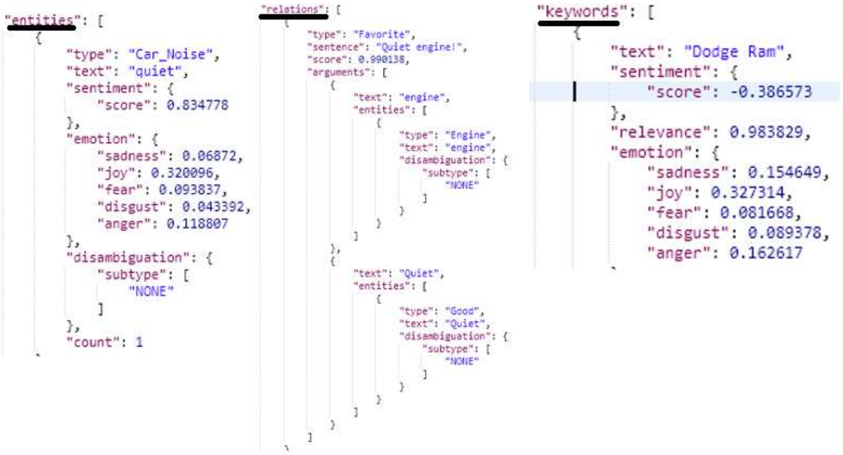
Regarding the dataset of images and to classify a model of a car in a certain category, a machine learning tool was used, using Visual Recognition technique. For this, a classifier was developed that from the image of the car can associate it with a category. Initially, a set of images with the corresponding car category was provided, so the tool can learn how to classify images into categories. To validate this classification, a set of tests was performed, for example, for an image that refers to the "dodge_pickup_3500" model, resulting in a classification in three categories: "Coupe" with a precision of 0.56, "Pickup_truck "with precision of 0.99 and" Sport_utility "with precision 0.84. The chosen category was “Pickup_truck”, because it obtains the best result.

Regarding the comment dataset, a set of tests was carried out with an excerpt of comments about cars, to extract information from these data. It was possible to identify a set of information as: entities, keywords and relations. This information was possible to obtain, because in the Watson Knowledge Studio tool (associated as a supplement to the Natural Language Understanding tool) exists a machine learning model capable of recognizing a set of patterns. A set of entities and relationships for the created model are identified, to be able to recognize these patterns in the Watson Knowledge Studio tool. A set of eight training files containing excerpts of comments for the model to be able to recognize entities and relationships are used. Next, the machine learning model created in the Watson Knowledge Studio is used in the Natural Language Understanding tool, where associated with the different relations, keywords and entities will recognize a set of information using Natural Language Processing technique and associate to each keyword and entity the sentimental and emotional analysis.

An excerpt of comments on the model "dodge_pickup_3500" is used as a test and the result obtained is shown in figure 2. As shown in figure 2, it was possible to extract a set of information from this data, namely keywords, entities and relations. It was possible to obtain the sentimental analysis through the variable score, and emotional analysis through the variables Emotion_sadness, Emotion_Joy, Emotion_fear, Emotion_disgust and Emotion_anger, of the entities and keywords.

For example, it was possible to identify the entity "Car_Noise" and associate to this entity the variable "score" sentimental analysis with the precision of 0.83, in which it is identified as being positive. Note this value varies between 1 and -1. In the keywords was identified for example "dodge Ram" and emotional analysis of that word obtain the following values of emotions: sadness with precision of 0.15; joy with precision of

0.32; fear with precision of 0.08; disgust with precision of 0.07; and anger with precision of 0.12. Note the range of emotion values varies between 0 and 1. In relations it was possible to identify the relationship between two distinct entities being "engine" and "Quiet".



```
"entities": [
    {
        "type": "Car_Noise",
        "text": "quiet",
        "sentiment": {
            "score": 0.834778
        },
        "emotion": {
            "sadness": 0.06872,
            "joy": 0.320096,
            "fear": 0.093837,
            "disgust": 0.043392,
            "anger": 0.118807
        },
        "disambiguation": {
            "subtype": [
                "NONE"
            ]
        },
        "count": 1
    ,
```

```
"relations": [
    {
        "type": "Favorite",
        "sentence": "Quiet engine!",
        "score": 0.990138,
        "arguments": [
            {
                "text": "engine",
                "entities": [
                    {
                        "type": "Engine",
                        "text": "engine",
                        "disambiguation": {
                            "subtype": [
                                "NONE"
                            ]
                        }
                    }
                ]
            },
            {
                "text": "Quiet",
                "entities": [
                    {
                        "type": "Good",
                        "text": "Quiet",
                        "disambiguation": {
                            "subtype": [
                                "NONE"
                            ]
                        }
                    }
                ]
            }
        ]
    ,
```

```
"keywords": [
    {
        "text": "Dodge Ram",
        "sentiment": {
            "score": -0.386573
        },
        "relevance": 0.983829,
        "emotion": {
            "sadness": 0.154649,
            "joy": 0.327314,
            "fear": 0.081668,
            "disgust": 0.089378,
            "anger": 0.162617
        ,
```

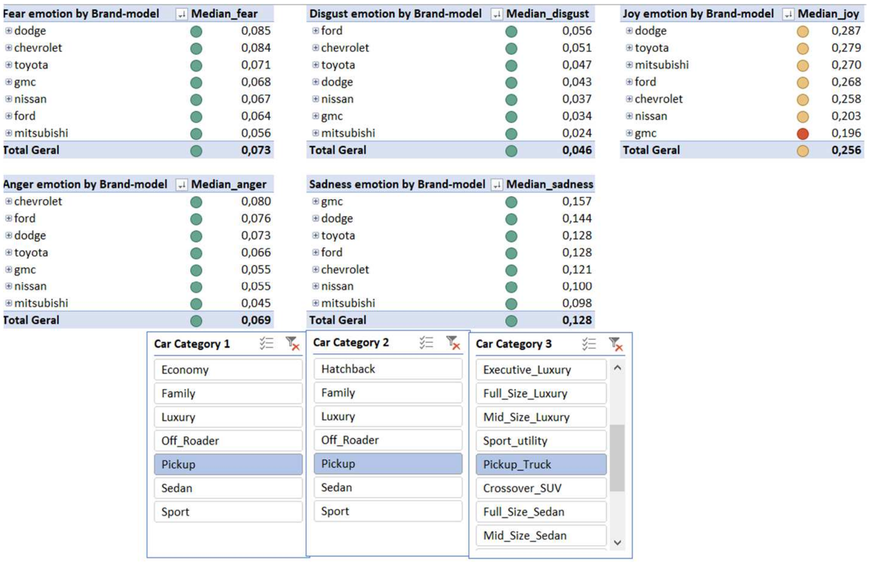
Fig. 2. Entities, relations and keywords

4 Results discussion

This section presents an analysis performed with the OLAP cubes, and its visualization through dashboards. A dashboard composed of emotional analysis was developed using the Entities cube, followed by a dashboard for sentimental analysis and other features of a car model using the three cubes generated, namely Keywords, Entities and Relations Cube.

Figure 3 presents five types of emotions associated with a set of car models. The choice of these car models is due the fact that a segmentation of data divided as “Car Category 1”, “Car Category 2” and “Car Category 3” was created, of the car models that belong to "Car Category 1" "Pickup", "Car Category 2" "Pickup" and "Car Category 3" "Pickup_Truck". For each of the five tables, an indicator value was used to measure the emotion of the car model and a colored mark was also added to understand if the value is acceptable (green color), neutral (yellow color) or bad (red color). The tables that measure the emotions "anger", "sadness", "fear" and "disgust" do not have values with significant impact, so their analysis in this case may not have impact in the decision-making process. The joy emotion has a greater impact than the others, but even so, most car models have a value considered neutral, only the “gmc” car model was below normal.

At last, the figure 4 illustrates a deeper analysis of several factors to a specific car model. In this figure, a set of tables with several factors associated with a car model of the brand “volvo” and model “s40” is presented.



Fear emotion by Brand-model		Median_fear
dodge	🟢	0,085
chevrolet	🟢	0,084
toyota	🟢	0,071
gmc	🟢	0,068
nissan	🟢	0,067
ford	🟢	0,064
mitsubishi	🟢	0,056
Total Geral	🟢	**0,073**

Disgust emotion by Brand-model		Median_disgust
ford	🟢	0,056
chevrolet	🟢	0,051
toyota	🟢	0,047
dodge	🟢	0,043
nissan	🟢	0,037
gmc	🟢	0,034
mitsubishi	🟢	0,024
Total Geral	🟢	**0,046**

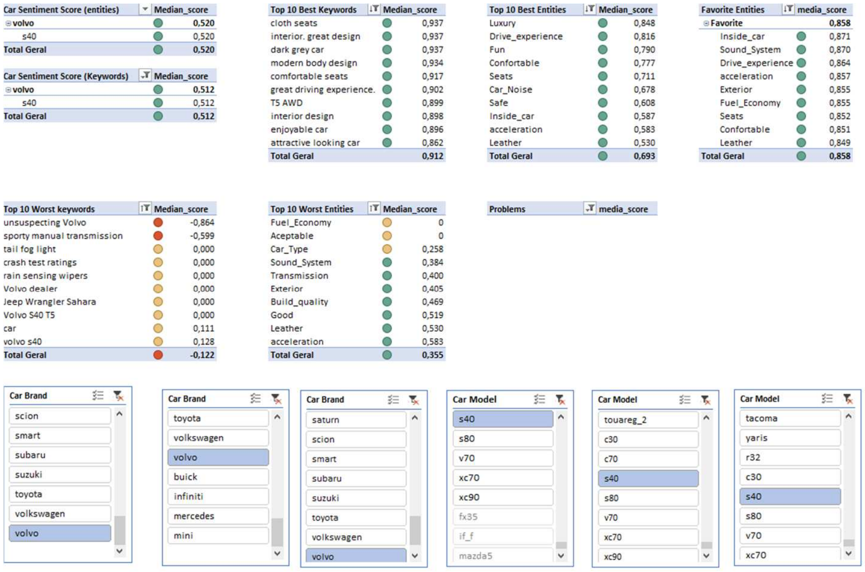
Joy emotion by Brand-model		Median_joy
dodge	🟡	0,287
toyota	🟡	0,279
mitsubishi	🟡	0,270
ford	🟡	0,268
chevrolet	🟡	0,258
nissan	🟡	0,203
gmc	🔴	0,196
Total Geral	🟡	**0,256**

Anger emotion by Brand-model		Median_anger
chevrolet	🟢	0,080
ford	🟢	0,076
dodge	🟢	0,073
toyota	🟢	0,066
gmc	🟢	0,055
nissan	🟢	0,055
mitsubishi	🟢	0,045
Total Geral	🟢	**0,069**

Sadness emotion by Brand-model		Median_sadness
gmc	🟢	0,157
dodge	🟢	0,144
toyota	🟢	0,128
ford	🟢	0,128
chevrolet	🟢	0,121
nissan	🟢	0,100
mitsubishi	🟢	0,098
Total Geral	🟢	**0,128**

Car Category 1	Car Category 2	Car Category 3
Economy	Hatchback	Executive_Luxury
Family	Family	Full_Size_Luxury
Luxury	Luxury	Mid_Size_Luxury
Off_Roader	Off_Roader	Sport_utility
[✓ Pickup]	[✓ Pickup]	[✓ Pickup_Truck]
Sedan	Sedan	Crossover_SUV
Sport	Sport	Full_Size_Sedan
		Mid_Size_Sedan

Fig. 3. Emotional analysis



Car Sentiment Score (entities)		Median_score
volvo	🟢	**0,520**
s40	🟢	0,520
Total Geral	🟢	**0,520**

Car Sentiment Score (Keywords)		Median_score
volvo	🟢	**0,512**
s40	🟢	0,512
Total Geral	🟢	**0,512**

Top 10 Best Keywords		Median_score
cloth seats	🟢	0,937
interior. great design	🟢	0,937
dark grey car	🟢	0,937
modern body design	🟢	0,934
comfortable seats	🟢	0,917
great driving experience.	🟢	0,902
T5 AWD	🟢	0,899
interior design	🟢	0,898
enjoyable car	🟢	0,896
attractive looking car	🟢	0,862
Total Geral		**0,912**

Top 10 Best Entities		Median_score
Luxury	🟢	0,848
Drive_experience	🟢	0,816
Fun	🟢	0,790
Confortable	🟢	0,777
Seats	🟢	0,711
Car_Noise	🟢	0,678
Safe	🟢	0,608
Inside_car	🟢	0,587
acceleration	🟢	0,583
Leather	🟢	0,530
Total Geral	🟢	**0,693**

Favorite Entities		media_score
Favorite		**0,858**
Inside_car	🟢	0,871
Sound_System	🟢	0,870
Drive_experience	🟢	0,864
acceleration	🟢	0,857
Exterior	🟢	0,855
Fuel_Economy	🟢	0,855
Seats	🟢	0,852
Confortable	🟢	0,851
Leather	🟢	0,849
Total Geral	🟢	**0,858**

Top 10 Worst keywords		Median_score
unsuspecting Volvo	🔴	-0,864
sporty manual transmission	🔴	-0,599
tail fog light	🟡	0,000
crash test ratings	🟡	0,000
rain sensing wipers	🟡	0,000
Volvo dealer	🟡	0,000
Jeep Wrangler Sahara	🟡	0,000
Volvo S40 T5	🟡	0,000
car	🟡	0,111
volvo s40	🟡	0,128
Total Geral	🔴	**-0,122**

Top 10 Worst Entities		Median_score
Fuel_Economy	🟡	0
Aceptable	🟡	0
Car_Type	🟡	0,258
Sound_System	🟢	0,384
Transmission	🟢	0,400
Exterior	🟢	0,405
Build_quality	🟢	0,469
Good	🟢	0,519
Leather	🟢	0,530
acceleration	🟢	0,583
Total Geral	🟢	**0,355**

Problems	media_score

Car Brand	Car Brand	Car Brand	Car Model	Car Model	Car Model
scion	toyota	saturn	[✓ s40]	touareg_2	tacoma
smart	volkswagen	scion	s80	c30	yaris
subaru	[✓ volvo]	smart	v70	c70	r32
suzuki	buick	subaru	xc70	[✓ s40]	c30
toyota	infiniti	suzuki	xc90	s80	[✓ s40]
volkswagen	mercedes	toyota	fx35	v70	s80
[✓ volvo]	mini	volkswagen	if_f	xc70	v70
		[✓ volvo]	mazda5	xc90	xc70

Fig. 4. Sentimental car analysis

In the upper left corner of Figure 4, the score obtained from the sentimental analysis of the specific car model is shown, either from the Entities or the Keywords cube. It is a pretty good value, that means people got a good impression of the car model. In the three tables on the right, the “top 10 keywords”, “top 10 entities” and “top 10 favorite entities” are shown. It is easy to check that people liked the interior, the seats, the acceleration and some other features. In the three tables below, “top 10 worst keywords”, “top 10 worst entities” and “problems” are measured using the “median_score” that measures the sentiment. In these three tables referenced above, there are not many components referenced as negative, other than the transmission of the car, and as neutral the fuel consumption of the car.

In these different analyses presented above, the resulted obtained were quite satisfactory, since it was possible to extract relevant information about the cars as: sentimental and emotional analysis; what problems/weakness and components did people most like/strengths; made it possible to perform a deep analysis of a car model.

5 Conclusions, Limitations and Future Work

5.1 Conclusions

This paper proposed an experimental work to understand if it is possible to extract information insights from data variety, i.e., unstructured and semi-structured. It was quickly realized that to take advantage of unstructured and semi-structured data, it was necessary to find a way to transform and store them. For this, an architecture solution was defined with the objective of extracting useful information from unstructured and semi-structured data. Next, the architecture solution was evaluated and tested, which allowed to realize that it can extract relevant information from unstructured and semi-structured data.

An analytical platform was developed to carry out different analyzes of the data obtained from the experiment. In the analytical platform, it was possible to verify a set of relevant information such as: sentimental and emotional analysis; what are the problems/strengths associated with a car; and a detailed analysis of a car.

5.2 Limitations

In this paper the following limitation is identified, the experiment created cannot be used as a generic solution. Consequently, for each specific case, it is necessary to calibrate the tools to the dataset available and to the results pretended.

5.3 Future Work

As future work, it will be relevant to carry out a study using unstructured and semi-structured data in a real organization, to show that the data variety can reveal useful information’s for analysis and, depending on data formats and types, explore other techniques of data analysis to understand what type of information can be obtained.

Acknowledgments

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT - Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

References

[1] H. Baars and H.-G. Kemper, "Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework," *Inf. Syst. Manag.*, vol. 25, no. 2, pp. 132–148, Mar. 2008.

[2] T. Asai, K. Abe, S. Kawasoe, H. Arimura, H. Sakamoto, and S. Arikawa, "Efficient Substructure Discovery from Large Semi-structured Data," *IEICE Trans. Inf. Syst.*, pp. 2754–2763, 2004.

[3] P. Russom, "BIG DATA ANALYTICS BIG DATA A N A LY TIC S TDWI BEST PRACTICES REPORT Introduction to Big Data Analytics," 2011.

[4] X. Dong *et al.*, "Knowledge vault," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 601–610.

[5] L. Li, Y. Yao, J. Tang, W. Fan, and H. Tong, "QUINT," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016, pp. 985–994.

[6] Y. Zhou, L. Liu, and D. Buttler, "Integrating Vertex-centric Clustering with Edge-centric Clustering for Meta Path Graph Analysis," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 1563–1572.

[7] B. Bi, H. Ma, B.-J. (Paul) Hsu, W. Chu, K. Wang, and J. Cho, "Learning to Recommend Related Entities to Search Users," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 2015, pp. 139–148.

[8] R. Blanco, G. Ottaviano, and E. Meij, "Fast and Space-Efficient Entity Linking for Queries," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 2015, pp. 179–188.

[9] K. L. Caballero Barajas and R. Akella, "Dynamically Modeling Patient's Health State from Electronic Medical Records," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 69–78.

[10] M. Kokkodis, P. Papadimitriou, and P. G. Ipeirotis, "Hiring Behavior Models for Online Labor Markets," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 2015, pp. 223–232.

[11] N. Zalmout and M. M. Ghanem, "Multivariate adaptive community detection in Twitter," *Int. J. Big Data Intell.*, vol. 3, no. 4, p. 239, 2016.

[12] T. P. Oliveira, J. S. Barbar, and A. S. Soares, "Computer network traffic prediction: a comparison between traditional and deep learning neural networks," *Int. J. Big Data Intell.*, vol. 3, no. 1, p. 28, 2016.

[13] N. Makrynioti *et al.*, "PaloPro: a platform for knowledge extraction from big social data and the news," *Int. J. Big Data Intell.*, vol. 4, no. 1, p. 3, 2017.

[14] G. Dimitrakopoulos, V. Chatzigiannakis, and L. Tsitouras, "A knowledge-based integrated framework for increasing social management intelligence," *Int. J. Big Data Intell.*, vol. 4, no. 1, p. 36, 2017.

[15] S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil, "People on drugs," in

Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14, 2014, pp. 65–74.

[16] V. Shashidhar, N. Pandey, and V. Aggarwal, "Spoken English Grading," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 2089–2097.

[17] Z. Chen and B. Liu, "Mining topics in documents," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 1116–1125.

[18] S. Wang, Z. Chen, G. Fei, B. Liu, and S. Emery, "Targeted Topic Modeling for Focused Analysis," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, pp. 1235–1244, 2016.

[19] T. Kurashima, T. Iwata, N. Takaya, and H. Sawada, "Probabilistic latent network visualization," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 1236–1245.

[20] E. Schubert, M. Weiler, and H.-P. Kriegel, "SigniTrend," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 871–880.

[21] M. Nagarajan *et al.*, "Predicting Future Scientific Discoveries Based on a Networked Analysis of the Past Literature," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 2015, pp. 2019–2028.

[22] B. Geerdink, "A reference architecture for big data solutions - introducing a model to perform predictive analytics using big data technology," *Int. J. Big Data Intell.*, vol. 2, no. 4, p. 236, 2015.

[23] M. Papaioannou, E. Plum, E. T. F. Rogers, J. Valente, and N. I. Zheludev, "All-Optical Image Recognition Using Metamaterials," in *Frontiers in Optics 2016*, 2016, p. FF5G.7.

[24] K. Ganesan and C. Zhai, "Opinion-based entity ranking," *Inf. Retr. Boston.*, vol. 15, no. 2, pp. 116–150, Apr. 2012.