


Ontology-based Approach for Business Opportunities Recognition

Vinicius Ferreira Salgado¹, Diego Bernardes de Lima Santos¹,
Frederico Giffoni de Carvalho Dutra²^a, Fernando Silva Parreiras³^b
and Wladimir Cardoso Brandão¹^c

¹Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil

²Companhia Energética de Minas Gerais (CEMIG), Belo Horizonte, Brazil

³Laboratory for Advanced Information Systems, FUMEC University, Belo Horizonte, Brazil

Keywords: Business Opportunity, Information Extraction, Web Crawler, Ontology.


Abstract: The Web is the main source of business related information due to its accessibility, diversity and huge size, resulting from the high degree of collective engagement. However, extracting relevant information from this vast environment to use in decision making by organizational staff is a great challenge. In particular, the gathering of information related to business opportunities and its effective treatment to extract pieces of useful information to predict consumer and market behavior is essential for organizational survival. Although some approaches for handling business related information from Web have been proposed in literature, they under exploit contextual semantic patterns for information extraction, e.g., the set of properties related to the business opportunity topic. The present article proposes an ontology-based approach to recognize business opportunities from business related news extracted from the Web. Experimental results show that of our approach can effectively recognize business opportunities, reaching up to 90% of accuracy.


1 INTRODUCTION


The leading search engine has reported the index of more than one trillion uniquely addressable documents and the processing of more than 40 thousand user queries each second (Cutts, 2012). The accessibility, diversity and massive-scale nature make the Web the main source of information and services for business, demanding efficient approaches to retrieve and filter relevant content for decision making by organizational staff. Particularly, organizations increasingly use the Web to improve their performance, mining business opportunities that can result in higher profits. The United States Federal Trade Commission¹ defines a business opportunity as an investment that allows the beginning of a business, usually involving the sale or lease of a product or service that enable the purchaser-licensee to begin a business. Broadly, a business opportunity is a situation in which

it is possible for people and enterprises to buy, sell, exchange, lease or acquire products and services.

Crawling business information from the Web with no proper treatment does not provide organizations with the capacity to make assertive decisions. Additionally, crawling business information from Web pages demands filtering those exclusively related to business opportunities, recognize entities related to the opportunities, estimate important missing values and rank the opportunities according to a business criteria that optimize decision making by the organizational staff. Therefore, an ontology-based approach to extract information related to business opportunities from Web is paramount for organizational effectiveness. In this article, one proposes BOR, an ontology-based approach for business opportunities recognition. Particularly, the ontology is used to drive the extraction of information on business opportunities from business related news collected from Web. One assesses the effectiveness of the proposed approach by using it in a real organizational scenario to build a business opportunity dataset for a large Brazilian energy company. The main contributions of this article are: i) BOR, an ontology-based approach for busi-

^a <https://orcid.org/0000-0002-8666-0354>

^b <https://orcid.org/0000-0002-9832-1501>

^c <https://orcid.org/0000-0002-1523-1616>

¹<https://www.ftc.gov>

ness opportunities recognition from business news; ii) a throughout evaluation of BOR by assessing its effectiveness in a real organizational scenario and; iii) a business opportunity dataset with public business news extracted from the Web.

The present article is organized as follows: Section 2 presents literature review. Section 3 presents related work. Section 4 presents OntoBE, the ontology used by the proposed approach to drive the business opportunities recognition process. Section 5, presents BOR, the ontology-based approach to recognize business opportunities from business related news. Section 6 presents the experimental setup, followed by experimental results in Section 7. Finally, Section 8 concludes, pointing directions for future work.

2 BACKGROUND

2.1 Web Crawlers

Information retrieval (IR) is the research field that investigates the representation, storage, organization and access to information items to provide easy access for users to information of their interest (Baeza-Yates and Ribeiro-Neto, 2011). In particular, information items typically correspond to text documents. Thereby, an information retrieval system (IRS) retrieves relevant documents related to a user information need expressed by a query. IRS should not only to decide which documents to retrieve, or how to extract relevant pieces of information from such documents, but mainly to decide what is relevant for users (Brandão et al., 2014).

Web crawlers collect documents from the Web as fast as possible to build a comprehensive local corpus of documents (Pant et al., 2004). For this, they send requests for documents to web servers and process the responses to download and store the collected documents into a corpus (Dr. K. Iyakutti, 2017).

2.2 Web Extractors

Information extraction (IE) is the task of automatically extracting structured information from unstructured or semi-structured documents (Yu et al., 2014). In most cases, this task concerns processing human language through natural language processing (NLP). In particular, an information extraction system (IES) transforms the raw material collected by IRS, refining and reducing it to a germ of the original text. It starts with a collection of relevant text, such as newspaper and journal articles, transforming them into information that is more readily digested and analyzed by

isolating relevant text fragments, extracting relevant information from the fragments and piecing together targeted information in a coherent framework (Cowie and Lehnert, 1996). Thereby, web extractors are IES that extracts pieces of information from web documents. While web crawlers collect documents from Web, web extractors retrieves pieces of relevant information from collected documents.

One of the most known application of web extractors is the extraction of entities, such as persons, organizations, locations and monetary values, from text documents, a task called Named Entity Recognition and Classification (NERC) (Nadeau and Sekine, 2007).

2.3 Text Classification

Text classification or categorization, and document classification or categorization, refer to the task of assigning a document, or a piece of text, to one or more classes or categories (Mladeni et al., 2010). Text classification can be performed either through manual annotation or by automatic labeling. With the growing scale of text data in industrial applications, automatic text classification is becoming increasingly important (Minaee et al., 2020). They are particularly useful in the processing of information extracted from the Web, composed of a huge volume of textual documents. Text classification is also a challenging task due to the great diversity of languages and dialects (Lai et al., 2020).

Usually, text classification approaches are composed by distinct components, such as text extractors, content transformers, classifiers and evaluators. In particular, the initial entry consists of a set of raw text data $D = \{d_1, d_2, \dots, d_N\}$ usually extracted from a set of unstructured documents (Aggarwal and Zhai, 2012). Then a set of transformation procedures, referred as text pre-processing, can be performed on D , such as tokenization, lower case conversion, special character removal, accent replacement, stop word removal, stemming and lemmatization. Next, a classifier implemented from a large set of classification algorithms is applied to the D to determine the classes associated with each element. Finally, the classifier's effectiveness is estimated using statistic procedures and effectiveness metrics (Aggarwal and Zhai, 2012).

3 RELATED WORK

The h-TechSight system detects changes and trends in business related information to monitor business markets over a period of time (Kokossis et al., 2005). The

authors present experimental results focusing in job advertisement and argue that their system has been tested by real users in industry, increasing the efficiency of acquiring knowledge and supporting industry projects. In the same vein, an ontology-based approach for information extraction (Saggion et al., 2007) focus in e-business, recognizing relevant concepts in business documents, such as acquisition, partnerships, contracts and investments. The authors present an automatic annotator of information extracted from the Web that identifies the link between ontology and annotated text, and an application that extracts business related information per location.

The MBOI approach (Bai et al., 2004; Tajarobi et al., 2005) collects and classifies business related documents (calls for tenders) from Web by using a classification model based on language modeling with unigrams. The authors argue that MBOI was used by several companies that reported a significant improvement in their business activities. In an extended work (Paradis et al., 2005) the authors improve MBOI by incorporating text entity extractors for locations, organizations, dates and money.

Web news portals expose their content through different platforms to reach more readers, such as social networks. Thus, the search for information on these platforms becomes an interesting objective. For instance, business opportunities can be found on several crowdfunding sites. Crowdfunding is the practice of funding a venture by raising money from a large number of people. Kickstarter (An et al., 2014) crawls business posts from Twitter to recommend investors for crowdfunding projects. The recommendation approach uses text classification with supervised learning to evaluate the content of tweets, resulting in a effective textual classification.

Similarly to the other approaches previously reported in literature (Duarte et al., 2007; Pirovani and Oliveira, 2018) where the authors present an outperforming effectiveness of web extractors for NLP tasks in Portuguese, in this article one address the problem of crawling business related information from Portuguese news in the Web and extract from them business related information. But different from them, one use an enterprise ontology to drive the crawling and extraction of business opportunities.

4 THE ONTOBE ONTOLOGY

Ontologies enable the explicit definition of the logical structure of concepts and their relationships to generate common understanding about a specific topic, also reducing development time and cost for the topic

modeling and improving data quality. An ontology should be independent from a computer or social context, be consistent and also hold the lowest number of claims (Gruber, 1993). The BOR approach uses the OntoBE ontology (Falci et al., 2020) to drive the business opportunities recognition process. OntoBE is applied on the crawler, filter and in the classification steps, providing improvements in the accuracy of these components. The ontology does not represent all business strategies, but rather the universe of business based on new developments and expansion of productive capacity, resulting in news which is more relevant for the business opportunities context. Figure 1 presents the main concepts and conceptual relationships provided by OntoBE.

OntoBE focuses in relationships between people, processes and technologies and was designed to support decision making, such as the case of customers needing to search for business expansion attributes and researchers needing to provide statistics and explain interactions of people and processes of business expansion information gathering. OntoBE addresses particular functional requirements, such as general news, information on companies and industrial sectors, company directories, product, biographical, financial, investment, legal and statistical information, and market research. The information on companies and industrial sectors corresponds to academic journals related to business publications, financial publications and newspapers dedicated to business. As previously mentioned events covered by OntoBE potentially incorporate business opportunities. For instance the acquisition, merge or expansion of a company, the growth happened due to the financial amount invested increasing its production capacity possibly generating new jobs, aiming an improvement in revenue.

5 THE BOR APPROACH

Figure 2 presents the BOR architecture and from it one observe that first, the crawler component extracts news from Web,s starting with a list of seed URLs. The URLs in the list contain addresses of business related sites and portals. The crawler component extracts information such as author, date of last modification, description, text, title and the URL of the crawled news and store them in the Business News database. Particularly, the database also stores all the original news. The OntoBE provides business concepts that supply the crawling strategy.

Second, the filter component performs the following procedures over the data stored in the Business News database: removal of duplicate links, removal

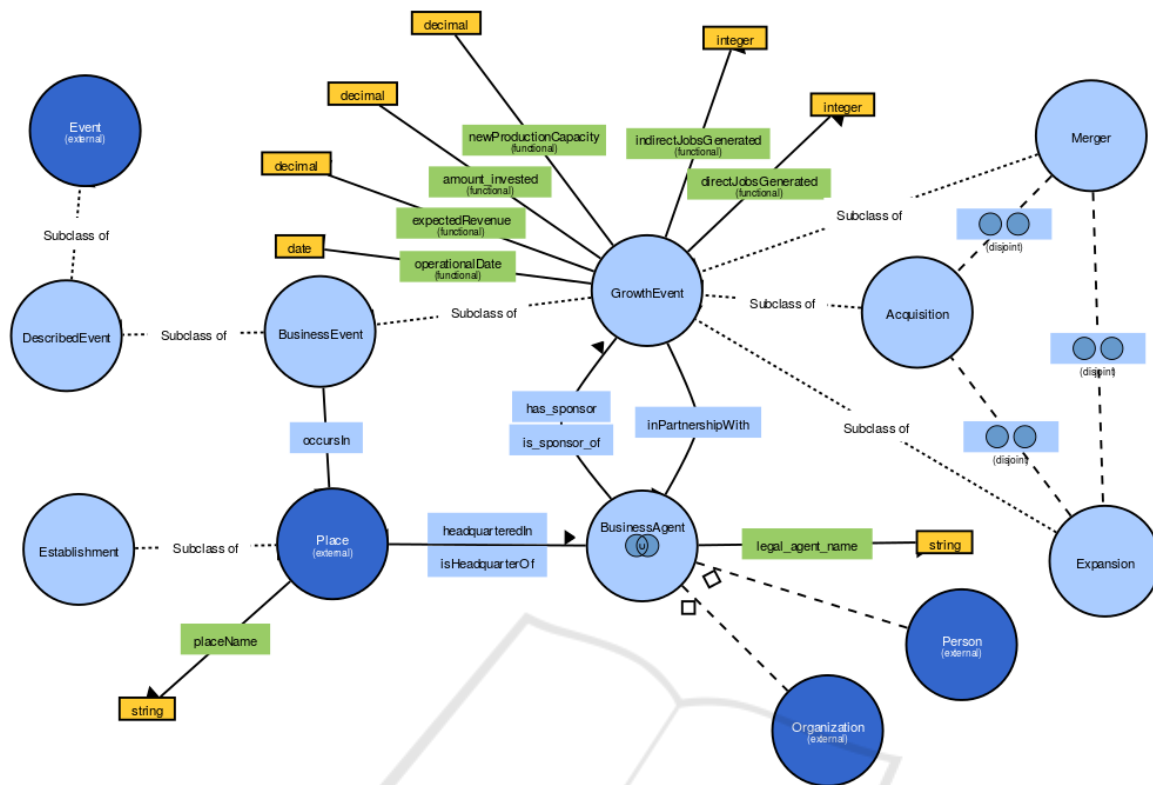


Figure 1: The OntoBE ontology.

of invalid links, removal of news that doesn't have title or full text. The objective of this step is to build a pre-processed news repository with relevant content that can be effectively used for text classification. All filtered content are stored in the Business Content repository. In particular, the Business Content repository stores the title, description, URL, status, the queue for adding related entities and the news full text. Once again, the OntoBE provides business concepts and taxonomy that supports the content filtering.

Third, the classifier component performs the classification of news. In particular, it learns how to classify business news from labeled examples extracted from the Business Content repository. For the initial classification, the news was manually labeled in order to enrich the automatic classification. The manual classification were made by specialists and analyzes the words contained in the title, description and full text. After labeling, the text is classified through the text classification technique and stored in Business Opportunities News for future entity extraction.

Finally, the extractor component recognize entities in business opportunities news. The objective of this step is to extract OntoBE entities from classified news to highlight business opportunities.

6 EXPERIMENTS

This section presents the experiments one carried out to evaluate the BOR approach, including experimental setup and procedures. In particular, the experiments answer the following research questions: i) How effective is BOR to collect and filter business related news from Web? ii) Which textual features provide positive impact on the classification performance? iii) How does BOR perform to recognize business opportunities from business related news?

6.1 Crawler Setup

The crawler component uses two different strategies to collect news from Web: vertical and horizontal. The vertical strategy uses a seed of Brazilian news portals, going in-depth search on the sites. The horizontal strategy uses a set of keywords to collect news pages from Web by using a programmable search engine². The domains for the vertical crawling and the keywords for the horizontal crawling related to business related news were constantly evaluated and updated to enrich the Business News database. In partic-

²<https://developers.google.com/custom-search>

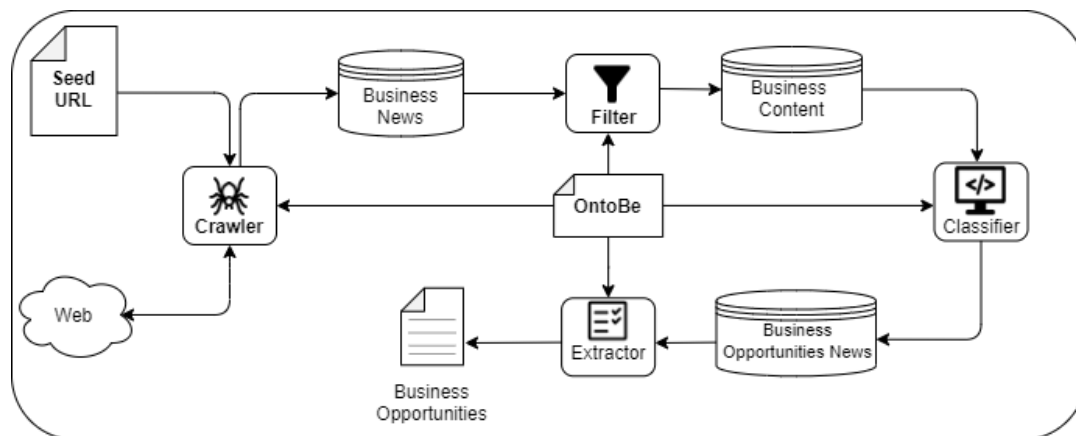


Figure 2: The BOR architecture.

ular, domains are automatically changed to get news in sites where there are more business related news. Keywords are automatically updated according to the most used terms in business related news. The crawler is able to search for data and recognize content fields such as author, download date, modified date, published date, description, filename, image URL, language, source domain, page title, URL and full text.

6.2 Filter Setup

In the filtering step one define requirements for relevant news to be recognized as business related news. A reference database provided by R&D CEMIG was used for this. It contains business related news as defined by OntoBE presented in Figure 1 and follows the definition of business opportunity. In particular, first a check is performed on the language that must be set as Portuguese, to get news in Portuguese. Then, the process described in Section 5 its applied. Finally, the news is stored in the Business Content repository with ID: unique identifier for the news, the same one defined on the Business News database; Title: the title of the news; Description: the description of the news; Full Text: the full text (the entire body) of the news; URL: the news URL from where it was crawled; Status: the news class to be set by the classifier, e.g. “null” for non-classified, positive for business related news and negative for business unrelated news; Entities: a list with entities to be extracted from the news.

6.3 Classifier Setup

The classifier is responsible to associate news to classes based on the news content and considering the business context. News text should go through a process of normalization. These process consists of

steps as show in Section 2.3, also removing the quotation marks that is commonly used in Portuguese, besides removing large spaces between text segments and also line breaks. Particularly, the classification evaluation can be defined as follows: 1) Reading and processing the news in the Business Content repository; 2) Evaluation with four different sources of textual features extracted from the news: i) Title; ii) Title + Description; iii) Title + Full Text; iv) Title + Description + Full text; 3) Evaluation of the effectiveness of each classifier with different training and test configurations; 4) Identification of the most effective approaches to identify business related news.

The classification process is carried out initially with a quantity of news and, to simulate the operating environment, it was necessary to include more news and perform the reclassification, with the intention of leaving the classifier with a better accuracy. To evaluate the classification, different algorithms were used to generate the models, particularly the algorithm RF (Random Forest) with the maximum depth of the tree in 5, the number of trees in the forest in 10 and 1 of features to consider when looking for the best split. The algorithm SVM (Support Vector Machine) with linear kernel with tolerance for stopping criteria in $1e-4$ and the algorithm NN (Neural Network) implemented using a multi-layer perceptron with regularization L2 in 1 and with a maximum number of interactions in 1000. Finally, after the classification, the news are available for the extraction of entities.

6.4 Extractor Setup

The extractor component is responsible to recognize different entities in the news such as organizations, people and locations. The process consists of extracting entities based on the OntoBE ontology, where they are initially extracted as common entities, for ex-

ample, investment extraction begins with three extractions ('\$', 'SYM'), ('26', 'NUM'), ('billion', 'NUM') resulting in the entity ('\$26 billion', 'AMOUNT INVESTED'). For a better understanding, the extraction process can be defined as follows: 1) Reading the pre-classified business related news; 2) Extract entities driven by the OntoBE ontology, resulting in business opportunities entities; 3) Evaluation of the effectiveness of the entity extractor.

7 EXPERIMENTAL RESULTS

This section presents the experimental results that support the answers for the three research questions presented in Section 6. In particular, one build different datasets to support experiments³, presenting experimental results of two cycles of tests.

7.1 Crawling Effectiveness

In this section, one address the first research question by assessing the effectiveness of the crawler and filter components. The average crawling rate was 150 documents per minute and the requests were distributed in time to avoid web servers overload. Table 1 presents the amount of news in each cycle.

Table 1: Number of news.

Process	Quantity	
	First Cycle	Second Cycle
Crawling	1,466	10,834
Filtering	504	3,702
Valid links	962	7,132

The removal step consists of eliminating some news after applying the filter. After this process the result of valid links is obtained. Then, they will initially be classified manually, in order to create a labeled basis for the classification process.

7.2 Classification Effectiveness

In this section, one address the second research question, assessing approaches to classify business related news. Table 2 and 3 show the accuracy in first cycle for each algorithm used to filter news with different resources for each training and test configuration. One evaluates four different sources of textual resources extracted from the business related pages: i) Title (TO); ii) Title + Description (TD); iii) Title + Full Text (TF); iv) Title + Description + Full

³<http://doi.org/10.5281/zenodo.4019968>

text (ALL). Significance is verified with a two-tailed paired t-test (Jain, 1991), with the symbol ▲ (▼) denoting a significant increase (decrease) at the $p < 0.05$ level, and the symbol ● denoting no significant difference.

The result of first cycle shows that the RF classifier is less effective than the others, as it contains few words related to the business context. In addition, Linear SVM classifier outperforms the others with an accuracy from 94% to 100%, depending on the number of instances used in the training. In order to carry out the evolution of the classification by adding new business related news, Table 4 and 5 present the results for the second cycle. In this scenario, the Linear SVM classifier outperforms the other classifiers. These results show that the algorithm is capable of associating relevant words to the business context. Remembering our goals, these observations attest to the effectiveness of our approach to collect and filter business related news.

7.3 Extractor Effectiveness

In this section, one address the third research question by assessing the effectiveness of our approach to recognize entities in business related news. In particular, BOR uses a supervised learning model to recognize entities in Portuguese.

The result of extraction shows that not all entities were extracted from news related to business. The amount invested and the local information are the most important for classified news. To circumvent the problem of null fields, the database "business entities", as shown in Figure 2, stores all news with their respective entities. Table 6 shows the accuracy for the entities of OntoBE. The information as investing company (*legal_agent_name*) and local (*placeName*) is present in more than 90% of the news. Both information is crucial for further research, if necessary. The information from entities about the expected return on investment, the beginning of the new operation and new production capacity impacted the final accuracy by not containing much information in the news. The news available does not have details necessary to fill these entities. For the BOR architecture, the information extracted according to the OntoBE ontology reaches an accuracy of 36% of 423 news classified as business related.

8 CONCLUSION

This article introduced BOR, the ontology-based approach to recognize business opportunities from busi-

Table 2: Classification accuracy in the first cycle of news of TO and TD.

Train/Test	TO			TD		
	RF	SVM	NN	RF	SVM	NN
90/10	0.6610 (±2.93e-5)	0.8252 (±1.57e-5) ▲	0.8252 (±1.57e-5) ▲●	0.6184 (±6.63e-6)	0.8463 (±4.01e-5) ▲	0.8463 (±4.01e-5) ▲●
80/20	0.6461 (±2.55e-6)	0.8860 (±8.39e-6) ▲	0.8860 (±8.39e-6) ▲●	0.6062 (±3.24e-7)	0.9276 (±1.27e-6) ▲	0.9174 (±3.15e-6) ▲▼
70/30	0.5950 (±2.12e-6)	0.8958 (±3.39e-6) ▲	0.8709 (±3.15e-6) ▲▼	0.5847 (±7.28e-7)	0.9655 (±1.90e-6) ▲	0.9344 (±1.49e-6) ▲●
60/40	0.5843 (±9.61e-7)	0.9142 (±1.19e-6) ▲	0.9090 (±1.58e-6) ▲▼	0.5922 (±4.03e-7)	0.9714 (±6.47e-7) ▲	0.9688 (±6.65e-7) ▲●
50/50	0.5881 (±1.99e-6)	0.9296 (±4.98e-7) ▲	0.9230 (±5.87e-7) ▲●	0.6154 (±3.24e-6)	0.9631 (±9.84e-8) ▲	0.9730 (±3.39e-7) ▲▼
40/60	0.5761 (±3.57e-8)	0.9446 (±5.29e-8) ▲	0.9308 (±1.65e-7) ▲▼	0.5847 (±8.56e-8)	0.9809 (±1.78e-7) ▲	0.9740 (±2.32e-7) ▲▼
30/70	0.5861 (±1.72e-8)	0.9450 (±1.29e-7) ▲	0.9371 (±1.20e-7) ▲●	0.5801 (±5.10e-9)	0.9822 (±5.64e-8) ▲	0.9792 (±1.15e-7) ▲▼
20/80	0.5766 (±3.20e-9)	0.9506 (±4.30e-8) ▲	0.9046 (±2.25e-7) ▲▼	0.5805 (±3.56e-8)	0.9844 (±3.57e-8) ▲	0.9597 (±7.25e-8) ▲▼
10/90	0.5768 (±6.04e-8)	0.9572 (±5.47e-8) ▲	0.9491 (±2.79e-8) ▲▲	0.5848 (±2.20e-7)	0.9838 (±1.75e-8) ▲	0.9826 (±1.53e-8) ▲▼

Table 3: Classification accuracy in the first cycle of news of TF and ALL.

Train/Test	TF			ALL		
	RF	SVM	NN	RF	SVM	NN
90/10	0.6084 (±4.27e-6)	1.0000 ▲	0.9689 (±6.62e-6) ▲▼	0.6594 (±8.02e-6)	0.9794 (±6.50e-6) ▲	0.9689 (±6.62e-6) ▲▼
80/20	0.6164 (±6.24e-7)	0.9894 (±1.62e-6) ▲	0.9842 (±3.64e-6) ▲●	0.6112 (±1.05e-6)	0.9894 (±1.62e-6) ▲	0.9842 (±3.65e-6) ▲▼
70/30	0.6124 (±8.10e-6)	0.9896 (±1.43e-7) ▲	0.9826 (±2.37e-7) ▲●	0.6055 (±3.67e-7)	0.9861 (±9.64e-8) ▲	0.9792 (±9.36e-8) ▲●
60/40	0.5999 (±2.57e-6)	0.9896 (±1.23e-7) ▲	0.9870 (±1.75e-7) ▲●	0.6207 (±2.05e-6)	0.9896 (±1.23e-7) ▲	0.9870 (±1.75e-7) ▲●
50/50	0.5738 (±2.27e-7)	0.9874 (±1.38e-7) ▲	0.9895 (±1.21e-7) ▲●	0.5904 (±4.44e-7)	0.9916 (±1.38e-7) ▲	0.9895 (±1.21e-7) ▲▼
40/60	0.6282 (±2.81e-6)	0.9913 (±4.25e-8) ▲	0.9878 (±5.47e-8) ▲●	0.5899 (±5.79e-7)	0.9913 (±4.25e-8) ▲	0.9861 (±5.47e-8) ▲▼
30/70	0.5845 (±3.77e-8)	0.9881 (±2.02e-8) ▲	0.9865 (±2.94e-8) ▲●	0.5994 (±3.54e-7)	0.9880 (±2.00e-8) ▲	0.9865 (±2.94e-8) ▲●
20/80	0.5779 (±7.75e-9)	0.9883 (±1.09e-8) ▲	0.9844 (±2.02e-8) ▲▼	0.5818 (±4.65e-9)	0.9883 (±3.41e-8) ▲	0.9805 (±2.33e-8) ▲▼
10/90	0.5773 (±4.52e-8)	0.9895 (±2.05e-9) ▲	0.9884 (±1.03e-8) ▲●	0.5727 (±5.56e-9)	0.9895 (±7.23e-9) ▲	0.9884 (±1.03e-8) ▲●

Table 4: Classification accuracy in the second cycle of news of TO and TD.

Train/Test	TO			TD		
	RF	SVM	NN	RF	SVM	NN
90/10	0.9285 (±4.36e-09)	0.9453 (±4.64e-09) ▲	0.9341 (±6.51e-09) ▲▼	0.9285 (±4.36e-09)	0.9481 (±6.57e-09) ▲	0.9327 (±2.96e-08) ▲▼
80/20	0.9382 (±5.45e-10)	0.9628 (±5.48e-09) ▲	0.9445 (±1.41e-09) ▲▼	0.9382 (±5.45e-10)	0.9663 (±7.35e-09) ▲	0.9438 (±2.94e-09) ▲▼
70/30	0.9359 (±1.51e-10)	0.9672 (±1.65e-09) ▲	0.9434 (±1.75e-09) ▲▼	0.9359 (±1.51e-10)	0.9700 (±3.19e-09) ▲	0.9443 (±8.11e-10) ▲▼
60/40	0.9386 (±6.79e-13)	0.9715 (±2.97e-10) ▲	0.9442 (±7.93e-10) ▲●	0.9386 (±6.79e-13)	0.9750 (±4.71e-10) ▲	0.9466 (±4.07e-10) ▲▼
50/50	0.9441 (±1.75e-11)	0.9761 (±6.30e-10) ▲	0.9464 (±1.78e-10) ▲▼	0.9441 (±1.75e-11)	0.9797 (±1.02e-10) ▲	0.9483 (±1.34e-10) ▲▼
40/60	0.9410 (±1.11e-11)	0.9779 (±2.81e-10) ▲	0.9413 (±2.81e-11) ▲▼	0.9410 (±1.11e-11)	0.9808 (±1.79e-10) ▲	0.9422 (±1.14e-10) ▲▼
30/70	0.9404 (±5.53e-12)	0.9803 (±2.75e-10) ▲	0.9513 (±3.73e-10) ▲▼	0.9404 (±5.53e-12)	0.9813 (±1.05e-10) ▲	0.9541 (±2.36e-10) ▲▼
20/80	0.9400 (±4.85e-12)	0.9815 (±2.26e-10) ▲	0.9503 (±1.35e-10) ▲▼	0.9400 (±4.85e-12)	0.9852 (±1.76e-10) ▲	0.9540 (±2.47e-10) ▲▼
10/90	0.9407 (±3.34e-14)	0.9828 (±2.08e-10) ▲	0.9513 (±7.49e-11) ▲▼	0.9407 (±3.34e-14)	0.9859 (±1.26e-10) ▲	0.9544 (±5.42e-11) ▲▼

Table 5: Classification accuracy in the second cycle of news of TF and ALL.

Train/Test	TF			ALL		
	RF	SVM	NN	RF	SVM	NN
90/10	0.9285 (±4.36e-09)	0.9635 (±4.89e-09) ▲	0.9383 (±6.03e-08) ▲▼	0.9285 (±4.36e-09)	0.9593 (±4.61e-09) ▲	0.9397 (±5.14e-08) ▲▼
80/20	0.9382 (±5.45e-10)	0.9768 (±2.59e-09) ▲	0.9487 (±5.45e-09) ▲▼	0.9382 (±5.45e-10)	0.9810 (±4.54e-09) ▲	0.9530 (±9.49e-09) ▲▼
70/30	0.9359 (±1.51e-10)	0.9831 (±3.68e-09) ▲	0.9504 (±3.40e-10) ▲▼	0.9359 (±1.51e-10)	0.9840 (±4.89e-09) ▲	0.9541 (±3.86e-10) ▲▼
60/40	0.9386 (±6.79e-13)	0.9866 (±7.48e-10) ▲	0.9564 (±1.19e-09) ▲▼	0.9386 (±6.79e-13)	0.9876 (±7.80e-10) ▲	0.9554 (±9.27e-10) ▲▼
50/50	0.9441 (±1.75e-11)	0.9878 (±2.21e-10) ▲	0.9486 (±8.69e-10) ▲▼	0.9441 (±1.75e-11)	0.9889 (±1.73e-10) ▲	0.9483 (±4.97e-10) ▲▼
40/60	0.9410 (±1.11e-11)	0.9906 (±2.51e-10) ▲	0.9422 (±7.66e-11) ▲▼	0.9410 (±1.11e-11)	0.9899 (±2.21e-10) ▲	0.9443 (±2.19e-10) ▲▼
30/70	0.9404 (±6.41e-12)	0.9907 (±1.76e-10) ▲	0.9557 (±4.43e-10) ▲▼	0.9404 (±6.41e-12)	0.9901 (±9.03e-11) ▲	0.9593 (±4.70e-10) ▲▼
20/80	0.9400 (±4.85e-12)	0.9923 (±2.12e-10) ▲	0.9568 (±6.05e-10) ▲▼	0.9400 (±4.85e-12)	0.9924 (±1.85e-10) ▲	0.9572 (±2.19e-10) ▲▼
10/90	0.9407 (±3.34e-14)	0.9924 (±4.35e-11) ▲	0.9589 (±5.41e-10) ▲▼	0.9407 (±3.34e-14)	0.9923 (±5.79e-11) ▲	0.9620 (±3.19e-10) ▲▼

Table 6: Accuracy of entities.

Entity	Quantity	Accuracy
legal_agent_name	385	0.91
placeName	420	0.99
amount_invested	222	0.52
expectedRevenue	2	0.005
operationalDate	16	0.04
newProductionCapacity	23	0.05
directJobsGenerated	102	0.24
indirectJobsGenerated	45	0.11

ness related news from Web. Particularly, BOR is a supervised learning approach that exploits semantic features from Web news, automatically labeling page

content such as title, description and full text. In contrast to the supervised approaches in the literature, BOR not only exploits enterprise ontologies, but also uses OntoBE to drive the crawling, classification and extraction process. The proposed approach was evaluated using three classification steps, increasing the amount of news extracted by simulating a growth in news published on internet portals. The results of this assessment attest to the effectiveness of our learning approach for recognizing business opportunities, with effectiveness reaching up to 90%. In addition, when analyzing our extraction of entities, was demonstrate the robustness of the recognition of business opportunities where they obtained the information defined

by the BOR ontology. Finally, performance analyses on our classification methods, showed that they are particularly suitable for textual classification.

For the future, there is still room for further improvements, such as exploit deep neural network algorithms, such as those based in the Transformer architecture, for named entity extraction. Another plan is to assess the effectiveness of alternative learning techniques for textual classification, as well as the use of additional resources, particularly by adding new labeled bases.

ACKNOWLEDGEMENTS

The present work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Financing Code 001. The authors thank the partial support of the CNPq (Brazilian National Council for Scientific and Technological Development), FAPEMIG (Foundation for Research and Scientific and Technological Development of Minas Gerais), CEMIG, FUMEC, LIAISE and PUC Minas.

REFERENCES

- Aggarwal, C. C. and Zhai, C. (2012). *A Survey of Text Classification Algorithms*, pages 163–222. Springer US.
- An, J., Quercia, D., and Crowcroft, J. (2014). Recommending investors for crowdfunding projects. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW'14, pages 261–270.
- Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern information retrieval: the concepts and technology behind search*. Pearson Education.
- Bai, J., Paradis, F., and Yun Nie, J. (2004). Web-supported matching and classification of business opportunities. In *Proceedings of the 2nd International Workshop on Web-based Support Systems*, WSS'04, pages 28–36.
- Brandão, W. C., Santos, R. L. T., Ziviani, N., Moura, E. S., and Silva, A. S. (2014). Learning to expand queries using entities. *Journal of the Association for Information Science and Technology*, 9:1870–1883.
- Cowie, J. and Lehnert, W. (1996). Information extraction. *Communications of the ACM*, pages 80–91.
- Cutts, M. (2012). Spotlight keynote. In *Proceedings of Search Engines Strategies*, SES'12.
- Dr. K. Iyakutti, J. U. (2017). Mining association rules for web crawling using genetic algorithm. *International Journal of Engineering and Computer Science*, pages 2635–2640.
- Duarte, J., Cavalcante, R., and Milidiú, R. (2007). Machine learning algorithms for portuguese named entity recognition. *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial*, pages 67–75.
- Falci, D., Dutra, F., Brandão, W., Ferreira, E., and Parreiras, F. (2020). Integrating ontologies for business expansion information gathering. In *Proceedings of the 13rd Brazilian Seminar on Ontologies*, ONTOBRAS'20.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, pages 199–220.
- Jain, R. (1991). *The art of computer systems performance analysis - techniques for experimental design, measurement, simulation, and modeling*. Wiley-Interscience.
- Kokossis, A., Bañares-Alcántara, R., Jiménez, L., and Linke, P. (2005). h-TechSight: a knowledge management platform for technology intensive industries. In *Proceedings of the 38th European Symposium on Computer-Aided Process Engineering*, ESCAPE'05, pages 1345–1350.
- Lai, Y.-A., Zhu, X., Zhang, Y., and Diab, M. (2020). Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, LREC'20, pages 1739–1746.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2020). Deep learning based text classification: A comprehensive review. *CoRR*, abs/2004.03705.
- Mladeni, D., Brank, J., and Grobelnik, M. (2010). *Document Classification*, pages 289–293. Springer US.
- Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, pages 3–26.
- Pant, G., Srinivasan, P., and Menczer, F. (2004). Crawling the Web. In *In Web dynamics: Adapting to change in content, size, topology and use*, pages 153–178. Springer-Verlag New York, Inc.
- Paradis, F., Nie, J.-Y., and Tajarobi, A. (2005). Discovery of business opportunities on the internet with information extraction. In *Proceedings of the Workshop on Multi-Agent Information Retrieval and Recommender Systems*, IJCAI'05.
- Pirovani, J. and Oliveira, E. (2018). Portuguese named entity recognition using conditional random fields and local grammars. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC'18.
- Saggion, H., Funk, A., Maynard, D., and Bontcheva, K. (2007). Ontology-based information extraction for business intelligence. In *Proceedings of the 6th International Semantic Web Conference*, ISWC'07, pages 843–856. Springer Berlin Heidelberg.
- Tajarobi, A., Garneau, J.-F., and Paradis, F. (2005). MBOI: Discovery of business opportunities on the internet. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, HLT-Demo'05, pages 30–31.
- Yu, H., Guo, J., Yu, Z., Xian, Y., and Yan, X. (2014). A novel method for extracting entity data from deep web precisely. In *Proceedings of the 26th Chinese Control and Decision Conference*, CCDC'14, pages 5049–5053.