A Cognitive Automation Approach for a Smart Lending and Early Warning Application

[Industrial and Application paper]

Ermelinda Oro
High Performance Computing and
Networking Institute of the National
Research Council
Altilia.ai
Rende (CS), Italy
linda.oro@icar.cnr.it

Massimo Ruffolo
High Performance Computing and
Networking Institute of the National
Research Council
Altilia.ai
Rende (CS), Italy
massimo.ruffolo@icar.cnr.it

Fausto Pupo
Altilia.ai
Rende (CS), Italy
fausto.pupo@altiliagroup.com

ABSTRACT

The rapid development of Internet and the dissemination of information and documents through a myriad of heterogeneous data sources is having an ever-increasing impact on the financial domain. *Corporate and Investment Banks* (CIBs) need to improve and automate business and decision-making processes simplifying the way they access data sources to get alternative data and answers. Manual or traditional approaches to data gathering are not sufficient to effectively and efficiently exploit information contained in all available data sources and represent a bottleneck to processes automation. This paper presents a *cognitive automation* approach, that makes use of Artificial Intelligence (AI) algorithms for automatically and efficiently searching, reading and understanding documents and contents intended to humans. The paper also presents the system that implements the proposed approach by an application in the area of financial risk evaluation and lending automation. The presented approach allows CIBs to obtain answers and analysis useful to improve the ability of different bank areas to manage lending processes, forecast situations involving risks, facilitate lead generation, and develop customized marketing and sales strategies.

KEYWORDS

Augmented Intelligence, Machine Reading Comprehension, Question Answering, Cognitive Automation, Heterogeneous Data, Financial Services, Smart Lending, Early Warning, Information Extraction, Natural Language Processing, Document Layout Analysis.

1 INTRODUCTION

Financial organizations are constantly looking for innovative ways to generate opportunities, automate and optimize business and decision-making processes, reduce risks and mitigate adverse events. In order to build long-term partnerships with their customers, the *Corporate and Investment Banks* (CIBs) need to develop customized marketing and sales strategies, and at the same time manage financial risks, based on a deep knowledge of corporate customers and markets in which they operate. The answers to CIBs' questions about the entities involved in the business and decision-making processes must be sought within a myriad of heterogeneous data sources. Financial markets change rapidly, therefore CIBs need to quickly process big data available in both traditional data sources (such as financial statements) and alternative sources of information (such as social and online media, corporate websites and online financial document repositories).

Traditional approaches are not sufficient to effectively and efficiently exploit information contained in these sources. Indeed, the ability to select, collect, analyze and interpret big data requires *Artificial Intelligence* (AI) algorithms capable of automatically and efficiently searching, reading and understanding documents and content designed for humans.

In this paper, we present a *cognitive automation* approach and the related system, along with a financial application, that automate and simplify business and decision-making tasks and processes requiring human cognitive abilities. Presented cognitive automation approach allows CIBs to obtain answers and analysis useful to improve the ability of different bank areas to manage lending processes, forecast situations involving risks, facilitate lead generation, and optimize sales activities. Examples of required answers, alternatively referred to as *data points* in this paper, are: entities and relationships between them, yes/no answers, sentiments, perceptions, and opinions.

The rest of the paper is organized as follows: Section 2 describes related work useful to comprehend modules of the proposed system. Section 3 introduces the proposed approach and related system. Section 4 presents how we solve some needs of CIBs by implementing a smart lending and early-warning application. Finally, section 5 concludes the work.

2 RELATED WORK

The proposed approach and system encompass strong and hard capabilities in *machine reading comprehension* (MRC) that exploit *pre-trained language models* and *human-in-the-loop* machine learning. In this section, we briefly review related work regarding these main aspects.

Machine Reading Comprehension. Machine Reading Comprehension (MRC) is the ability to answer questions asked in natural language by automatically reading from texts. The objective is to greatly simplify the way in which humans interrogate large volumes of information sources [3]. MRC is related to Natural Language Processing (NLP) and more specifically to Natural Language Understanding (NLU), which refers the ability of machines to understand natural language. NLU is considered an AI-hard problem and all its activities can be thought within a MRC framework [10]. MRC allows for exploring many aspects of language understanding, simply by posing questions. MRC can also be seen as the extended task of question answering (QA).

Recently, MRC methods have attracted a lot of attention among researchers and scholars around the world. Indeed, there have been many new datasets for reading comprehension developed in recent years, such as: SQuAD [22], NEWSQA [26], SearchQA [6], TriviaQA [9], HotpotQA [28], the latter requires multi-hop reasoning over the paragraphs, and ReCoRD [30] and COSMOS QA [8] that are designed for challenging reading comprehension with commonsense reasoning. However, these datasets mainly concern with understanding general text, and they are not related to specific knowledge domains. With deep learning (DL), end-to-end models have produced promising results on some MRC tasks. Unlike traditional machine learning, these models do not need to engineer complex features. Deep learning techniques for MRC have achieved very high performances on large standard datasets in general domains [4, 14, 29] and more recently, big successes have been obtained with approaches based on Pre-trained Language Models.

Pre-trained Language Models. We are entering the "Golden Age of NLP"[1]. With BERT of Google AI Language [5], initially published in 2018 as e-print version on ArXiv, which obtained outstanding performances in multiple NLP tasks (like sentiment analysis, question answering, sentence similarity), pre-training with fine-tuning has become one of the most effective and used method to solve NLP related problems. Compared to the word-level vectors (e.g. Word2Vec [13] released in 2013 and still quite popular, Glove [18], and FastText [1]) BERT trains sentence-level vectors and get more information from context. Before BERT, other pre-trained general language representations have been introduced. ELMO [19], which uses a bi-directional LSTM, generalizes traditional word embedding research along a different dimension extracting context-sensitive features. OpenAI GPT [21] demonstrates that greater results can be obtained by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task. ULMFiT [7] uses LSTM and produces contextual token representations. ULMFiT has been pre-trained from unlabeled text and fine-tuned for a supervised downstream task. Unlike previous papers, BERT uses a bi-directional Transformer. Transformers were introduced from Vaswani et at. [27]. After, a lot of BERT-based activities in natural language processing and understanding have shown even better results than BERT. The model ERNIE [31] is pre-trained by masking semantic units such as entity concepts, rather than tokens. Liu et al. [12] measure the impact of many key hyperparameters and training data size and present RoBERTa. Lan et al. [11] present ALBERT that implements two parameter-reduction techniques to lower memory consumption and increase the training speed of BERT. In this paper, we use a BERT-based MRC method that allows us the extraction of data points.

Human-in-the-Loop machine learning. Deep learning, in particular when it is applied to unstructured data, needs very large training sets to learn the parameters and hyperparameters, and the desired models [22]. Therefore, despite the obvious advantages of deep learning-based MRC systems, their use is often limited to an academic context where the performance of MRC techniques are tested on artificial datasets. These datasets poorly match the characteristics of the data of real business contexts, such as the financial sector, where a complex language with specialized terminology is used.

To facilitate the learning of MRC models in the financial domain, it is necessary to develop methods and interfaces for *human-in-the-loop machine learning*. Using these tools, humans can transfer domain knowledge to machines by annotating and validating datasets and models that can be used in the learning process. Currently, in the literature, there are some weakly supervised machine learning methods and systems that allow for creating annotated datasets from a *human-driven* perspective. For example, Snorkel[2] [23], based on data programming paradigm [24], is a recently proposed framework that enables users to generate large volumes of training data by writing labeling functions (such as rules and patterns) that capture domain knowledge. By using the data programming, such labeling functions can vary in accuracy and coverage, and they may be arbitrarily correlated. Other weakly supervised machine learning methods are for instance: Prodigy[3], Figure Eight[4], Amazon Mechanical Turk[5]. These methods can use and be combined with: (i) *transfer learning* [17] that exploits labeled data, parameters, or knowledge available in other tasks to reduce the need for labeled data for the specific new task, (ii) *active learning* [25] that select data points for human annotators to label, and (iii) *reinforcement learning* [20] that enables learning from feedback received through interactions with an external environment. Weakly human-driven methods can facilitate the adoption of MRC methods in complex domains, such as the financial one, in order to automate and simplify the extraction and interrogation of data of various formats in heterogeneous sources. For these reasons, in our approach, we implement human-driven annotation methods.

3 COGNITIVE AUTOMATION APPROACH

In this section we present the proposed approach useful to implement cognitive automation in decisional and operational business processes. Key steps of the presented approach are:

- Search, perform layout analysis, and classify documents.
- Dynamically exploit the knowledge of users for training and correction of the extraction algorithms thus enabling continuous learning.
- Extract answers about relevant questions concerning the entities involved in business processes by exploiting machines capabilities of reading and comprehend documents.
- Harmonize and store extracted information in knowledge graphs.
- Explore obtained information, and visualize synthetic and easily interpretable charts.

In the following, we describe modules of the system shown in figure 1 that implements the proposed approach.

Documents and Contents Gathering and Analysis. This module allows, through specific connectors and methods of web scraping and wrapping, the acquisition of heterogeneous contents and documents from different information sources. In order to obtain the machine-readable format, it processes image documents by using optical character recognition (OCR) algorithms. Then, it applies document layout analysis and understanding

[1] https://medium.com/@thresholdvc/neurips-2019-entering-the-golden-age-of-nlp-c8f8e4116f9d

[2] Snorkel https://www.snorkel.org/

[3] Prodigy https://prodi.gy

[4] Figure Eight https://www.figure-eight.com

[5] Amazon Mechanical Turk https://www.mturk.com

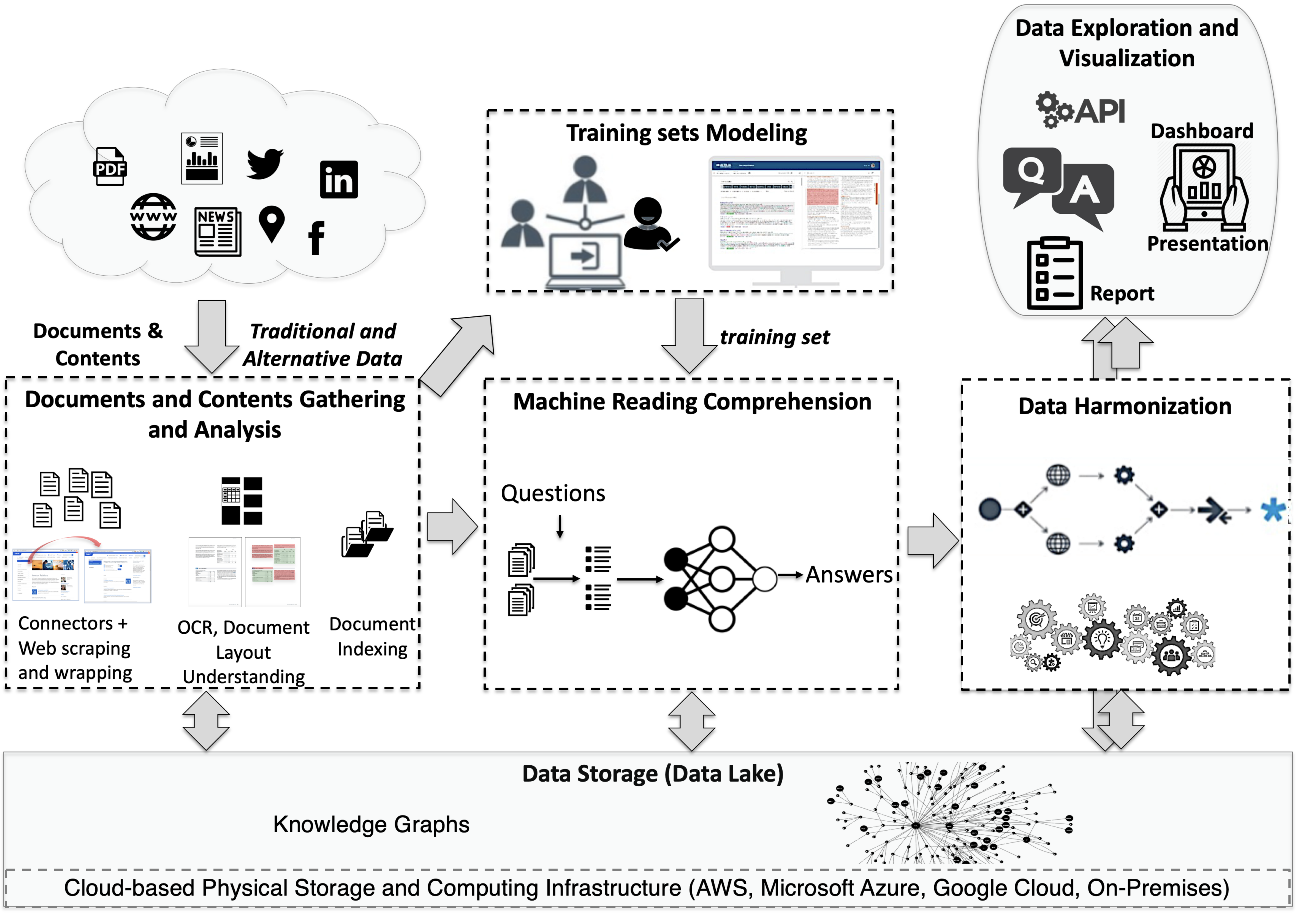

Data Exploration and Visualization
API
Dashboard
Presentation
Report
Training sets Modeling
Documents & Contents
Traditional and Alternative Data
training set
Documents and Contents Gathering and Analysis
Connectors + Web scraping and wrapping
OCR, Document Layout Understanding
Document Indexing
Machine Reading Comprehension
Questions
Answers
Data Harmonization
Data Storage (Data Lake)
Knowledge Graphs
Cloud-based Physical Storage and Computing Infrastructure (AWS, Microsoft Azure, Google Cloud, On-Premises)

Figure 1: Cognitive Automation System.

algorithms, also based on spatial reasoning [15, 16], to recognize structures of the documents (e.g.: columns, sections, tables, lists of records) and the reading order. Finally, the module enables for indexing documents and their portions.

Training sets Modeling. This module allows the *human-driven* annotation of portions of documents that answer specific questions exploiting a semi-automatic interactive and iterative process. This process involves the user by means of actions, mainly visual and/or based on simple rules, aimed at creating training sets for deep learning algorithms.

Machine Reading Comprehension (MRC). This module allows for learning models that extract data from documents in the form of answers to questions in natural language and it is based on different components:

(i) Retriever that selects a list of documents and portions that are most likely to contain the answer of a question obtained as input. It is implemented as a voting system that considers different versions of matching (e.g., based on Elasticsearch[6], DrQA [2] Reader that uses TF-IDF features exploiting uni-grams and bi-grams, and S-Reader [14] that uses different embeddings and hyperparameters with respect to DrQA).

(ii) Reader that takes as input the question and the portions chosen by the Retriever and outputs the most probable answers it can find. This sub-module is based on a pre-trained deep learning model. The model is essentially a PyTorch version of the well known NLP model BERT [5], which is made available by Hugging Face[7]. To fine-tune the model, created training sets in the modeling phase are exploited.

(iii) Selector that compares the answers' scores obtained by using an internal function and outputs the most likely answer according to the scores.

(iv) A graphical user interface that enables human-machine interaction used to implement reinforcement learning. By exploiting a graphical user interface that highlights results on portions of documents, users validate and give feedbacks to the deep learning algorithms that learn and improve performance by exploiting the user feedbacks.

Data Harmonization. This module enables the manipulation in a scalable way of data by using workflows based on Spark[8]. Workflows enable users to visually create complex processes that allows for gathering and processing data, performing data analysis, storing results in knowledge graphs, simply by combining and concatenating blocks. A block embeds algorithms that

[6] Elasticsearch https://www.elastic.co/

[7] Higging Face Transformers https://github.com/huggingface/transformers

[8] Spark https://spark.apache.org/

implement a specific task, for instance, the learned model for extracting data points, descriptive, predictive, and prescriptive analytics. For the same task different blocks that embed different logics (e.g.: various ways to collect data depending on the formats of sources) can be used.

Data Storage. Obtained results, including answers and metadata (e.g., the paragraphs where the answer was found and the title of the document), are stored into knowledge graphs (KGs). The current implementation of KGs is based on a multi-structured database that combines information retrieval capabilities with the ability to store data as graph databases.

Data Exploration and Visualization. Results can be explored through application programming interfaces (APIs) that allow integration with external applications, and they can be displayed in reports, dashboards, and presentations that visually track, analyze and show key performance indicators (KPI), metrics and key *data points*.

4 SMART LENDING AND EARLY WARNING APPLICATION

The rapid development of web content and the dissemination of information through social networks, blogs, and newspapers brought an ever-increasing impact on financial domain. How to rapidly and accurately mine the key information from big data is a challenging problem to study for researchers, and has become one of the key issues for investors and decision-makers. Indeed, the ability to automatically answer business questions enables *cognitive automation* in decisional and operational business processes in different *Corporate and Investment Banks* (CIBs) areas. CIBs need to decide if it is convenient to grant a loan to a company, to know the risk conditions of their customers portfolio, and to develop customized marketing and sales strategies. To this end, CIBs need to have a deep knowledge and to perform a careful evaluation of:

(i) corporate customers (such as, know board members, the environmental impact of the business, how they are perceived, the solidity of their business),
(ii) markets in which their customers are located and operate (e.g., solidity of the market, information about used commodities, competitors).

In practical terms, CIBs asked for a system capable to automatically: (i) answer to specific questions asked in natural language, i.e. extract *data points*, (ii) visualize queryable and navigable customer profiles that can be used for credit scoring, early warning, and marketing and sales activities.

In the following, we describe our solution that is based on the proposed approach and presented in the previous section 3.

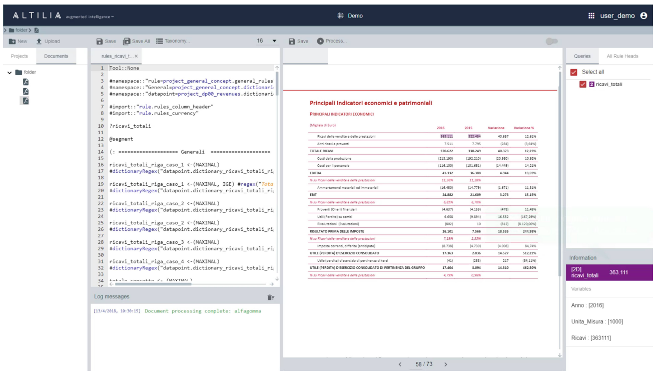
4.1 Documents and Contents Gathering and Analysis

Financial operators search for answers that can be obtained or inferred by reading and studying, even simultaneously, various information sources, such as financial documents (e.g., annual reports, 10-k forms, sustainability reports, notes to balance sheets), as well as web sources (e.g., news, blogs, social media). Examples of required answers (i.e., data points) are: the perception of a corporate brand on social media (customer brand perception), the geographical distribution of a company's debts, credits, and revenues, and the volume of R&D investments.

For the specific application, the proposed approach enables for extracting interesting data points in a scalable way from a huge amount of web sources related to a large number of companies. Data points enrich different aspects of customer profiles, such as Environment, Society, Governance (ESG) knowledge, which, for instance, can be used by credit scoring algorithms. The implemented web scraping tools are used to download news and financial documents from websites of companies or from the SEC (Securities and Exchange Commission) website, and to collect information and reviews from booking websites. These tools are flexible, easily configurable and maintainable. More in detail, the web scraping process consists of the definition of a configuration file for each different typology of websites to scrape. The wrapper uses DOM information and XPath along with similarities between different web sites reducing work needed to design wrappers. In addition, which kind of data/information to extract from the websites can be defined by a data model to fill. For instance, in order to collect information about restaurants and hotels the scraping tools navigate booking websites extracting reviews and attributes such as authors, title, date and all relevant info. In order to collect PDF documents (e.g., annual reports, 10-k forms, notes to balance sheets) the scraping tools navigate the companies' websites searching the sections investor relation and press release. Alternatively, the scraping tools download documents from financial document providers like SEC[9]. Downloaded PDF files are processed by using document layout analysis algorithms, even exploiting optical character recognition (OCR) techniques when needed, to extract portions (such as columns, paragraphs, tables, notes). Then, the different portions of documents (along with their relations, information of reading order, link to the original document and metadata) are stored in knowledge graphs and indexed in the system to be furthermore elaborated.

4.2 Training sets Modeling and MRC

During the training set modeling phase, a user can define labeling functions or visually annotate label-entity or question-answer pairs looking at input documents and information stored in the system. Figure 2 shows the graphical user interface that aids the creation of labeling functions.



```
Tool::None

#namespace::"rule=project_general_concept.general_rules
#namespace::"General=project_general_concept.dictionari
#namespace::"datapoint=project_dp00_revenues.dictionari

#import::"rule.rules_column_header"
#import::"rule.rules_currency"

?ricavi_totali

@segment

(: =================== Generali  ===================

ricavi_totali_riga_caso_1 <-(MAXIMAL)
#dictionaryRegex("datapoint.dictionary_ricavi_totali_ri

ricavi_totali_riga_caso_1 <-(MAXIMAL, IGE) #regex("Tota
#dictionaryRegex("datapoint.dictionary_ricavi_totali_ri

ricavi_totali_riga_caso_2 <-(MAXIMAL)
#dictionaryRegex("datapoint.dictionary_ricavi_totali_ri

ricavi_totali_riga_caso_2 <-(MAXIMAL)
#dictionaryRegex("datapoint.dictionary_ricavi_totali_ri

ricavi_totali_riga_caso_3 <-(MAXIMAL)
#dictionaryRegex("datapoint.dictionary_ricavi_totali_ri

ricavi_totali_riga_caso_4 <-(MAXIMAL)
#dictionaryRegex("datapoint.dictionary_ricavi_totali_ri

totale_concetto <- (MAXIMAL)
```

```
[13/4/2018, 10:30:15] Document processing complete: alfagomma
```

Principali Indicatori economici e patrimoniali

PRINCIPALI INDICATORI ECONOMICI

(Migliaia di Euro)	2016	2015	Variazione	Variazione %
Ricavi delle vendite e delle prestazioni	363.111	322.454	40.657	12,61%
Altri ricavi e proventi	7.511	7.795	(284)	(3,64%)
TOTALE RICAVI	370.622	330.249	40.373	12,23%
Costi della produzione	(213.190)	(192.210)	(20.980)	10,92%
Costi per il personale	(116.100)	(101.651)	(14.449)	14,21%
EBITDA	41.332	36.388	4.944	13,59%
% su Ricavi delle vendite e delle prestazioni	11,38%	11,28%		
Ammortamenti materiali ed immateriali	(16.450)	(14.779)	(1.671)	11,31%
EBIT	24.882	21.609	3.273	15,15%
% su Ricavi delle vendite e delle prestazioni	6,85%	6,70%		
Proventi (Oneri) finanziari	(4.637)	(4.159)	(478)	11,49%
Utili (Perdite) su cambi	6.658	(9.894)	16.552	(167,29%)
Rivalutazioni (Svalutazioni)	(802)	10	(812)	(8.120,00%)
RISULTATO PRIMA DELLE IMPOSTE	26.101	7.566	18.535	244,98%
% su Ricavi delle vendite e delle prestazioni	7,19%	2,35%		
Imposte correnti, differite (anticipate)	(8.738)	(4.730)	(4.008)	84,74%
UTILE (PERDITA) D'ESERCIZIO CONSOLIDATO	17.363	2.836	14.527	512,22%
Utile (perdita) d'esercizio di pertinenza di terzi	(41)	(258)	217	(84,11%)
UTILE (PERDITA) D'ESERCIZIO CONSOLIDATO DI PERTINENZA DEL GRUPPO	17.404	3.094	14.310	462,50%
% su Ricavi delle vendite e delle prestazioni	4,79%	0,96%		

☑ Select all
☑ 2 ricavi_totali

[2D] ricavi_totali 363.111

Anno : [2016]
Unita_Misura : [1000]
Ricavi : [363111]

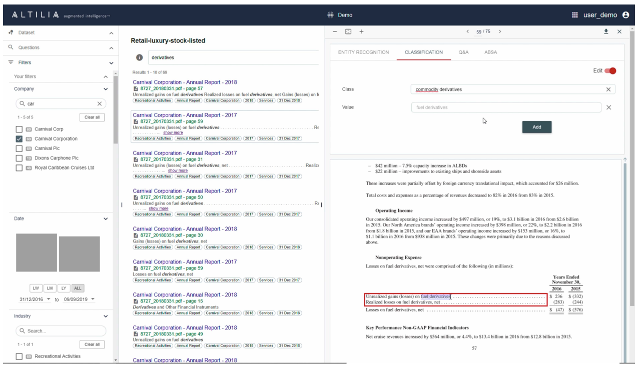
Figure 2: Labeling Functions GUI.

In the left part of the interface, the editor for defining labeling functions is displayed. These functions can exploit different

[9] https://www.sec.gov/edgar/searchedgar/companysearch.html

syntactic, spatial, and ontological information. In addition, they can use: (i) built-in that calls machine learning procedures or complex algorithms used as black-box, (ii) functions and concepts defined in other imported labeling files. The editor provides some facilities to simplify the writing of labeling functions exploiting relationships between label-value, titles-paragraphs or images-caption, table structures, and grammatical relationships like subject-verb-object (fact). At the upper right part of the interface, taxonomies of desired concepts to label are visualized. The GUI shows also the chosen PDF files used to visually evaluate the results of the executed labeling functions. Results details (attributes of the labeled concepts) can be visualized in the lower right-hand corner of the interface. In figure 2 the labeling functions annotate revenues in a financial statement.

In addition, as shown in figure 3, the GUI enables also to visually annotate texts, for instance, to assign labels, or to select answers of questions in the documents.



car

1 - 5 of 5

☐ Carnival Corp
☑ Carnival Corporation
☐ Carnival Plc
☐ Dixons Carphone Plc
☐ Royal Caribbean Cruises Ltd

31/12/2016 to 09/09/2019

1 - 1 of 1

☐ Recreational Activities

Retail-luxury-stock-listed

derivatives

Results 1 - 10 of 69

Carnival Corporation - Annual Report - 2018
8727_20180331.pdf - page 57
Unrealized gains on fuel *derivatives* Realized losses on fuel *derivatives*, net Gains (losses) on f
Recreational Activities | Annual Report | Carnival Corporation | 2018 | Services | 31 Dec 2018

Carnival Corporation - Annual Report - 2017
8727_20170331.pdf - page 59
Unrealized gains (losses) on fuel *derivatives* .... Re
show more
Recreational Activities | Annual Report | Carnival Corporation | 2017 | Services | 31 Dec 2017

Carnival Corporation - Annual Report - 2017
8727_20170331.pdf - page 31
Unrealized gains (losses) on fuel *derivatives*, net .... Realiz
show more
Recreational Activities | Annual Report | Carnival Corporation | 2017 | Services | 31 Dec 2017

Carnival Corporation - Annual Report - 2017
8727_20170331.pdf - page 50
Unrealized gains on fuel *derivatives* .... Re
show more
Recreational Activities | Annual Report | Carnival Corporation | 2017 | Services | 31 Dec 2017

Carnival Corporation - Annual Report - 2018
8727_20180331.pdf - page 30
Gains (losses) on fuel *derivatives*, net
Recreational Activities | Annual Report | Carnival Corporation | 2018 | Services | 31 Dec 2018

Carnival Corporation - Annual Report - 2017
8727_20170331.pdf - page 59
Losses on fuel *derivatives*, net
Recreational Activities | Annual Report | Carnival Corporation | 2017 | Services | 31 Dec 2017

Carnival Corporation - Annual Report - 2018
8727_20180331.pdf - page 15
Derivatives and Other Financial Instruments
Recreational Activities | Annual Report | Carnival Corporation | 2018 | Services | 31 Dec 2018

Carnival Corporation - Annual Report - 2018
8727_20180331.pdf - page 49
Unrealized gains on fuel *derivatives*
Recreational Activities | Annual Report | Carnival Corporation | 2018 | Services | 31 Dec 2018

Carnival Corporation - Annual Report - 2018

Class: commodity derivatives

Value: fuel derivatives

– $42 million – 7.5% capacity increase in ALBDs
– $22 million – improvements to existing ships and shoreside assets

These increases were partially offset by foreign currency translational impact, which accounted for $26 million.

Total costs and expenses as a percentage of revenues decreased to 82% in 2016 from 83% in 2015.

Operating Income

Our consolidated operating income increased by $497 million, or 19%, to $3.1 billion in 2016 from $2.6 billion in 2015. Our North America brands' operating income increased by $398 million, or 22%, to $2.2 billion in 2016 from $1.8 billion in 2015, and our EAA brands' operating income increased by $153 million, or 16%, to $1.1 billion in 2016 from $938 million in 2015. These changes were primarily due to the reasons discussed above.

Nonoperating Expense

Losses on fuel derivatives, net were comprised of the following (in millions):

	Years Ended November 30, 2016	2015
Unrealized gains (losses) on fuel derivatives	$ 236	$ (332)
Realized losses on fuel derivatives, net	(283)	(244)
Losses on fuel derivatives, net	$ (47)	$ (576)

Key Performance Non-GAAP Financial Indicators

Net cruise revenues increased by $564 million, or 4.4%, to $13.4 billion in 2016 from $12.8 billion in 2015.

57

Figure 3: A visual annotation of a concept related to the financial domain.

Created training sets are exploited within machine / deep learning algorithms, as described in section 3.

4.3 Data Harmonization and Storage

To scale-up KPIs extraction, a workflow can be designed, deployed in the cloud, and execute in parallel and scheduled way. In figure 4, the designed workflow enable to search and extract text portions from PDF documents related to the target questions.

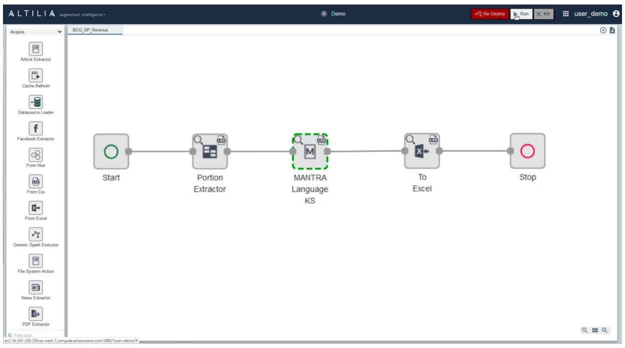


Figure 4: Workflow to scale-up KPIs extraction.

In the shown example, we are interested in extracting data points from balance sheets related to financial information (e.g., customer, industry, year, financial costs, commodities price, total revenues, EBIT, EBITDA) of more than 3000 companies. Information extracted are saved in knowledge graphs and can be provided in different formats selected by the customer (e.g., csv, excel, or json).

4.4 Data Exploration and Visualization

Banks are interested in creating reports, dashboards, and presentations to visualize customer profiles. In the following, we show some examples of dashboards and PowerPoint slides obtained by analyzing extracted data points related to a target company and considering peer companies used for benchmarking.

Figure 5 show a comparison of main financial data of the selected target client (e.g., revenue growth, EBITDA margin and growth, and net debt-to-EBITDA ratio) with the mean values of benchmarking companies (peers in the same industry of the target company). Target companies and peers can be dynamically selected to see real-time updates of charts.

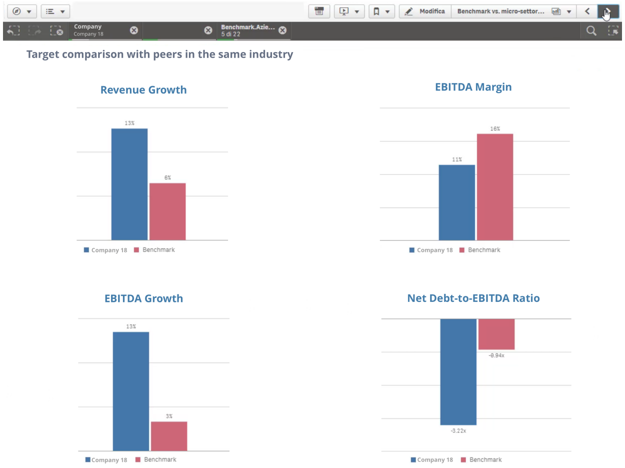


Figure 5: A comparison of main financial data between the target company and peers.

Figure 6 shows an exposure of the customer to financial risks in the form of a presentation slide exported by the system. In detail, the figure shows the exposure to interest rate changes on loan risk, to forex rates changes risk, and to commodity price variation risk also making comparisons with peers (benchmark percentages).

Figure 7 shows a deep dive on the foreign activities (forex risk) of the selected company (e.g.: revenue and credits/debit by country, non-euro revenues percentage of the total) and the forex derivatives it already has in its portfolio (i.e. derivatives usage table slitted for type of instruments).

5 CONCLUSION

In this paper, we presented a *cognitive automation* approach and the related system, along with a financial application, that enables CIBs to automatically: (i) extract data points from textual data sources, and (ii) visualize dashboards and presentations containing customers' data and comparisons between customers and their peers. The greater wealth and depth of information on risks and opportunities improve the ability to manage lending processes, provide real-time early warning, and help sales activities. In particular, the implemented solution enables: (i) Automatic, faster, and predictive credit/risk scoring (customer qualification) creation. (ii) Digitalization of lending processes (loans

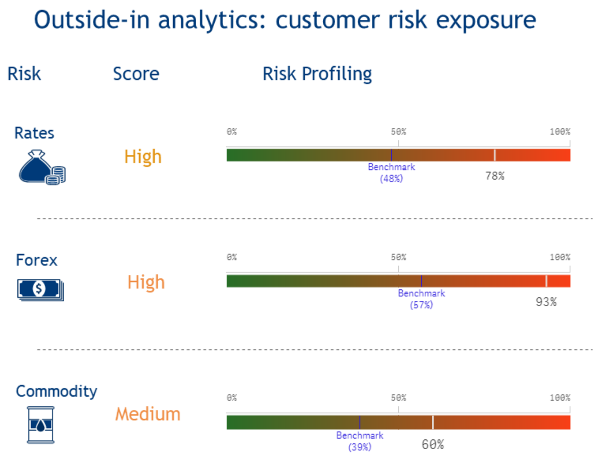


Figure 6: Exposure of a company to financial risks.

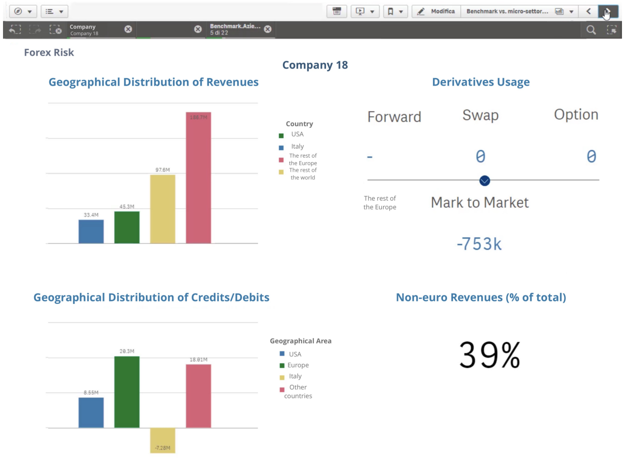


Figure 7: Deep dive on forex risk of a selected company

underwriting). (iii) Smarter and more effective early warnings. (iv) Reduction of losses due to unforeseen defaults. In this way, different areas of banks can benefit from developing customized marketing and sales strategies, as well as building efficient and effective lending processes, based on a deep knowledge of corporate customers and the market in which they operate.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[2] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *ICLR* (2017).

[3] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36 (2012), 1165–1188.

[4] Christopher Clark and Matt Gardner. 2017. Simple and Effective Multi-Paragraph Reading Comprehension. In *ACL*.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[6] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&amp;A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179* (2017).

[7] Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *ACL*.

[8] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *EMNLP/IJCNLP*.

[9] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*.

[10] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *International Conference on Machine Learning*. 1378–1387.

[11] Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv* abs/1909.11942 (2019).

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv* abs/1907.11692 (2019).

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[14] Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and Robust Question Answering from Minimal Context over Documents. *arXiv preprint arXiv:1805.08092* (2018).

[15] Ermelinda Oro and Massimo Ruffolo. 2017. A Method forWeb Content Extraction and Analysis in the Tourism Domain. In *International Conference on Enterprise Information Systems*, Vol. 2. SCITEPRESS, 365–370.

[16] Ermelinda Oro and Massimo Ruffolo. 2017. Object extraction from presentation-oriented documents using a semantic and spatial approach. US Patent 9,582,494.

[17] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22 (2010), 1345–1359.

[18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[19] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[20] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing* 2016, 1 (2016), 67.

[21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_ understanding_paper. pdf* (2018).

[22] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *EMNLP*.

[23] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment* 11, 3 (2017), 269–282.

[24] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*. 3567–3575.

[25] Burr Settles. 2009. *Active Learning Literature Survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[26] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A Machine Comprehension Dataset. In *Rep4NLP@ACL*.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*.

[28] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*.

[29] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *arXiv preprint arXiv:1804.09541* (2018).

[30] Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *ArXiv* abs/1810.12885 (2018).

[31] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL*.