05

BIG DATA ON MEDIA: NOT JUST BIG, BUT SMART

This article explores big data strategies in media industries and its users, suggesting that data, particularly news, is not only growing at a fast pace, but also getting smarter.

This article presents an innovative big data media project, “Máquina do Tempo”, a joint collaboration between SAPO and University of Porto, that led to an interactive online tool for navigation and exploration of 25 years of news articles, supported on automatically generated entities’ co-occurrences networks and rich profiles.

Jorge Teixeira *(jorge-teixeira@alticelabs.com)*
Marta Vicente Pinto *(marta-c-pinto@alticelabs.com)*

Big Data, Media, News, Machine Learning, Natural Language Processing, Information Extraction, Named Entity Recognition, Information Visualization, Text Mining, Computational Journalism

Introduction

The big data hype is said to present huge opportunities for CSPs. Most trends – in technology development, consumer behaviour, regulation and RDI investments forecasts – seem to corroborate that big data will play an important role in future revenue streams.

The advent of IoT, the "massification" of connected device ownership (either smartphones, tablets or other wearable devices), investments in ultra-fast networks and cloud enablement infrastructures put CSPs in a privileged position to lead this hype and start innovating with data. Having been transforming its business to adapt to the digital economy needs and to the virtualization and automation trends, CSPs do have the resources – Human and data sources- available to engage in this new knowledge field, though investment in developing and acquiring specific skills is obviously needed.

Big data *per se* has little value… but data processing applied to societal, business or organizational challenges can open a whole new business stream for CSPs: in business predictive analysis (advanced business intelligence); e-health (pandemic pattern detection); civil protection (early detection mechanisms); context aware marketing; available media information processing (extract the coherence and correlations among the amounts of media information available) or even typical CSP product enhancement (TV recommendations based on previous video consumption). Although having access to a huge amount of near real-time data, CSPs must comply with strict regulatory standards of privacy and security, legislation that is not applicable to OTTs, who use their client's data more or less freely, being able to deliver value added and tailored offerings to its clients. Besides this, legislation is being reshaped to include the new content delivery trends, which is good news for CSPs, who have the most to gain in this big (data) game.

CSPs have constant access to network traffic data, client consumption of data, a huge network of sensors and connected things and, now, how do

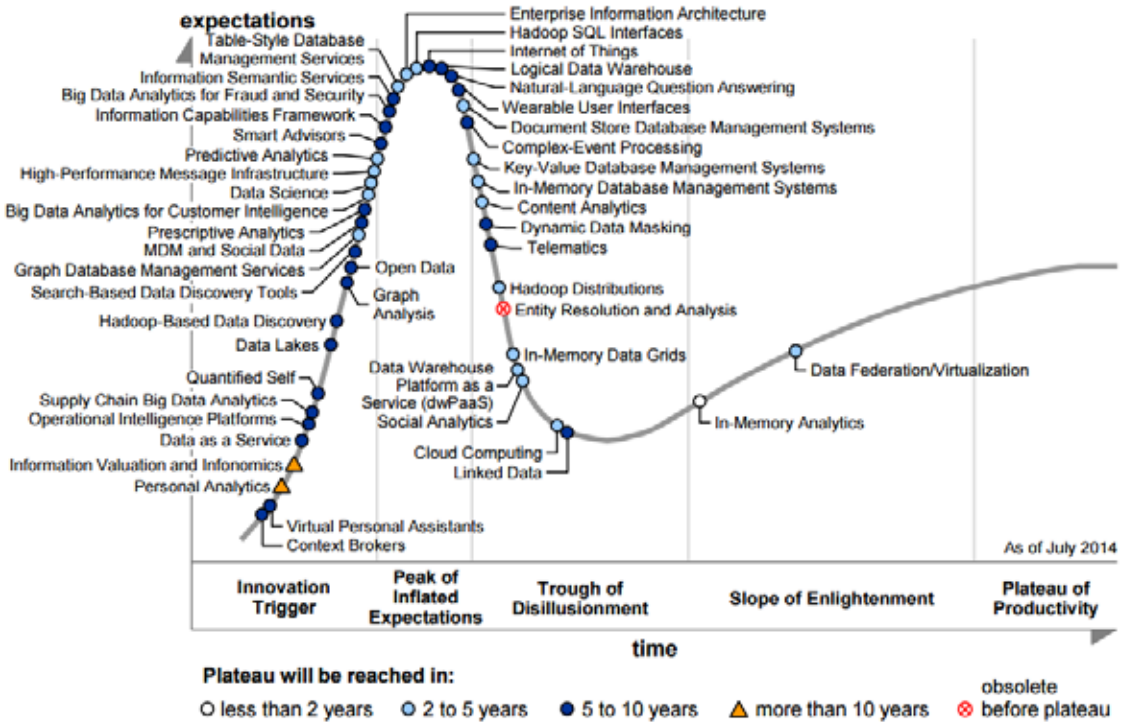

FIGURE 1 Hype Cycle for Big Data, 2014 [2]

they monetize this? Is the market ready for the big data-based offering? Are CSPs ready to deliver it?

In order to understand how CSPs will play a strategic role in this subject, it is essential to first acknowledge that – due to the already mentioned "minor" obstacles - the current moment is still an early stage in the big data hype: companies are still experimenting innovative big data based solutions, testing market acceptance. It is estimated that in the near year of 2017, "60% of big data projects will fail to go beyond piloting and experimentation and will be abandoned" [1] (refer to Figure 1).

So, it is time to test, experiment and gain critical knowledge in big data.

According to a study held by The Economist Intelligence Unit [3], the top two key data challenges in businesses are the quality, reliability or comprehensiveness of data and the lack of effective systems to gather and analyze data. Additionally, the same study results show that in the top three data insights critical to decision-making, two include "current status" (e.g. quality) and "qualitative" (e.g. customer experience). Quality of Service and user experience are, in fact, two of the most important aspects that need to be carefully evaluated in order to maximize the probability of success and revenue of a media product nowadays.

To look at the information available and understand the potential of a particular mix and match / combination of data sets is almost a transcendent art. So one shall start by using data that already grasps very well: SAPO made the PoC (the subject of this paper) using a star solution "SAPO news", a newsfeed tool built entirely in-house.

From the PoC materialized in the launch of "Máquina do Tempo" [4], SAPO climbed to a new era. It may seem innocent to correlate news, but a whole new world of possibilities was open by this research and testing: you can correlate people with less positive attitude towards society (terrorist associations), correlate news about brands or trendy goods / behaviours, correlate stock exchange information with political news in real time and trace information valuable to the society. It is truly information innovation.

The solution proposed in this article tackles both aspects by bringing new, innovative and high quality information to the end user (B2B or B2C), keeping and even improving the high standards of quality of service for SAPO/PT media news room, and at the same time providing such information through intuitive interfaces to engage the users at their maximum potential. As stated by Matthew Keylock [3], "If you don't engage with your best customers, they won't want to engage with you. Every decision either grows or erodes loyalty." Such engagement can be driven in many different ways, but with the advent of mobile and high-definition devices, data visualization is drastically increasing its importance in this field. The BBC media company is just one of many examples, where their mission for data journalism is to become visual journalism [3].

"Wikibon expects the Big Data market to top $84 billion in 2026, which represents a 17% compound annual growth rate" [5], and the CSPs are very well positioned to capture part of this value.

Related projects and State of the Art

Related work on big data, with special focus on media, may be divided into two major categories: products and projects from large media producers, typically news agencies and news editors; and smaller industrial and scientific projects focused on enriching media content available online.

Large industry related projects

Such industries envision retrieving to their customers the most informative and high-quality content and, therefore, the focus is on the data itself (e.g: news articles). Additionally, most of the effort is for the last minute news and for the "real-time" events, inevitably leaving aside rich stories told through the news, where the actual big data is.

New York Times [6] is one of the most important and long-lasting media producers, publishing news, and consequently generating data since 1851, with more than 13 million news articles available. TimesMachine [6] is a repository of "129 years of New York Times journalism, as it originally

appeared" and brings to the end user access to the digitized content of such data in a time-centric approach. Although this represents an unique and valuable resource, no further analysis and mining are performed on the top of this data. Coming from its RDI group, NYTLabs [7], Delta (a visualizing reader activity in real time), StreamTools (a graphical toolkit for data streams) or Kepler (semantic network visualization of topics) are examples of new and differentiated approaches for Big Data analysis on media. BBC Research & Development started a project [8] in 2011 focused on NLP (Natural Language Processing) and ML (Machine Learning) techniques used to spot connections, improve understanding and avoid information overload. Nevertheless, such data analysis dies at RDI Labs and never achieves its final destination, news readers, journalists or merely curious people.

The Guardian data blog [9] is considered to be one of the first systematic efforts to incorporate publicly available data sources into news reports. These reports are typically the result of large journalistic investigations of public data (news included) and have an important impact on the big data area. Nevertheless, Guardian's approach is not able to provide to the user freedom of choice, analysis and data exploration. Data analysis and results are subject to the media producer interests.

Exploratory related projects

Besides, smaller industrial and scientific projects from organizations that typically aspire to higher risk strategies, have enough space to explore and develop intelligent systems, supported on new approaches from big data and machine learning, aiming for an improved media content and user experience.

"News Explorer" [7], by IBM Watson, aims at extracting entities (persons, organizations and events) from online news and subsequently connect such entities based on their presence on news articles. Some user experience decisions on the web application disable the possibility of a richer exploration of information, diminishing the capabilities of exploring such a big data repository of knowledge.

"Libero 24x7" [10] aims to bring to its customer better and differentiated knowledge, by means of two interactive online tools: "timeline" and "grafo", a network of entities mentioned on news. Even though it allows free navigation through the news, its user interface is very poor and constrains the analysis of the data.

Recently there has been an increase of new companies, particular startups, targeting for ML and NLP as a service. AlchemyAPI [11], Luminoso [12] or Aylien [13] are examples of such boom in Software as a Service (SaaS) industries, where one of the application domains is media. Nevertheless, these industrial and scientific projects, smaller in scale and yet in an immature state, frequently fail on two essential aspects: (i) the quality and user experience regarding the interfaces with the users, ranging from web and mobile applications to infographics and interactive applications and (ii) the lack of information and extracted knowledge from data which is actually new.

State of the art

Although in different stages of maturity and complexity, many of these products and projects share similar scientific approaches to achieve their means.

Named Entity Recognition (NER) is a broad area of study that aims to identify and extract entities (person's names, organizations, etc.) from text. NER supported on machine learning is being studied by the research community for a long time [14] [15] [16] and most recent results point out to more accurate and language-independent methods [17], generic enough to be applied on agrammatical language such that from social media [18] [19]. Linguistic patterns are rule-based approaches for NER tasks [20] [21] [22], supported on dictionaries definitions and synonyms, among others. Although language (and domain) dependent, accuracy is most of the times higher when comparing with machine learning approaches.

Entity Disambiguation (NERD) is another problem yet to be fully solved. For example, "Costa" may both refer to "António Costa" or "Marco António Costa", Portuguese politicians, but also to "Costa Coffee", a

coffee shop chain in the UK. This task's major goal is to identify the true entity for each mention, based, for instance, on the information extracted from large encyclopedic collections [23] [24] or by its context [25] [26] [27].

News Media is an interesting market where quotations are highly relevant [28]. Quotations, with particular emphasis on those from public personalities, convey rich and important information, and most studies focus on direct quotations [29] [30]. Quotations identification is executed at the sentence level, both with hand-crafted rules and lexica [31] [32] [33] or machine learning approaches [28] [34]. Indirect quotations are clearly the most challenging ones, and besides that, they contain richer information, including for example, temporal and spatial modifiers.

Challenges for Smart Big Data on Media

Smart big data on media content embraces important challenges for both large media industries and smaller exploratory ones. First, working on big data environments, applying machine learning and artificial intelligence techniques and exploring their results in appropriate ways, involves complex, time-consuming and risky strategies that only smaller industry player or RDI Labs are able to tackle. Second, assuring high quality and innovative information and user experience are key aspects for larger media industry player but at the cost of less engaging and less intelligent outcomes.

The solution presented in the following section aims to fulfill this gap and improve user experience and B2B and B2C QoS, while enhancing journalist's tools for higher quality news and novel information, without neglecting the economic viability of the outcome.

Proposed Solution

The proposed solution comprehends three main pillars, as depicted in Figure 2:

- Machine learning for media;
- Big data infrastructure;
- User interface, all with common ground knowledge: news media.

This solution aims to give a contribution to the current state of the art on media projects dealing with big data scenarios, with a special focus on news. It aims to build ML models capable of extracting, classifying and correlating information available on news. To build such models, this solution uses ML algorithms that can learn from real data and predict on unseen data, using big data cloud-based infrastructures, capable of operating very large amounts of data, both historical (batch processing) and stream. Lastly, special attention is also given to the user interface and user experience, as this is the privileged mean of communication and knowledge sharing between end users and media producers.

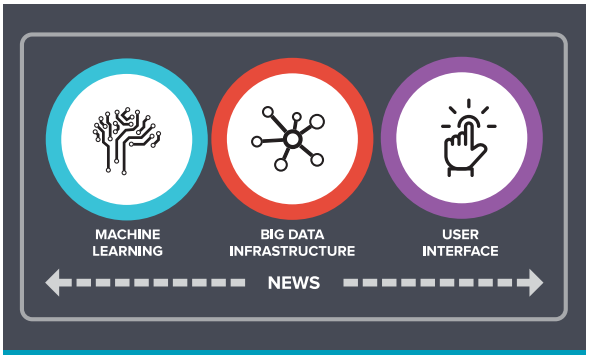

FIGURE 2 High-level domain solution

Machine Learning for Data Analysis

The architecture for the first pillar of the proposed solution is presented in Figure 3.

This workflow interconnects news articles, in its raw format, to the user interface where the final solution is available. For the sake of simplicity, news articles are assumed to be already stored in a database, although such process includes RSS feeds crawling and scraping, among other data cleaning and indexing techniques. NERD is the first step of the data processing pipeline, and this solution focuses on persons only. Although Wikipedia and Freebase [35] knowledge bases cover a large set of entities, namely persons, such sets are restricted to mediatic

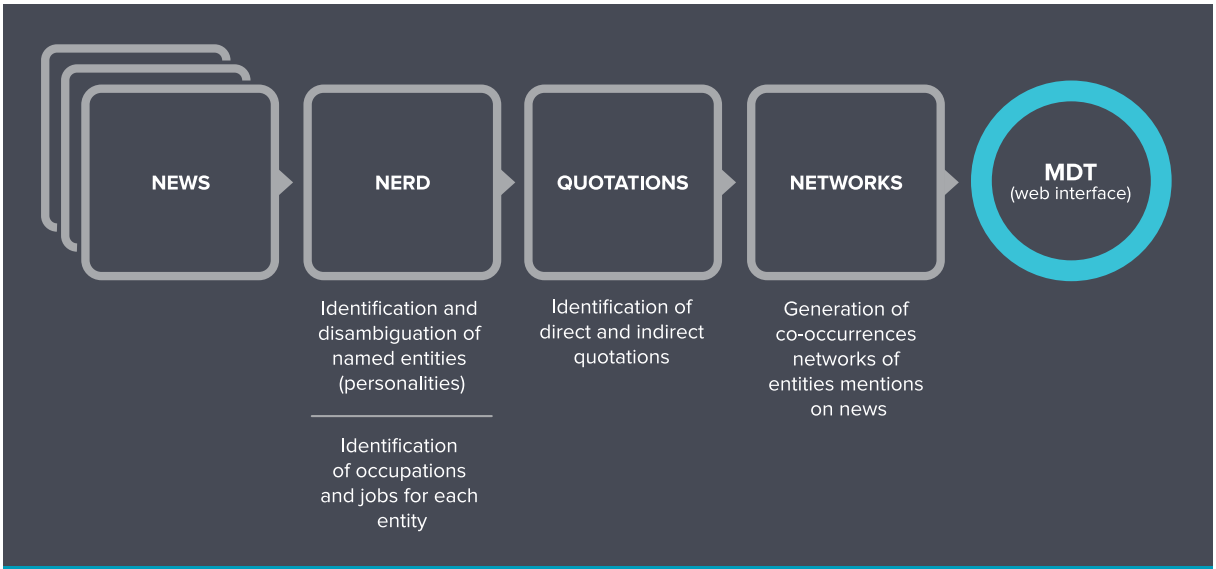

FIGURE 3 Pillar I - Machine Learning architecture

entities, are barely updated and do not include most of the persons mentioned on national scope news articles, such as Mayors, deputies, soccer coaches or CEOs of smaller companies. Moreover, manual based lists of entities do not satisfy our needs. The proposed solution for NERD is supported on ML, in particular in Conditional Random Fields (CRF) algorithm. CRFs are undirected statistical graphic models [36], and have shown that these models are well suited for sequence analysis, particularly on named entity recognition on newswire data. Our method is based on a bootstrapping approach for training a NER classifier with CRFs [37] and led to high-quality results (83% for precision and 68% for recall). Additionally, news articles convey excellent information regarding current and past jobs and roles of public personalities. Such information can be used to automatically generate micro-biographies with the support of apposition structures. For instance, "Barack Obama, president of USA, (...)" or "The Russian leader, Vladimir Putin, (...)" are examples of this linguistic structure. This property was explored such that, by using linguistic patterns, is it possible to build a high quality and specialized knowledge base [38]. Results achieved also point to a high-quality resource, with 97% precision and 59% recall values.

Automatic identification and extraction of quotations (refer to the second block in Figure 3) is a complex task due to a large number of linguistic variations that may exist at the language level. First, a non-neglectable amount of quotations omit the speaker, frequently being replaced by pronouns (e.g.: "He said..."). This process is referred as co-reference. Second, the speaker can be replaced by a different entity type, frequently an organization (e.g.: "Microsoft declared that..."), thus diverging from the true meaning of quotation. Also, speakers name can have different variations, such as short versions of the name (e.g.: "Barack Obama" and "Obama") or a job descriptor (e.g.: "The President of USA said that..."). Lastly, quotations are not necessarily bounded by single sentences and can occur in multiple sentences. Because of all these unsolved challenges on the quotations extraction topic, using a fully automatic machine learning approach for this problem would not retrieve high-quality results. Quotations' extraction is thus based on a set of linguistic patterns matched against news articles sentences. Results show higher precision values for direct quotations, as expected, but also considerably high (approximately 80%) precision values for indirect quotations extracted from single sentences.

The last core block from the architecture presented in Figure 3 refers to networks. Networks, also known as Graphs, allow the visual representation of relations,

without losing important characteristics of networks such as the strength of connections, centrality or clicks. For news media, networks of entities are clearly a meaningful approach to visualize and analyze information reported on news articles. This solution is based on networks of co-occurrences of persons on news. For each news article, if two or more persons are mentioned in that particular text, a co-occurrence relation is created. Nodes represent persons while edges represent co-occurrences. The size of each node is directly connected with the number of mentions of the person of news articles, and the thickness of the edges matches the number of news articles which both persons co-occur. Apart from node size and edge thickness, yet another variable can be introduced on the network, namely entity types or categories of documents, to name a few.

Although graphs may seem to be a good solution to visualize relationships between entities extracted from news, this approach quickly becomes non-informative as soon as the amount of data drastically increases – a big data scenario. The number of nodes and edges become so high that no visual patterns are possible to obtain from such networks, if no additional steps are taken into account. The proposed solution uses a two-folded approach to deal with this challenge. The first feature to highlight is the Force Atlas algorithm, used on networks. This algorithm [39] is supported on the physical behaviour of electrical charges and springs and aims to spatially distribute nodes and edges of the network based on the strength of the connections. It naturally enhances the creation of clusters (groups of connected nodes), which points to a particularly relevant aspect of such data visualization.

Second, both egocentric and global networks have a depth of 1.5, instead of typical unitary depth networks (star networks). Examples of both networks depths are presented in Figure 4.

Depth networks are a common measure of distance on graphs and represent the number of edges that needs to be transversed between any two nodes. This solution is based on 1.5 networks (right example in Figure 4), a balanced solution between unitary depth networks (left example in Figure 4), which are low informative networks, and networks with depth equal or higher than 2, which can easily saturate.

The third and last important feature included in this proposed solution relates to the classification of news articles. Classification refers to the process of automatically assigning one or more classes to a particular document (e.g.: sports, Europe or

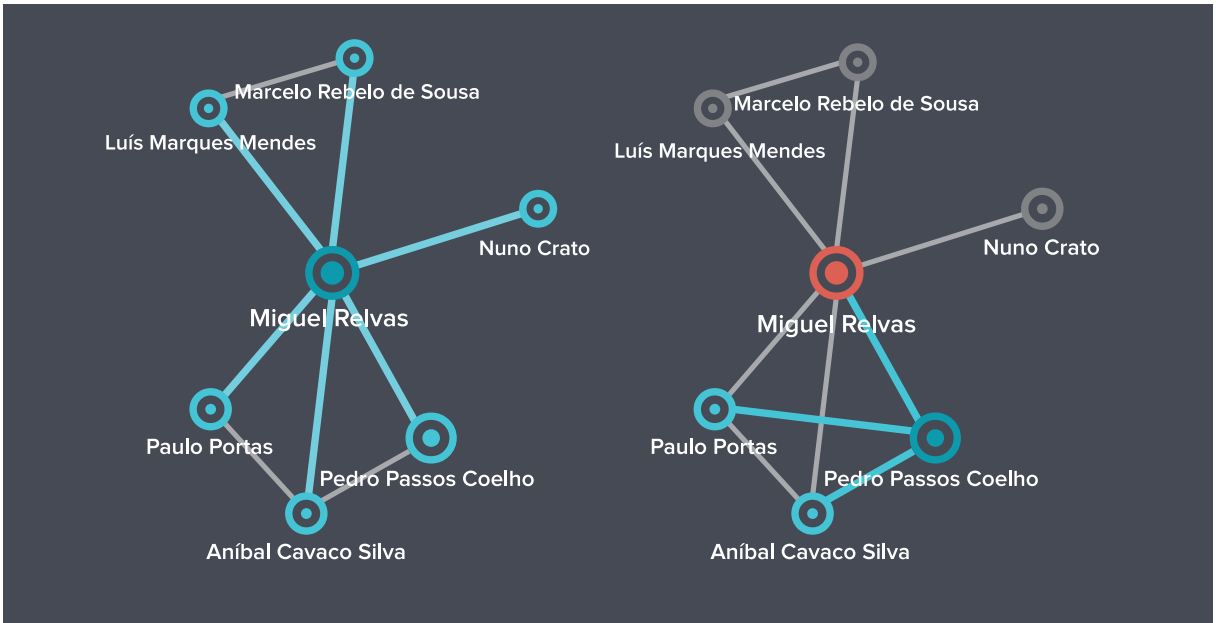

FIGURE 4 Egocentric networks with depth 1 (left side) and depth 1.5 (right side)

migrants). To tackle this classification problem, a multi-classifier approach was implemented: Support Vector Machines (SVM) and Nearest Neighbour (NN) classification models were built to deal with different specificities of this task [40] [41]. Much of the classification process is commonly performed by journalists and editors at the time of writing a news article. Such information is typically delivered to end-users through metadata (e.g.: tags associated with news articles). Machine learning is one of the approaches used to expand such list of metadata classes and thus enrich news articles with additional information. With an expanded list of classes for each news articles, it becomes possible to correlate entities with classes (e.g.: "António Costa" is frequently mentioned on politics news) and, ultimately, filter networks of entities based on news classes, allowing more detailed search and analysis of data.

An Infrastructure for Big Data on Media

The second large contributor for the proposed solution is the architecture, depicted in Figure 5. It represents the logical infrastructure of the proposed solution, from an engineering point-of-view.

From the flow of information represented in Figure 5, on the most left side of the architecture, news articles are consumed from a distributed messaging framework, SAPO Broker [42]. This framework, among other features, provide Publish-Subscribe and Point-to-Point messaging, guaranteed delivery and wildcard subscriptions. From the scalability perspective, there can be as many consumers as necessary, according to the throughput of data. Data is then routed to the Information Extraction Engine, where most of the information extraction, processing and classification takes place, as described in this section. IE engine is modular, such that information is extracted and processed in logically independent blocks. With this approach, in the event of an increase of the throughput of data, this represents an increase on the number of blocks of data to process, which, ultimately, means more computational power need (e.g.: additional server or virtual machine, more CPUs or an increase of

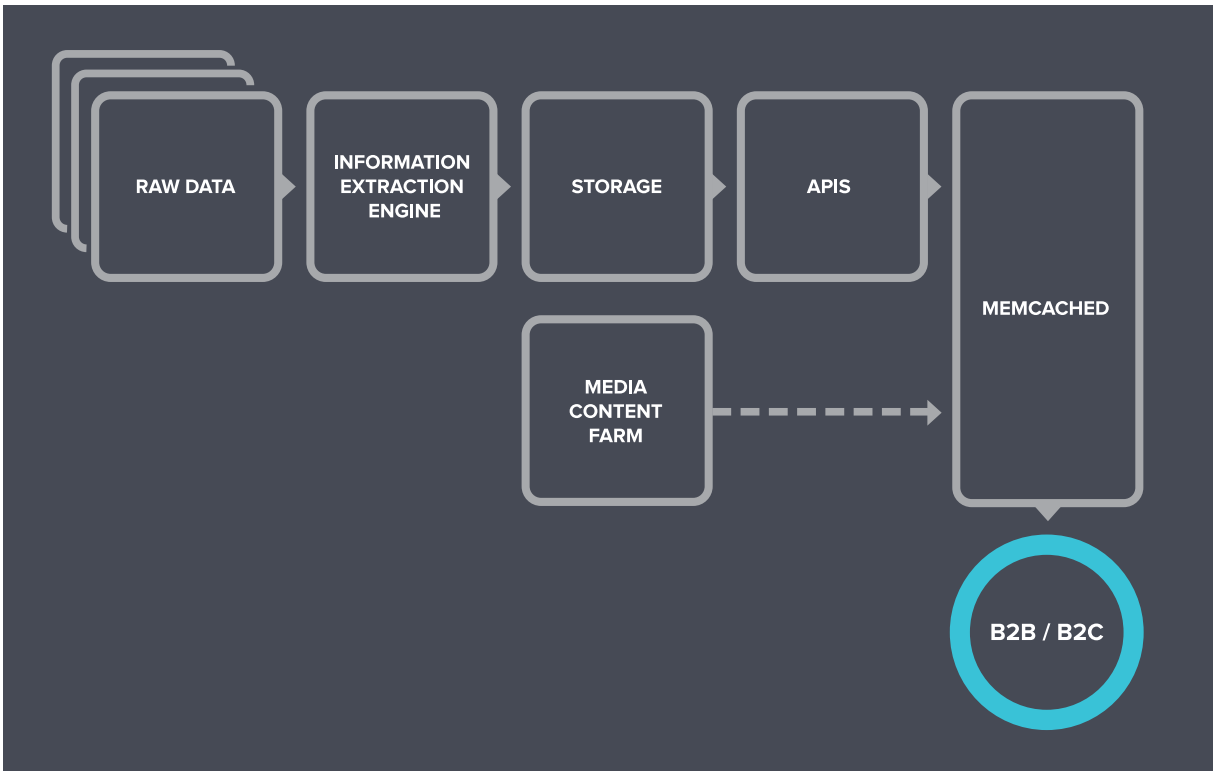

FIGURE 5 Pillar II - Big Data infrastructure

memory). Such information is subsequently stored and indexed in high performance and low latency systems such as SOLR. Additionally, these indexing systems are redundantly spread in the backend network for both performance and resilience issues. Moreover, nontextual data (typically images and videos) is stored in a separate key/value database for improved performance responses.

APIs account for the interface between storage and indexing systems, with fully processed data, and final user applications and services, for both B2B and B2C scenarios. Such APIs are specifically designed and optimized for pre-established usages to guarantee standardization of responses and sub-second responses. A further cache mechanism is the last building block of this architecture, positioned to ensure that repeated requests are quickly and seamlessly retrieved.

User Interface for Media

The last main pillar referred in the high-level domain solution from Figure 2 is the user interface. From a user-centric perspective, the user interface is the second biggest challenge for big data on media, and successful products and services in this field are still scarce (refer to section 2). There are two important aspects that impact on the quality of the user interface and consequently on the user satisfaction: (i) how the information is organized and (ii) how the user can interact with such information. Both aspects are crucial to engaging end-users and turn this approach into a successful product thus creating new opportunities for business.

In the following section will be presented “Máquina do Tempo”, a media product, developed by SAPO and maintained by Altice Labs in collaboration with the University of Porto. This product is the face of a big data and ML infrastructures with a user interface carefully designed to fulfill both user and business needs.

“Máquina do Tempo”

“Máquina do Tempo” (MdT) [4] is the completion of the solution presented in the previous chapter. It is an interactive online tool which allows users to navigate and explore news published during the last 25 years. Such large repository comprehends news from LUSA, a major Portuguese news agency, together with news published online by the main Portuguese news stream. With MdT users can analyze personalities and events from from historic records as reported on the news.

This project was developed within the scope of a collaboration between the SAPO R&D Laboratories at the University of Porto and SAPO Notícias and is the culmination of five years of research and development from both institutions.

MdT comprehends a set of approximately 8 million Portuguese news articles written between 1987 and today, representing around 160 million distinct sentences and more than one thousand million words.

MdT has more than 200 thousand distinct personalities and 5 million relationships, and about half a million direct and indirect quotations stated by these entities. On average, each entity has 4 different identified occupations during the 25 years’ timeframe of MdT. All this information is automatically extracted and indexed with ML and natural language approaches (refer to section 4).

MdT has essentially two means of interaction: a personality based and a timeline selection. Exploring MdT from personalities point of view allows users to access an automatically generated profile page for each entity, thus enabling user access to a wealth of information ranging from the entity photo and occupation to quotations extracted from news, most relevant news mentioning such personality as well as the deeply most connected entities on a specific timeframe. Figure 6 depicts the profile page from Marcelo Rebelo de Sousa, a Portuguese public personality frequently mentioned on news and with a long political record.

The header of each profile page includes, apart from name and occupation, statistics regarding the entity’s presence on news, such as the total number of mentions, quotations or relationships and a distribution over news categories/themes (e.g.: politics or sports). Additionally, users can explore the presence of each personality on news since 1990 at

FIGURE 6 MdT profile page

the distance of a click: using the timeline view, the date and time intervals can be changed according to user's needs and all information presented on each page is updated accordingly. By exploring the networks of personalities, users are able to navigate on egocentric co-occurrences networks for each specific profile page selection.

Moreover, global networks, as depicted in Figure 7, allow users to explore co-occurrences networks without any particular egocentric entity. This is particularly interesting to explore clusters of entities and their most relevance presence on different topics of news, which can be obtained with the support of different color nodes for each topic.

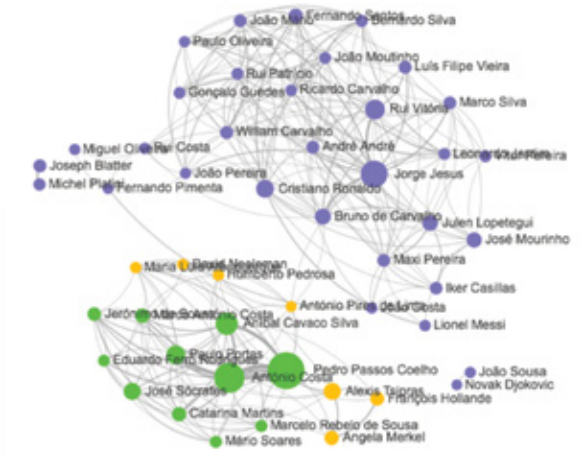

FIGURE 7 MdT co-occurrences network

As previously discussed (refer to sections 2 and 3), user engagement is one of the key data challenges in business. SAPO web portal, with approximately 30 million daily visits, is constantly pursuing innovative and positive experiences to improve and impact on its user engagement. Considering this as an opportunity for SAPO news services, SAPO developed a widget at SAPO web portal (its homepage), which is an entry point for MdT. The widget is presented as shown in Figure 8.

FIGURE 8 MdT widget at SAPO homepage

The idea behind this widget is simple: give clear and updated information to SAPO users about who is getting more attention on news. The result is a top three list with the persons most mentioned on news during the past three days. For each person, the user has a photo, to immediately associate the name with the person, as well as the current job/occupation most frequently referred by media. With this approach, we believe two goals were achieved: first, and based on users' feedback, we identify an improvement of user engagement by adding new approaches on top of traditional online news services; second, there was a significant increase on web traffic, as a consequence of the redirection of traffic from SAPO homepage to MdT, impacting directly on ads revenue, among other KPIs.

Conclusion and future outlook

Is data big? Yes. Big data has already arrived and media is well part of it. There are definitely no doubts concerning that. Is data getting smarter? The amount of data produced on a daily basis by media industries is too high, so that the benefits the society can take from it are still scarce. What to do with so much data is what actually bothers all, from scientists and journalists to entrepreneurs and business people.

The society and the industry are both still far from having intelligent news media software. The size of data is not, at this time, the greatest challenge. Such challenge is related with the language analysis and processing, as well as to prepare machines to actually extract novel and usefull knowledge from data. At the current stage, machine learning and artificial intelligence are able to enhance content and support humans on decision making. That seems to be all.

New York Times states that "The future of news is not an article" [43], and pinpoints some future directions such as enhanced tools for journalists and adaptive content, referring that the future of news is much more than a stream of articles and highlights the distinction between ephemeral content and evergreen content. The journalism business needs to create and deliver more high-quality information, instead of, for instance, publishing hundreds of articles a day, then starting all over the next day, recreating any redundant content each time.

SAPO's "Máquina do Tempo" is an approach to support journalists on their investigation, to give readers the opportunity to explore news without any barriers and to eventually support the society with new and richer information along time. Next developments on MdT will include additional entities types (organizations, locations and events), also more complex relations between entities [44], and an innovative approach to automatically build stories from news [45].

References

[1] Heudecker, Nick, Lakshmi Randall, Roxane Edjlali, Frank Buytendijk, Douglas Laney, Regina Casonato, Mark Beyer, and Merv Adrian, Predicts 2015: Big Data Challenges Move from Technology to the Organization, Gartner, 2014

[2] Buytendijk, Frank, Hype Cycle for Big Data, 2014, Gartner, 2014

[3] Stone, Martha L. Big Data for Media, Reuters Institute for the Study of Journalism, 2014

[4] SAPO, Máquina do Tempo, accessed December 17th, 2015, http://maquinadotempo.sapo.pt

[5] Kelly, Jeff, "Big Data Vendor Revenue and Market Forecast, 2011-2016", Wikibon, 2015

[6] New York Times, NY 2015, accessed December 17th, 2015, http://nytimes.com

[7] Explorer, News, News Explorer, accessed December 17th, 2015, http://news-explorer.mybluemix.net

[8] BBC, Natural Language Processing- Automated tools for text processing and analysis, accessed December 17th, 2015, http://www.bbc.co.uk/rd/projects/nlp

[9] The Guardian DataBlog, The Guardian, accessed December 17th, 2015, www.theguardian.com/data

[10] Libero, Libero 24x7, 2015, http://247.libero.it

[11] Alchemy, accessed December 17th, 2015, http://alchemyapi.com/products/alchemydata-news

[12] Luminoso, Luminoso: API, accessed December 17th, 2015, http://www.luminoso.com/products/api

[13] Aylien, accessed December 17th, 2015, http://aylien.com

[14] Nadeau, David, and Satoshi Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (1): 3-26, 2007

[15] Zhou, GuoDong, and Jian Su, Named entity recognition using an HMM-based chunk tagger, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 474-480, 2012

[16] Florian, Radu, Abe Ittycheriah, Hongyan Jing, and Tong Zhang, Named entity recognition through classifier combination, Proceedings of the seventh conference on Natural language learning at HLT-NAACL (Association for Computational Linguistics) 4: 168-171, 2013

[17] Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran, Learning multilingual named entity recognition from Wikipedia, A 194: 151-175, 2013, "Learning multilingual named entity recognition from Wikipedia", Artificial Intelligence (194): 151-175, 2013

[18] Ritter, Alan, Sam Clark, and Oren Etzioni, Named entity recognition in tweets: an experimental study, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 1524-1534, 2011

[19] Liu, Xiaohua, Shaodian Zhang, Furu Wei, and Ming Zhou, Recognizing named entities in tweets, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 359-367, 2011

[20] Hanisch, Daniel, Katrin Fundel, Heinz-Theodor Mevissen, Ralf Zimmer, and Juliane Fluck, ProMiner: rule-based protein and gene entity recognition, BMC bioinformatics 6 (1), 2005

[21] Abdallah, Sherief, Khaled Shaalan, and Muhammad Shoaib, Integrating rule-based system with classification for Arabic named entity recognition, Computational Linguistics and Intelligent Text Processing (Springer Berlin Heidelberg) 311-322, 2012

[22] Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky, Deterministic coreference resolution based on entity-centric, precision-ranked rules, Computational Linguistics 39 (4): 885-916, 2013

[23] Cucerzan, Silviu, Large-Scale Named Entity Disambiguation Based on Wikipedia Data", In EMNLP-CoNLL, 708-716, 2007

[24] Bunescu, Razvan C., and Marius Pasca, Using Encyclopedic Knowledge for Named entity Disambiguation, EACL, 9-16, 2006

[25] Li, Yang, Chi Wang, Fangqiu Han, Jiawei Han, Dan Roth, and Xifeng Yan, Mining evidences for named entity disambiguation", Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1070-1078, 2013

[26] Hoffart, Johannes, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum, Robust disambiguation of named entities in text, Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 782-792, 2011

[27] Saleiro, Pedro, Luís Rei, Arian Pasquali, Carlos Soares, Jorge Teixeira, Fábio Pinto, Mohammed Nozari, Catarina Félix, and Pedro Strech, POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter, CLEF Eval, 2013

[28] Pareti, Silvia, O'Keefe Timothy, Konstas Ioannis, James Curran, and Irena Koprinska, Automatically Detecting and Attributing Indirect Quotations, Proccedings of EMNLP, 989-999, 2013

[29] Elson, David K., and Kathleen McKeown, Automatic Attribution of Quoted Speech in Literary Narrative", AAAI, 2010

[30] Pouliquen, Bruno, Ralf Steinberger, and Clive Best, Automatic detection of quotations in multilingual news, Proceedings of Recent Advances in Natural Language Processing, 487-492, 2007

[31] de La Clergerie, Éric, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, and Victor Mignot, Extracting and visualizing quotations from news wires, In Human Language Technology, Challenges for Computer Science and Linguistics, Springer Berlin Heidelberg, 522-532, 2011

[32] Krestel, Ralf, Bergler Sabine, and Witte, René, Minding the source: Automatic tagging of reported speech in newspaper articles, Reporter 1 (5): 4, 2008

[33] Sarmento, Luís and Sérgio Nunes, Automatic Extraction of Quotes and Topics from News Feeds, 4th Doctoral Symposium on Informatics Engineering, DSIE, 2009

[34] Fernandes, William, Paulo Ducca, Eduardo Motta, and Luiz Milidiú Ruy, Quotation extraction for portuguese, Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, Cuiabá, 204-208, 2011

[35] http://www.freebase.com

[36] McCallum, Andrew, and Li Wei, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, Proceedings of the seventh conference on Natural language learning at HLT-NAACL, Association for Computational Linguistics, 2003

[37] Teixeira, Jorge, Luís Sarmento, and Eugénio Oliveira, A Bootstrapping Approach for Training a NER with Conditional Random Fields, Progress in Artificial Intelligence, Springer Berlin Heidelberg, 664-678, 2911, 2011

[38] Verbetes, Verbetes API, accessed December 17th, 2015, https://store.services.sapo.pt/en/cat/catalog/other/free-api-information-retrieval-verbetes

[39] Jacomy, Mathieu, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian, ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, PLoS ONE 9 (6), 2014

[40] Teixeira, Jorge, and Luís, Oliveira, Eugénio Sarmento, Semi-automatic creation of a reference news corpus for fine-grained multi-label scenarios, Proceedings of the 6th CISTI, IEEE, 1-7, 2011

[41] Sarmento, Luís, Sérgio Nunes, Jorge Teixeira, and Eugénio Oliveira, Propagating Fine-Grained Topic Labels in News Snippets, Proceedings of the 2009 IEEE/WIC, IEEE Computer Society, 515-518, 2009

[42] SAPO Broker, accessed December 17th, 2015, http://oss.sapo.pt/sapo-broker

[43] Lloyd, Alexis, "The Future of news is not an article", The New York Times R&D Lab, accessed December 17th, 2015, http://nytlabs.com/blog/2015/10/20/particles/

[44] Saleiro, Pedro, Jorge Teixeira, Carlos Soares, and Eugénio Oliveira, TimeMachine: Entity-centric Search and Visualization of News Archives (to be published)", Proccedings of ECIR, Springer, 2016

[45] Abreu, Carla, Jorge Teixeira, and Eugénio Oliveira, ENCADEAr: Encadeamento automático de notícias, Oslo Studies in Language 1 (7), 2015