Patterns Recognition Approach for Newspapers Analytics: Case of Algerian PDF Newspapers

Dihia LANASRI[1]

[1]PROXYLAN EPE/SPA, Benaknoun, Algiers, Algeria

Abstract

Nowadays, companies are convinced that putting people at the heart of their businesses is vital for their competition performances and success. To achieve this goal, a deep understanding of people's published content is mandatory to evaluate their satisfaction, frustration, and interestingness, and eventually recommend them other items and services.

Media Analytics or News analytics is the main solution used to collect, process and analyze the different news and content published by people on media, social networks, etc. to provide the required insights and metrics which help in data-driven decision making.

Press newspapers, precisely, PDF newspapers are one of the valuable and rich generated content that should be collected and analyzed by companies and organizations due to the capital of information they contain. Newspaper Analytics is a promising field of study, however, a reduced number of works are interested in this topic.

The lack of work dealing with this kind of content (PDF newspapers) for data analytics motivates us to propose: (1) An end-to-end approach which allows conducting a newspaper analytics applied to PDF newspapers; (2) A detailed approach for PDF newspaper pattern recognition, for Latin and Arabic PDF, is presented.

This approach consists in identifying the different blocks of press articles in a PDF page, and their associated metadata (authors, title). Once boundings of each article is detected, the recognized articles and metadata can be extracted using the OCR techniques (which is out of the scope of this paper).

This approach is validated through different experiments applied on our proper constructed dataset. This later contains more than 1.500 PDFs collected from different Algerian newspapers in Latin and Arabic languages, during five months. The obtained results are promising and allowed us to develop a tool which presents the results of PDF newspaper analytics to end-users through dashboards and KPIs (key performance indicators) to keep an eye on their presence in the media and their reputation.

Keywords

Newspaper Analytics, Pattern Recognition, Computer Vision, Deep Learning

1. Introduction

News analytics, Media analytics or Media Intelligence is a really promising field which needs more attention regarding its added value for companies, industries, organizations and governments. News analytics refers to the different tools, solutions, metrics and indicators used to analyze the huge amount of published news, comments, posts, reviews, etc., in their different formats mainly on websites and social networks. The advents made by web2.0 have encouraged

RIF'23: The 12th Seminary of Computer Science Research at Feminine, March 09, 2023, Constantine, Algeria

dihia.lanasri@proxylan.dz (D. LANASRI)

0000-0002-3794-844X (D. LANASRI)

people to share more content on the web. According to Datareportal[1] special report published in July 2020, important raise is experienced in social media activities during the beginning of the COVID-19.

These active users on social media have become more powerful and companies become more careful with this users content. This content expresses satisfaction, frustration, or delights user sentiments [1] and has a big impact on influencing customers' decisions [2]. As a consequence, companies are spending a lot of effort in making and improving their business by deeply analyzing and understanding this content [3].

Gathering and analyzing this kind of data is very important mainly for top management to keep an eye on the evolution of the market, build competitive advantages, analyze the evolution of business, understand the requirements of their customers or their prospects, etc.

News analytics is based on complex techniques of natural language processing, advanced linguistic analysis to process and understand the context of text, then derive interesting qualitative and quantitative insights about targeted audience, opinion analysis, fake news detection, context analysis, etc. This information is valuable to help managers to build their data-driven strategies, improve their risk management tools, and make better business decisions.

Newspaper analytics is a sub-field of news analytics which focuses on the analysis of data published in press newspapers by journalists in the form of articles. For many years, specific entities are created inside companies and organizations to collect, read and analyze the different articles published in plenty of newspapers. Specific roles are hired just to perform this burden task of manually selecting, then cutting the portions of newspaper that may interest the managers. The emergence of the web and the development of technology has reduced the difficulty of this task thanks to the availability of electronic newspapers in HTML or PDF formats.

Many solutions are provided to collect, process and analyze the content of web news based on advanced analytics techniques like machine learning and deep learning, NLP. Hundreds of paid solutions are provided by vendors to facilitate this task like Refinitiv Machine Readable News Analytics[2]. However, a lack of works and solutions dealing with newspapers, mainly the PDF format, is identified in industrial and academic fields. Even if some open source and paid solutions allow extracting the whole content of this type of documents like (pdf2text python library) in a raw shuffled format, but this is not really suitable for PDF Newspaper. This latter is characterized by its specific format of articles blocks. Where in each page, many blocks are dumped and each block represents the text of the article, its authors and its title.

Identifying the parts and the boundings of each article like illustrated in figure 1 with its related metadata like title and author is an imperative requirement, in order to be able in the future to extract articles one by one in text format from each PDF page. Once extracted, different NLP and text processing techniques can be applied on text to understand and analyze it. Plenty of usage cases can be proposed like sentiment analysis [4] of a given article, opinion mining, Topic modeling, etc.

The lack of academic and industrial works dealing with this issue motivates our proposal.

[1] https://datareportal.com/reports/digital-2020-july-global-statshot

[2] https://www.refinitiv.com/en/financial-data/financial-news-coverage/political-news-feeds-analysis/news-analytics

يوم حاسم في سباق التشريعيات

تنقضي هذا الأحد مهلة دراسة ملفات الترشح لتشريعيات 12 جوان المقبل والتي ستفصل فيها السلطة الوطنية المستقلة للانتخابات، ويترقب آلاف المترشحين المحتملين ما تحمله لهم الساعات القادمة من مستجدات في ظل سقوط أسماء معروفة من السباق الانتخابي المبرمج بعد شهر من الآن. • ف. زينب

أحصت السلطة الوطنية المستقلة للانتخابات في هذا الصدد ما لا يقل عن 24,214 ملف ترشح تنقضي مهلة دراستها هذا الأحد، وذلك قبل إعطاء إشارة انطلاق الحملة الانتخابية، علما أن إيداع الملفات تم في 27 أفريل المنصرم واستفاد المترشحون من تمديد للآجال لمدة خمسة أيام اضافية بعد استشارة مجلس الدولة والمجلس الدستوري وأخذ رأي مجلس الوزراء.

وكان المجلس الدستوري قد أكد "دستورية" أحكام الأمر الموقع من قبل رئيس الجمهورية، والذي يقضي بتمديد آجال إيداع ملفات الترشح لتشريعيات 12 جوان لكونها "لا تمس بالضمانات الدستورية لممارسة المواطن لحقه في الترشح".

وتحسبا لهذا الموعد الانتخابي، سيعقد رئيس السلطة الوطنية المستقلة للانتخابات، محمد شرفي، غدا الأحد، لقاء مع رؤساء الأحزاب السياسية بمقر السلطة وذلك استجابة لطلبهم.

وتشير الأرقام الأخيرة التي أعلنت عنها السلطة إلى أن "العدد الإجمالي للمترشحين بلغ 2,400 مترشح منهم 1,180 قائمة حزبية و1,220 قائمة حرة".

كما أودع 39 حزبا سياسيا ملفات الترشح عبر 58 مندوبية للسلطة الوطنية للانتخابات، بينما تقدمت الجالية الوطنية المقيمة بالخارج بـ65 قائمة من بينها 61 تابعة للأحزاب.

وحددت السلطة الوطنية المستقلة للانتخابات الشروط الواجب توفيرها من طرف الأحزاب السياسية لقبول إيداع قوائم الترشيحات لتشريعيات 12 جوان القادم ومن ضمنها تزكية القائمة بـ25,000 توقيع للناخبين عبر 23 ولاية على أن لا يقل العدد الأدنى من التوقيعات في كل ولاية 300 توقيع.

أما بالنسبة للقوائم المستقلة، فتنص المادة 316 من القانون العضوي للانتخابات على أنه "يجب أن تدعم كل قائمة بـ100 توقيع على الأقل، عن كل مقعد مطلوب شغله من ناخبي الدائرة الانتخابية المعنية".

وفيما يتصل بالقوائم الانتخابية في الخارج، ينص قانون نظام الانتخابات على أن قوائم المترشحين تقدم "إما تحت رعاية حزب سياسي أو عدة أحزاب سياسية (دون اشتراط التوقيعات) وإما بعنوان قائمة حرة تكون مدعمة بـ200 توقيع على الأقل عن كل مقعد مطلوب شغله من توقيعات ناخبي الدائرة الانتخابية المعنية".

وفي حال تم رفض ملف ترشح احد المترشحين، بعد دراسته، يمكن للمعني تقديم طعن على مستوى المحكمة الإدارية المختصة إقليميا في مدة اقصاها 3 أيام من تاريخ التبليغ بالرفض وفق المادة 98 من القانون العضوي للانتخابات.

وبالنسبة للجالية الوطنية بالخارج، يكون قرار الرفض قابلا للطعن بالنسبة لمترشحي الدوائر الانتخابية بالخارج امام المحكمة الإدارية بالجزائر العاصمة خلال خمسة أيام ابتداء من تاريخ التبليغ.

وبالمقابل، "تفصل المحكمة الإدارية في الطعن خلال خمسة (5) أيام كاملة ابتداء من تاريخ تسجيل الطعن ويبلغ الحكم تلقائيا وفور صدوره بأي وسيلة قانونية إلى الأطراف المعنية وحسب الحالة إلى الوالي أو رئيس الممثلية الدبلوماسية أو القنصلية قصد تنفيذه"، علما أن الحكم يكون "غير قابل لأي شكل من أشكال الطعن"، وفق نفس المادة.

جبهة المستقبل تبرز أهمية الانتخابات التشريعية

أبرز رئيس حزب جبهة المستقبل، عبد العزيز بلعيد، أمس السبت بالجزائر العاصمة، أهمية الانتخابات التشريعية المقبلة في بناء الجزائر الجديدة.

وخلال ندوة صحفية، أكد السيد بلعيد أن حزبه الذي اختار كشعار للحملة الانتخابية "انتخب المستقبل"، تمكن من جمع استمارات الترشح في 61 منطقة منها 4 بالخارج، وأن س7,84 بالمائة من المترشحين هم جامعيين و32 بالمائة هم نساء"، ملفتا النظر لبعض المشاكل التي التقى بها حزبه على غرار إقصاء عدد من المناضلين بحجج. حسبه. "غير منطقية".

وأضاف في هذا الصدد، أنه تم وضع طعون على مستوى الولايات ومجلس الدولة في انتظار رد العدالة، التي. كما أكد. "سنلتزم بكل قراراتها"، معبرا في نفس الوقت، عن أمله أن تكون السلطة الوطنية المستقلة للانتخابات "منتخبة غير معينة" مستقبلا، قائلا: "حتى ولو كان عملها مثالي سنناضل من أجل ذلك في إطار الديموقراطية".

كانت وراء 1000 حادث مرور خلال 3 أشهر

670 قتيل على طرقات الجزائر بسبب الدراجات

كشفت المكلفة بالإعلام على مستوى المندوبية الوطنية للأمن عبر الطرق، أمس السبت، بالبليدة، أنه تم تسجل أكثر من ألف حادث مروري وقع على مستوى شبكة الطرقات الوطنية تسبب فيها سائقي الدراجات النارية خلال الثلاثي الأول من السنة الجارية.

وأوضحت السيدة فاطمة خلاف لوكالة انباء الجزائرية، على هامش فعاليات يوم تحسيسي موجه لفئة سائقي الدراجات النارية أن المندوبية الوطنية للأمن عبر الطرق سجلت خلال الثلاثي الأول 1202 حادث تورط فيها سائقي الدراجات النارية وهذا بين 5874 حادث مروري سجل عبر الطرقات الوطنية خلال السنة الجارية وأسفر عن وفاة 670 شخص وإصابة 7747 آخرين.

واستنادا لذات المتحدثة فإن نحو 20 بالمائة من مجمل حوادث المرور التي وقعت خلال الفترة المذكورة، تسبب فيها هذا الصنف من السائقين، لافتة إلى أن أبرز التجاوزات المرتكبة من طرفهم هي السرعة المفرطة والتجاوزات الخطيرة وعدم إرتداء الخوذة الواقية للرأس التي تعد من أهم وسائل الحماية التي يتخلى عنها نسبة معتبرة منهم بالرغم من دورها الهام في حمايتهم من خطر الموت أو الإصابة بعاهات مستديمة.

ويهدف الحد من حوادث المرور لا سيما تلك التي يتورط فيها سائقو الدراجات النارية، سطرت المندوبية الوطنية للأمن عبر الطرق برنامجا خاصا يتضمن تنظيم حملات تحسيسية على مستوى الولايات التي تعرف إنتشارا كبيرا للدراجات النارية على رأسها ولاية البليدة وكذا الجزائر العاصمة وتيبازة وغرداية . وفي هذا الصدد، تم على مستوى حظيرة المركب الرياضي مصطفى تشاكر بالبليدة تنظيم يوم تحسيسي لفائدة هذا الصنف من السائقين، وذلك بالتنسيق مع نادي الدراجات النارية للولاية وبمشاركة عدة نوادي أخرى من ولايات مجاورة على غرار الجزائر العاصمة وتيبازة وكذا الأجهزة الأمنية ممثلة في الأمن والدرك الوطنيين والحماية المدنية.

وفي هذا الاطار، صرح رئيس نادي الدراجات النارية لولاية البليدة، فيصل مالك أن القائمين على هذه المبادرة التوعوية حرصوا على تحسيس سائقي الدراجات النارية حول ضرورة إيلاء أهمية لعدد من العوامل التي تعد من بين أبرز الأسباب المؤدية إلى وقوع حوادث المرور والمتمثلة خاصة في تفادي السرعة المفرطة والحرص على الصيانة الدورية للمركبة بالإضافة إلى تنبيههم حول دور الخوذة في حماية حياتهم.

كما تخلل هذا اليوم التحسيسي أيضا تنظيم معرض للدراجات النارية بمختلف أنواعها وأحجامها وكذا تقديم عروض تعكس السياقة السليمة للدراجة النارية.

خط جديد للسكك الحديدية:

أول رحلة بين المسيلة والعاصمة هذا الاثنين

سيدخل خط جديد للسكك الحديدية يربط بين المسيلة والجزائر العاصمة حيز الخدمة "قريبا" بعد تعليق لحركة القطارات على هذا الاتجاه دام أكثر من أربع سنوات، حسب ما علم أمس السبت من مديرية النقل.

وستكون أول رحلة لهذا القطار يوم الاثنين المقبل على الساعة 5 صباحا انطلاقا من المسيلة مرورا بولايات برج بوعريريج، البويرة وبومرداس وصولا إلى الجزائر العاصمة على الساعة 8 صباحا، وفقا لما أفاد به نفس المصدر، مضيفا أن رحلة العودة لهذا القطار مبرمجة في ذات اليوم على الساعة 16 زوالا على أن يعود إلى ولاية المسيلة على الساعة الثامنة ليلا.

وسيتم توفير قطارات من نوع متطور لنقل المسافرين عبر هذا الخط استنادا لما أكدته مديرية النقل، لافتة إلى أن وضعه حيز الخدمة يندرج في إطار تنويع الخدمات المقدمة من طرف الشركة الوطنية للسكك الحديدية.

ومن أجل تشجيع مستعملي هذه الوسيلة للنقل، تقترح الشركة الوطنية للسكك الحديدية تخفيضا يصل إلى 60 بالمائة على سعر التذكرة، مثلما تمت الاشارة اليه.

للتذكير فقد تم تعليق حركة النقل عبر خط السكة الحديدية الرابط بين المسيلة والجزائر العاصمة منذ أزيد من أربع سنوات لأسباب تتعلق بأسعار التذكرة ومواقيت مرور القطار الذي كان سابقا ينطلق من ولاية باتنة ويمر بالمسيلة على الساعة التاسعة صباحا، الشيء الذي لم يكن يناسب المسافرين.

الدكتور عبد الكريم تفرقنيت:

"التصدي للفايك نيوز يكون بسن قوانين ردعية"

أفاد أستاذ الإعلام بجامعة الجزائر 03 الدكتور عبد الكريم تفرقينت خلال نزوله ضيفا على ندوة فكرية نظمتها جمعية الصحفيين والمراسلين لولاية البليدة بمناسبة اليوم العالمي لحرية التعبير بأن "الفايك نيوز" "الأخبار الكاذبة". تحول إلى هاجس حتى في الدول الأوربية، مشيرا إلى ظهور صفحات إعلامية لمكافحة الفايك نيوز، ويتمثل دور هذه الصفحات في متابعة الأخبار المغلوطة وتصحيحها، إلى جانب مبادرات من إدارات مواقع التواصل الاجتماعي كالفايسبوك من أجل إدارة إنسانية لمكافحة الفايك نيوز، وذلك بوضع شروط معينة على المستخدمين، في حين هذه الشروط حسبه صعبة التنفيذ وتجعل المستخدمين يتخلون عن استخدام هذه المواقع ويلجأون إلى مواقع أخرى، كما كشف الدكتور تفرقنيت عن تجربة لمجموعة من الخبراء الأوربيين الذين يعتمدون على خوارزميات للكشف عن الأخبار الملفقة.

وصرح الدكتور تفرقنيت بأن مواجهة الأخبار الملفقة عبر مواقع التواصل الاجتماعي أو ما يعرف بالفايك نيوز يكون بسن قوانين ردعية، لافتا إلى أن كل الدول بما فيها الدول الأوربية لجأت إما لقانون العقوبات أو سن قوانين إضافية أخرى لمكافحة الأخبار الملفقة عبر مواقع التواصل الإجتماعي.

وتحدث ضيف جمعية الصحفيين بالبليدة عن المسؤولية الإجتماعية في الوسائط الجديدة، مفيدا بأن المسؤولية صعب تحديدها عبر مواقع التواصل الاجتماعي، وتتم حاليا بإستخدام الرقابة رغم أنها مكلفة ماليا، مشيرا إلى أن صعوبة تحديد المسؤولية في مواقع التواصل الاجتماعي يعود لكونها مجهولة الهوية على عكس ما هو موجود في الصحافة.

وقال تفرقنيت بأن الوسائط الجديدة المتمثلة في مختلف مواقع التواصل الاجتماعي أحدثت تحولا كبيرا في مجال إبداء الرأي، كما ساعدت في الحشد والتعبئة في الإضرابات والاحتجاجات، بالإضافة إلى ذلك ساهمت في المشاركة في الحياة السياسية، بعد أن أصبحت هذه المواقع مفتوحة للجميع. وفي نفس الوقت ساهمت الوسائط الجديدة حسبه في تشكيل الرأي العام الالكتروني، مؤكدا بأن الوسائط الجديدة غيرت المشهد الإعلامي وانتقلت إلى التفاعلية والآنية والظرفية.

و ذكر الأستاذ تفرقنيت من جهة أخرى بأن الوسائط الجديدة لا يمكنها أن تعوض الصحافة والإعلام، مضيفا بأن الإعلام يحتاج إلى محترفين ومصداقية في نقل الخبر، وأن وظيفة مواقع التواصل الاجتماعي هي وظيفة اتصالية ولا تؤدي دور إعلامي، وما ينشر من أخبار عبر ما يسمى بصحافة المواطن يفتقد حسبه للاحترافية في نقل الخبر كما تفتقد للمصداقية، مضيفا بأن الوظيفة الإعلامية التي تقوم بها وسائل الإعلام المختلفة لا يمكن أن تلغيها الوسائط الجديدة، بل تستمر هذه الوسائل في الوجود وظهور أي وسيلة لا يلغي أي نوع آخر، وقال الوسائط الجديدة تؤدي وظيفة جيدة من حيث الاتصال، في حين ناقصة من حيث وظيفتها الإعلامية.

• ب.ر. حنان

Figure 1: Newspaper PDF Blocks -Akhbar Alyoum; May 08, 2021-

In this paper, we aim to propose: (1) An end-to-end approach which allows conducting a newspaper analytics applied to PDF newspapers; (2) A detailed approach for PDF newspaper pattern recognition, for Latin and Arabic PDF. To achieve these goals, we propose a complete approach which allows us to identify in each PDF newspaper page the different articles associated with their authors and titles. Deep learning models are trained to recognize these patterns in PDFs.

To validate our proposal, some experiments are conducted on our own constructed dataset composed of more than 1.500 PDF newspapers in Arabic and Latin languages and collected from many Algerian newspapers like Echourouk, Ennahar, El Watan, Jeune indépendant, etc. Interesting results are achieved which encourage us to consider this solution in our newspaper analytics tool which returns many KPIs and graphs via an interactive dashboard to be used by many companies and institutions.

This paper is organized as follows, section 2 summarizes the related work, section 3 presents the detailed approach and deep learning model, section 4 details the conducted experiments and achieved results, section 5 concludes the paper.

2. Related Work

The literature is abundant of works related to news and media analytics [5], computer vision and image patterns recognition [6], in both academic and industrial fields. However, a restricted number of works are interested in press newspaper data analysis despite their added value [7]. In this section, we highlight the main works and solutions related to press newspaper analysis approaches (news discourse analysis is out of the scope of our paper), and the main pattern recognition used solutions.

2.1. Newspaper analytics approaches

Press News analysis passes through many steps principally subscribed in NLP discipline. Many frameworks and approaches are proposed to analyze the news as a type of text [8] characterized by its language, structure and context. Linguistic and grammatical operators are applied on press news to extract valuable semantic properties [9] that may be used for different analytics purposes like sentiment analysis [4], financial reporting [10], Crime analytics [11] where data mining and lexicon based techniques are used.

In general, analyzing news requires the creation of a rich dataset [12], this dataset resulted from manual or automatic collection of articles, headlines [13], content [14] from archives to be used for different analytic purposes. The direct use of these datasets, in most cases, is not advised, text preprocessing is an imperative step to be performed in order to enhance the quality of text, extract features and tokens and vectorize extracted text to facilitate machine understanding.

[13] proposed a complete text mining approach to analyze the text headlines extracted from the newspaper front pages. Word cloud, word sentiment analysis and word clustering case studies are performed after manual collection and preprocessing of headlines. Newspaper articles are also used to visually explore social networks [14]. The text articles were extracted manually using a German articles corpora.

[15] proposed a complete tool to analyze and visualize the newspaper front pages. This tool

downloaded images of newspaper front pages, imported them into a local software application, and hand-coded them for coverage measures. However, this tool did not extract the content and cannot identify the blocks of articles in the whole newspaper. It focuses on coloring the areas and calculating their surfaces to show the coverage rate.
We found one work [7] which proposed a mechanism to identify photos areas and text areas in electronic newspapers. However, this work did not focus on identifying the block of each article and its metadata, it colored the whole text area with a given color and image area with another color.

2.2. Pattern recognition solutions

Image pattern recognition is a field of computer vision which studies how automated systems can process and understand digital images or videos with the aim of making them work as similarly as possible to humans [16]. Many solutions are proposed for image pattern recognition. On one hand, some approaches are traditional and called *feature descriptors* for the extraction of image features like: edges, corners, colors. SIFT (Scale-Invariant Feature Transform) [17], SURF (Speeded-Up Robust Features) [18] and BRIEF (Binary Robust Independent Elementary Features) [19]. They use a series of mathematical approximations to learn a representation of the image. However, these solutions are complex and need more expertise [20]. On the other hand, many deep learning based solutions are proposed in literature where artificial neural networks (ANN) and Convolution Neural networks (CNN) are designed to train needed models which automatically extract image features .

To summarize, many works proposed approaches to analyze news or media content. However, few works are interested in press newspapers for data analytics. Moreover, the analysis of the literature shows the absence of works dealing with PDF newspapers or proposing pattern recognition solutions to identify the blocks of press articles in each PDF page. These findings motivate our proposal.

3. Newspaper Patterns Detection Approach

To fill-in the gap in academic and industrial fields, we propose in this section, a complete approach which allows identifying in each PDF newspaper page the different articles associated with their authors and titles. This solution is widely required in the context of newspaper analytics.

This approach allows us to extract and analyze the content of press PDF Newspapers, understand their purpose, then use some metrics to present a dashboard for top management to help them making decisions. This solution focuses on press newspaper pattern recognition applied to PDF newspapers in Latin or Arabic languages.

Our approach is illustrated in figure 2, where the different steps are defined as follows:

1- Requirements Analysis: In any data-driven decision making or data analytics solution, defining and understanding the business requirements is an imperative step. These steps allow us to define the needs and expectations of this product. The developers should analyze the analytics requirements expressed by the end-users. Generally, these requirements are given as: (1) The list of newspapers needed to be analyzed (e.x. Echourouk, Ennahar, El watan, etc.);

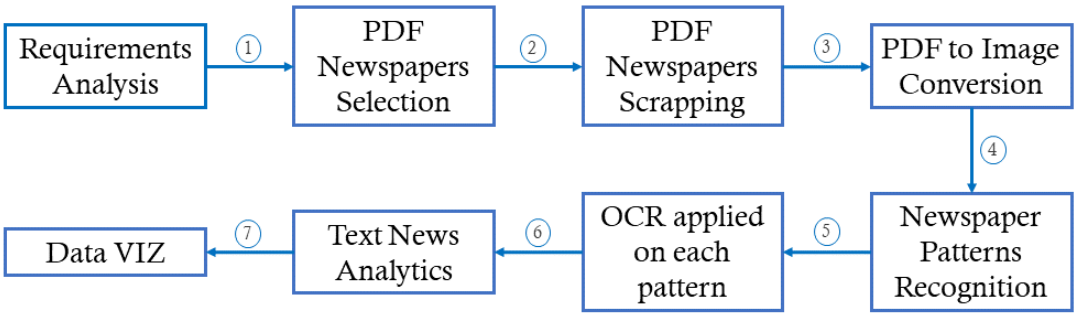

Figure 2: Newspaper Analytics Approach

(2) The frequency of dashboard updates, i.e. how many times the dashboard should be updated by new articles (e.x. update the dashboard every day at midnight.). This information is important because it allows us to define the automatic jobs wich will be run at this time of the day in order to bring the new articles and new PDFs;
(3) The list of metrics, KPIs and graphs that should be included in the dashabord for an effective data analysis;
(4) The target keywords to look for (e.x. gather articles talking about Algeria and Economy). The newspaper collection is targeted action, we bring just what we need in order to optimize processing and storage.

2- PDF Newspapers Selection: Once the requirements are well established, a list of the different newspapers that should be collected is defined and ranked by priority. At this step, developers are called to prepare the list of links of these PDF newspapers like Echourouk (https://www.echoroukonline.com/), Ennahar (https://www.ennaharonline.com/), etc. To define these links, a small manual work should be performed by engineers.

3- PDF Newspapers Scrapping: At this step, we define a python script which is used to scrape and download the different PDF newspapers then save them in a local repository. The python script is automatically executed (via a scheduled job). Scraping is an interesting way to collect data. Many libraries are available and they are open source to achieve these goals. Selenium is one of these well-known libraries widely used by data engineers.

4- PDF to Image Conversion: The different collected PDFs are automatically converted into images. This consists of converting each page of the PDF into a JPG image and storing them in a local repository needed for the next step.

5- Newspaper Patterns Recognition: This step is the main one in the process, where complex deep learning development is required to identify the different articles blocks with their metadata (authors and title). To achieve this goal, we define the approach illustrated in Figure

3. This step is splitted into different sub-steps detailed in what follows and it is an iterative solution, at each cycle, many enhancements and improvements can be added and the whole cycle should be executed.

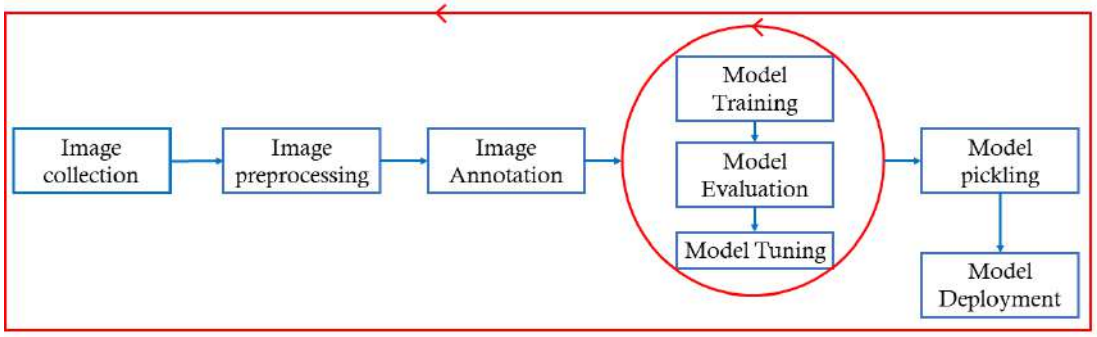

Figure 3: Pattern Recognition applied to newspaper images

5.1. Image Collection: in order to develop a deep learning model subscribed to the computer vision field, we need a voluminous dataset of annotated images. The different boundings of articles, titles, authors should be drawn manually and extract their positions.
5.2. Image Preprocessing: to achieve best results of accuracy and precision, the collected images which will be used in the training phase should be of a good quality. The image quality criteria that we defined are: clearness, luminosity, contrast, zoom. To enhance the image quality, we developed a python script based on the PIL library which curates images based on these criteria.
5.3. Image Annotation: Once the quality is treated, these images are annotated using the image annotation tool. This later allows drawing the boundaries of the article content, title and authors (as described in Figure 1). Then yolo files are exported to be used in the next step. The yolo file contains the 4 positions of each drawn boundary for each image extracted from the newpaper.
5.4. Deep learning model development: this is the core of the solution. At this phase, a python code is written to train, evaluate and tune the parameters of the deep learning model used to learn the article patterns obtained from the previous step. The deep learning model is based on Keras Tensorflow, where a CNN model is trained on the annotated dataset. We use some metrics to evaluate this model like precision, recall, loss. The CNN model we developed is composed of many layers. Three main layers should be present in any CNN model: (1) a convolutional layer is the main building block of a CNN. It contains a set of filters (or kernels), parameters of which are to be learned throughout the training. The size of the filters is usually smaller than the actual image. Each filter convolves with the image and creates an activation map[21]; (2) a pooling layer which reduces the dimensions of data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer., and (3) a fully connected layer which connects every neuron in one layer to every neuron in another layer.
5.5. Model Pickling: once we achieve good results, this model is packaged or pickled to be used to identify the patterns of any PDF newspaper.
5.6. Model Deployment: the model is deployed and consumed as an API. We give an image of a newspaper PDF to the API, which will identify the patterns and return the positions of each

boundary.

6- OCR applied on each pattern: Once these patterns are detected by the previous model, OCR techniques can be used to convert the patterns into text. At this level, some available libraries can be used. In the other case, we develop a deep learning model which recognizes the different characters of the images and transcripts them into text.

7- Text News Analytics: The extracted text from each page is then analyzed, where we verify if the defined keywords (at step 1) are contained in the text. If the text contains one of the keywords, it is stored in a database with some other metadata (publication date, authors, title, location of the image, page number, Name of the newspaper, newspaper size). Moreover, we developed an excel macro which highlights the keywords on the image.

8- Data VIZ: This last step is used to consume the data stored on the database to present it in the dashboard. We define a set of interesting graphs and metrics needed in the newspaper analytics field. Moreover, using this dashboard, the end-user can click on a given image location to display the complete newspaper page with highlighted keywords.

4. Experiments and Results

Our motivation for these experiments is guided by a real project of newspaper analytics in the R&D&I of our company. We have been working on this project for more than three years. Many deep experiments were conducted in these years before obtaining the accepted results presented in this paper.
We collected our newspapers from 10 Arabic and Latin Algerian PDF newspapers like echourouk, ennahar, el watan, el massa, depeche kabylie, etc. during 5 months. At the end, we obtained more than 1.500 PDF. Each PDF was converted into a list of JPG images to construct a dataset of more than 150.000 images (each newspaper contains an average number of 10 pages).
These images should be manually annotated to train our model of pattern detection. For this end, we hired a group of 4 annotators to do this work, we obtained 10.000 annotated images. The process of annotation is detailed in the previous section. In our case, we used an online tool which helped us to annotate the boundaries of articles, authors, titles of each image and convert this annotation into a yolo file containing the pixel positions of these boundaries.

We developed our CNN model using Tensorflow in order to allow the detection of patterns in newspapers. This model was composed of many layers mainly (convolution layer, pooling layer and fully connected layer). The obtained yolo files were splitted into training, testing and validation datasets.
To obtain better performance of processing, we used our internal servers equipped with many GPUs. The training phase takes around 16 hours.
We obtained a model with a training accuracy of 85 % and inference accuracy of 81 %. We used this model to infer results from unseen images. The example of detected patterns on new images is given in figure 5 and figure 4.

The obtained results are promising. More experiments are conducted now to improve the

Figure 4: Newspaper PDF detected patterns using our solution -Jeune independant; April 21, 2021- (to zoom the image click on: https://bit.ly/3GrPkEq)

accuracy of the model. New images are also annotated for better results. The finality of this work is: after detecting these blocks of data (article, title, authors) and an OCR is developed (out of the scope) to transform them into text then analyze it using sentiment analysis models, topic modeling, opinion mining, etc. Moreover, the obtained results helped us to create dashboards given to the end-user to evaluate the presence and reputation of the media.

A limitation, even if our work returned good results in terms of precision but still many enhancements are needed. In some cases the boundaries are not detected exactly which make the OCR not really performing. In case of complex articles where the text is splitted into many positions, our model cannot detect the article as one block. All these issues are the objects of our improvements.

Using an OCR model, we extracted the text from images to feed a database of newspaper data using MongoDB database. This latter is used to feed a dashboard containing many metrics and graphs useful for newspaper analytics in different companies.

5. Conclusion

News / Press data are valuable sources of information which are widely considered in academic and industrial worlds. The correct exploitation of these data helps in improving many services and satisfying the end-users. PDF newspapers are one of these important sources which require more attention for effective analytics.

The exploitation and analysis of PDF newspapers provide companies and institutions with many elements of information about their brand image, their presence in media, their coverage by newspapers, etc. It may also have other advantages like: security management, risk analysis, competitors tracking, etc. All these data help in proactive decision-making and prevent actions.

Having these motivations in mind, and because of the lack identified in academic and industrial worlds, we propose a complete end-to-end approach for newspaper analytics. Moreover, a special focus is given to the newspaper pattern recognition module, which consists of recognizing in each PDF newspaper page, the different blocks of articles and their associated metadata (authors and title) using advanced deep learning techniques. This information is required for future use, after applying OCR techniques to transform the detected patterns into text.

Interesting experiments are conducted on our own constructed dataset which is composed of thousands of images extracted from different Algerian Latin and Arabic collected PDF newspapers. These results allowed us to trust our approach and used it in our developed tool dedicated to our customers. Our newspaper analytics tool provides an interactive dashboard with very interesting KPI required by the end user to analyze their coverage by newspapers.

As a perspective, we are working on the OCR module which allows extracting text from each block of the detected article and their associated metadata. Interesting metrics can be derived from text analysis of the extracted articles to be used for a rich newspaper analytics experience.

Acknowledgments

We would like to thank Mr Farid GHANEM and Mr Samir TAGZOUT for their efforts for the success of this project and to achieve the defined objectives.
We would also like to thank our intern students [ABOUD Ibrahim, KETFI Hibet Allah, BOUGUESSA Wail] who helped us to prepare and annotate the dataset of images.

References

[1] K. S. Kumar, J. Desai, J. Majumdar, Opinion mining and sentiment analysis on online customer review, in: 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), IEEE, 2016, pp. 1–4.

[2] H. Zhang, L. Zhao, S. Gupta, The role of online product recommendations on customer decision making and loyalty in social shopping communities, International Journal of Information Management 38 (2018) 150–166.

[3] J. Krumm, N. Davies, C. Narayanaswami, User-generated content, IEEE Pervasive Computing 7 (2008) 10–11.

[4] S. Taj, B. B. Shaikh, A. F. Meghji, Sentiment analysis of news articles: a lexicon based approach, in: 2019 2nd international conference on computing, mathematics and engineering technologies (iCoMET), IEEE, 2019, pp. 1–5.
[5] D. Zeng, H. Chen, R. Lusch, S.-H. Li, Social media analytics and intelligence, IEEE Intelligent Systems 25 (2010) 13–16.
[6] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proceedings of the IEEE 98 (2010) 1031–1044.
[7] P. Fyfe, Q. Ge, Image analytics and the nineteenth-century illustrated newspaper, Journal of Cultural Analytics 3 (2018) 11032.
[8] T. A. Van Dijk, News as discourse, Routledge, 2013.
[9] T. A. Van Dijk, News analysis: Case studies of international and national news in the press, Routledge, 2013.
[10] G. Mitra, L. Mitra, The handbook of news analytics in finance, John Wiley & Sons, 2011.
[11] I. Jayaweera, C. Sajeewa, S. Liyanage, T. Wijewardane, I. Perera, A. Wijayasiri, Crime analytics: Analysis of crimes through newspaper articles, in: 2015 Moratuwa Engineering Research Conference (MERCon), IEEE, 2015, pp. 277–282.
[12] M. Reason, B. García, Approaches to the newspaper archive: content analysis and press coverage of glasgow's year of culture, Media, Culture & Society 29 (2007) 304–331.
[13] A. Hossain, M. Karimuzzaman, M. M. Hossain, A. Rahman, Text mining and sentiment analysis of newspaper headlines, Information 12 (2021) 414.
[14] A. Kochtchi, T. v. Landesberger, C. Biemann, Networks of names: Visual exploration and semi-automatic tagging of social networks from newspaper articles, in: Computer graphics forum, volume 33, Wiley Online Library, 2014, pp. 211–220.
[15] S. Costanza-Chock, P. Rey-Mazon, Pageonex: New approaches to newspaper front page analysis, International Journal of Communication 10 (2016) 28.
[16] C. Orhei, M. Mocofan, S. Vert, R. Vasiu, End-to-end computer vision framework, in: 2020 International Symposium on Electronics and Telecommunications (ISETC), IEEE, 2020, pp. 1–4.
[17] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2004) 91–110.
[18] H. Bay, T. Tuytelaars, L. V. Gool, Surf: Speeded up robust features, in: European conference on computer vision, Springer, 2006, pp. 404–417.
[19] M. Calonder, V. Lepetit, C. Strecha, P. Fua, Brief: Binary robust independent elementary features, in: European conference on computer vision, Springer, 2010, pp. 778–792.
[20] S. Khan, H. Rahmani, S. A. A. Shah, M. Bennamoun, A guide to convolutional neural networks for computer vision, Synthesis lectures on computer vision 8 (2018) 1–207.
[21] S. Mostafa, F.-X. Wu, Diagnosis of autism spectrum disorder with convolutional autoencoder and structural mri images, in: Neural Engineering Techniques for Autism Spectrum Disorder, Elsevier, 2021, pp. 23–38.

BELKACEM SAHLI :

«La lutte contre la corruption est avant tout une question de culture»

C'EST la salle du Petit Théâtre de la maison de la Culture Mouloud-Mammeri de Tizi Ouzou que le secrétaire général de l'Alliance nationale démocratique (ANR), Belkacem Sahli, a choisi, hier, pour développer un discours sur la situation du pays, avec pour comme toile de fond les élections législatives prévues pour le 12 juin prochain, et ce en dépit d'une assistance très peu nombreuse (20 personnes dont cinq policiers).

L'orateur n'a pas caché la situation délicate que vit le pays, et ce tant sur les plans économique, social et politique. Il faut reconnaître que Belkacem Sahli, contrairement à ses discours d'autrefois, marqués par une volonté de caresser les décideurs du pays dans le sens du poil, s'est montré hier plus dur et plus explicite. En effet, le secrétaire général de l'ANR n'a pas caché son mécontentement devant «la mise à l'écart» des partis politiques pour faire monter à leur détriment ces organisations dites de la société civile. «Si l'on ferme les médias aux personnels politiques et qu'on ne leur assure pas une réelle formation politique, il ne peut être attendu, dans ces conditions, que la faiblesse de leur rendement», a-t-il assuré.

Abordant le volet portant sur la Constitution, l'orateur a reconnu que son contenu est de qualité mais «son efficacité ne peut être constatable qu'à long terme». S'agissant de l'Observatoire de la société civile, Belkacem Sahli, après avoir relevé ses missions et son rôle que, théoriquement, il devrait jouer dans la société, a salué l'initiative mais a conditionné sa réussite par la compétence de ses membres et l'efficacité de ses mécanismes fonctionnels. «Dans ce sens, nous souhaitons voir ses membres élus et non désignés par l'appareil d'Etat», a-t-il ajouté.

«L'ANR ne peut que saluer une société civile bien réelle dans la mesure où elle assure un contre-pouvoir», a-t-il expliqué. A une question posée par un membre de l'assistance concernant la meilleure d'assurer une réelle lutte contre la corruption, le secrétaire général de l'ANR a répondu que la lutte contre la corruption «est avant tout une question de culture. C'est pendant sa tendre enfance qu'on doit apprendre au citoyen le négativisme de la corruption et le positivisme de se contenter de ce que l'on a gagné». Il a ajouté que la lutte contre ce phénomène destructeur dépend également de l'indépendance de la justice, de la solidité des institutions de la République et, bien sûr, de la bonne moralité de la société.

Enfin, Belkacem Sahli a cité la numérisation comme excellent outil de lutte contre la corruption dès lors que tous les marchés et tous les dossiers susceptibles de contenir le sceau financier seront identifiés et connus de toutes les parties concernées.

Cependant, l'orateur a admis qu'il ne sera jamais mis fin, dans l'absolu, à la corruption. «Toujours est-il cependant, avec l'indépendance de la justice et la politique de la transparence, ce phénomène sera considérablement réduit», a relevé l'orateur, avant de promettre à l'assistance qu'il reviendra à Tizi Ouzou dans le cadre de la campagne électorale pour les législatives.

Saïd Tissegouine

LÉGISLATIVES ANTICIPÉES DU 12 JUIN PROCHAIN

Candidature libre : mirage et désenchantement

C'est un insurmontable parcours de combattants auquel sont livrés les candidats indépendants pour les prochaines élections législatives anticipées prévues le 12 juin prochain. Difficultés d'obtenir des parrainages, lenteurs de l'administration et un financement rachitique. Le Jeune Indépendant a suivi dans cet hypothétique périple vers l'hémicycle de Zighoud Youcef, des candidats libres armés de leur foi et de leur espoir face à des mastodontes électoraux rompues aux rouages de l'administration et de l'art de la persuasion. Ils sortiront de cette expérience métamorphosés.

Rabie ne sera pas candidat indépendant aux futures législatives du 12 juin. Il vient de renoncer à cause d'une affaire de paperasse. Un document fiscal qui lui fait défaut. Les services des impôts se montrent inébranlables et ne veulent rien savoir. Rabie doit payer des amendes et effacer un petit reliquat. Pour lui, un simple extrait de rôles vaut de l'or.

Le rendez-vous avec son destin politique est raté. Il devra encore patienter cinq ans. La courte expérience de sa candidature avortée fut un véritable parcours de combattant. Presque un défi. «Je ne pensais pas que la politique est une affaire de connaissance, de réseaux, d'argent. Je voulais tenter la députation, c'est le désenchantement».

Philosophe en ces jours de jeûne, Rabie freine ses ardeurs politiques et hirakistes. «Le mieux est de prendre tout cela comme une leçon. Une candidature indépendante pour la députation est un rêve fou. C'est aussi un mirage. Désormais, je ne croirais plus aux promesses et aux paroles».

Comme Rabie, Messaoud T. confie sa vaine tentative de briguer un siège à la future Assemblée populaire nationale. Gérant et propriétaire d'une petite entreprise, il ne se doutait pas un seul instant que cette aventure est l'une des plus compliquée. « Je pensais avoir vu et vécu tous les enfers de la bureaucratie, quand j'ai lancé il y a quatre ans ma boite. Avec cette affaire des législatives et des listes indépendantes, j'ai dû baisser les bras et abandonner. Il y a trop d'obstacles et d'embûches.»

Messaoud pointe du doigt tous les lourdeurs et les entraves bureaucratiques. Selon lui, l'affaire des législatives a été une opération qui manquait du temps, mal préparée par l'administration et l'Autorité nationale indépendante chargée des élections.

«Chaque jour, on apprend une nouveauté dans la règlementation. Il y a trop de procédures, trop d'exigences. J'aurais dû choisir un parti politique. C'est plus facile et moins contraignant».

Issu des gestations du mouvement citoyen du hirak, des marches hebdomadaires et des slogans des vendredis, Messaoud avoue qu'il ignore les dispositifs de la loi électorale et des complexes conditions d'éligibilité exigées dans ce processus. Pour lui, comme pour d'autres jeunes, ce sont les promesses d'aides aux financements étatiques dans sa campagne qui l'ont poussé à tenter le diable.

Dépité, Messaoud pense que la loi a donné des facilitations aux partis, mais pas aux candidats libres ou aux listes indépendantes. Rien qu'avec cette histoire de parrainage, de validation des formulaires des signatures, il tourna en rond pendant des jours et des nuits.

Certains citoyens estiment que la difficulté demeure cette quête de signatures en faveur des candidats. "L'opération est délicate et très ardue. Pour gagner les élections, il faut gagner d'abord ces signatures. Ce n'est pas une bataille aisée même pour les formations politiques", nous dit Brahim, un politique aguerri, qui a fait ses classes au sein du vieux parti le FLN.

Il signale que les jeunes ne sont pas totalement portés sur la chose politique et ignorent presque tout de l'action électorale. Il a indiqué que même les partis politiques les plus connus n'arrivent pas à trouver le bon créneau pour mobiliser au moins une centaine de jeunes par commune, qui serait la base arrière de tout activité politique.

L'ARGENT LE NERF DE LA CAMPAGNE

Brahim pense aussi que durant les dernières années, la vie politique comme les partis politiques n'ont pas fonctionné réellement, se contentant des quotas attribués à chaque élection législative ou communale. Les débats politiques n'existaient pas. Les médias étaient fermés aux émissions d'actualité et cela ne permettait pas d'émergence de nouvelles figures, de découvrir des têtes. Même la proximité a disparu.

Cependant, et contrairement aux partis politiques, qui connaissent le milieu de la bureaucratie et les toiles de l'administration, possédant des cadres aguerris et rompus à ces parcours électoraux, ces jeunes candidats libres, novices pour la plupart, semblent perdus et affaiblis à l'avance par la compétition.

Certains guidés par l'ambition s'en mêlent et s'engagent dans certaines formations lilliputiennes. D'autres actifs dans le mouvement associatif s'appuient sur leurs camarades de la société civile pour confectionner une liste et la déposer dans les délais. Mais ce n'est pas tout.

Car d'abord, une candidature à la députation est une affaire de financement et de grosses dépenses notamment face aux imprévus. Tout dépend de la taille des wilayas, des agglomérations, des régions et de sa démographie.

La nouvelle loi sur le régime électoral a voulu mettre des paravents sur l'influence de l'argent dans les choix des candidatures, elle oblige les candidats libres à chercher de l'argent pour financer les premières démarches à travers les quartiers, les cités populaires, les villages et autres hameaux. En dépit des facilitations offertes par l'Etat pour encourager les jeunes à s'intégrer dans la vie politique, à travers la gratuité des salles destinées aux meetings et autre rassemblements, ainsi que la disponibilité des panneaux publicitaires, des candidats trouvent des difficultés pour former leurs listes de candidatures, pour engranger le nombre requis des parrainages.

La crainte de retomber dans les mêmes travers des précédentes expériences pèse lourdement. Certes, il n'y a plus de hiérarchisation, plus de tête de liste, le mode du scrutin a permis à l'électeur d'être libre de choisir le candidat qu'il préfère sur une longue liste.

Le gros nabab qui achète les sièges, les candidats, les parrainages et les voix de l'urne le jour du vote, n'existe plus avec le nouveau dispositif.

La représentation proportionnelle au vote préférentiel sur une liste ouverte est un acquis immense pour les jeunes candidats. Or, c'est la confection d'un programme commun, une liste commune et une campagne collective qui deviennent une mission impossible.

Avant, le jeu était fermé pour les ingénieurs, les médecins, les architectes, les enseignants universitaires, les jeunes chercheurs et les intellectuels libres, car les partis politiques les refusaient et s'en méfient. Libérer toute cette énergie de cadres en si peu de temps pour rentabiliser une élection et bien encadrer un futur Parlement est certes louable, comme initiative politique.

L'idée pourrait bien être comme une émanation de l'esprit du hirak populaire et démocratique.

Mais, le temps court dans l'annonce de la dissolution de l'APN et l'organisation des élections législatives anticipées suffiront -il pour changer les mentalités, et les pratiques qui ont fait reculer l'Algérie?

Mohamed Kouini

Figure 5: Newspaper PDF detected patterns using our solution -Le Jeune Indépendant; April 18, 2021- (to zoom the image click on: https://bit.ly/3QlHDEm)