Analyzing the Needs of the Offshore Sector in Morocco by Mining Job Ads

Imane Khaouja[1,2], Ibrahim Rahhal[1,2], Mehdi Elouali[1], Ghita Mezzour[1], Ismail Kassou[2], Kathleen M Carley[3]

[1]Université Internationale de Rabat, FIL, TICLab, Morocco

[2]Université Mohammed V Rabat, Morocco

[3]Institute for Software Research

Carnegie Mellon University Pittsburgh, PA, 15213 USA

Abstract—The offshore sector creates a large number of job opportunities in Morocco. Analyzing job ads related to that sector can help universities adapt their curricula in order to produce more employable graduates. Unfortunately, analyzing these ads is challenging because they are mostly non-structured. Most prior work, however, focuses on analyzing structured and semi-structured job ads through keyword search and regular expressions.

In this work, we collect and analyze job ads related to the offshore sector in Morocco over the period February-August 2017. We use a variety of machine learning and text mining techniques to process these ads. We examine the natural languages, programming languages, education level, and years of experience needed. We also examine contract type and salary when available. Our results reveal that French is the most needed language in offshore jobs, but that other foreign languages (English and Spanish) are also needed. We also find that the most needed offshore IT jobs are development and web design jobs. The most needed programming languages for these jobs are Java, SQL, JavaScript and PHP.

Keywords-Job market needs; Text mining; Non-structured text; Offshore.

I. INTRODUCTION

The offshore sector is a leading sector for job creation [1] in Morocco. That sector encompasses two major areas: Business Process Outsourcing (BPO) and Information Technology Outsourcing (ITO) [2]. BPO includes general administrative activities and functions, activities for management of client relationships and business-specific activities, while ITO combines infrastructure management, software development and software maintenance. Understanding the needs of the offshore sector could help universities adapt their curricula in order to increase the employability of their graduates within that sector. Unemployment of young graduates in Morocco is as high as 18% [3], partly due to a mismatch between education and the job market needs [4].

An interesting approach to identify the needs of the offshore sector is to mine online job ads pertaining to that sector. Job ads contain interesting information such as skills, education level and years of experience. Unfortunately, extracting such information is complicated because the majority of ads pertaining to the offshore sector in Morocco are either non-structured or semi-structured. Moreover, selecting job ads related to the offshore sector out of all job ads is non-trivial because that sector is diverse. However, most prior work focuses on mining structured and semi-structured job ads, where extracting information from job ads is relatively straightforward. Moreover, the majority of prior work uses simple keyword search to identify job ads pertaining to the sector of interest to that work.

In this paper, we collect all job ads from 10 popular job posting websites in Casablanca, the offshore hub in Morocco. We train a machine learning classifier to identify ads pertaining to the offshore sector. The machine learning classifier was trained on a subset of the data that we manually labeled and achieved 90% accuracy. Subsequently, we remove duplicate job ads using Simhash, an algorithm for detecting near duplicates. Finally, we use regular expressions to extract skills and other relevant information from ads and used Latent Semantic Indexing (LSI) to homogenize job titles (for example, both "HR assistant" and "Human Resources assistant" are mapped to "Human Resources assistant").

Our results reveal that offshore ITO jobs require significant higher education (50% of these jobs require 5 years of higher education). Moreover, the list of top programming languages required by these jobs seems to differ from the list of top programming languages internationally and from the focus of Moroccan university curricula. For example, JavaScript and PHP are among the top needed programming languages, whereas C and python are not. Finally, ITO jobs also require foreign languages where the focus is mainly on French and English.

We informed our department about our findings and we are in the process of adapting our curricula to better match the needs of the offshore market. For example, the focus on JavaScript and PHP will increase in the upcoming years. Such change in the curricula will not be a huge burden since JavaScript and PHP are relatively simple languages. On the other hand, we believe that such increased focus on these languages will improve the employability of our graduates within the offshore sector.

On the contrary, we find that the most BPO jobs are call center jobs. These jobs require little higher education and provide decent salaries (almost the double of minimal wage). The main focus of these jobs is foreign languages, mainly French. This means that these jobs could be interesting for a significant number of Moroccan youth that have no higher education. These young people would only need to take regular

or online courses to help them master French.

II. RELATED WORK

Several studies have examined the job skills required for the job market. In the last decade mining tools were used to catch the job market requirements and needs. The most studies examined the IT sector and competencies ontology.

In Huang et al. study [5], they compare the IT job skills from the academic and practitioner literature as well as job advertisements using meta-analysis. Not only their job ads sample is small (240 ads) but also, they synthesized only IT skills. In this paper, we study a sector that encompasses several areas.

Litecky et al. [6] extracted job ads with computing program degrees from five websites using keywords, then they extracted skills using a thesaurus. Crawling websites using keywords can miss some job ads since the keyword list can not be exhaustive. This paper hopes to remedy this situation by gathering all job ads then filtering them using machine learning tools.

Janotk [7] showed Labor Market Trends in the Area of Information Technology in Slovakia by analyzing the data and surveys from a job portal. He analysed job ads from only one website which facilitate the analysis. In our analysis, we deal with several websites which allow catching the job market dynamics online. As each website has its own specificities, more attention is paid to the standardization of the extracted information from job ads.

In [8], They went from candidates to build a knowledge base of skills in order to extract top skills form job ads. Their study harvest skills from candidate profiles on social media. In fact, this information remains inaccessible in current professional networks. Also, they overlook the correlation between attributes. Our work tries to investigate the interaction between the extracted features.

III. DATA

In order to identify websites that contain job advertisements, we used Optioncarriere.ma a job aggregator. This website helps us identify the most frequently used websites of job ads in Morocco. We scrapped these websites weekly between February 2017 to August 2017 specifying the region of Casablanca (see fig. 1).

In fact, crawling websites using keywords can miss some job ads since the keyword list can not be exhaustive, especially in the offshore sector that contains multiple areas and a multitude of occupations. Also, most collected job ads are submitted in a non-structured format which lacks a predefined category where the research by keywords could be done. We remedy this situation by gathering all job ads regardless of their type, then we filter them using a machine learning approach to automatically choose the offshore job ads.

At this step, we collected all the job ads from 10 websites as shown in table I. Some of these websites have a non-structured format. We extracted 128.000 job advertisements counting duplicates where the majority of the data collected have a non-structured format and need some preprocessing.

TABLE I: distribution of job ads

Website	% of job ads	Data type
www.avito.ma	46.3%	Non-structured
www.jobs-ma.com	32.9%	Non-structured
www.emploi.ma	4.7%	Semi-structured
www.moncallcenter.ma	3.6%	Semi-structured
www.rekrute.com	2.9%	Semi-structured
www.jobmaroc.com	2.2%	Semi-structured
www.embauchemaroc.com	1.5%	Semi-structured
www.m-job.ma	1.4%	Semi-structured
www.marocemploi.net	0.8%	Non-structured
www.recruteonline.com	0.3%	Semi-structured

The structured documents are defined as documents having a significant structure [9]. Semi-structured documents are text files that contain data that does not have to conform to a fixed schema, however have some organizational properties that make it easier to analyze. Non-structured text lacks these structuring characteristics. In job ads case, structured job ad refers to an ad where all job ad characteristics are set apart (Job description, Enterprise description, Qualification, Mission and other features). Non-structured job ad is formatted in free text, it contains little information about the job ad characteristics. like Avito (see table I) which is a general ad website that is very popular in Morocco, also employers publish their job ads, in a non-structured manner, in order to target more candidates.

IV. DATA PREPROCESSING

A. Deduplication

As recruiters tend to submit their job ads in different online platform to target more candidates, we gather the same job ads several times since we collect job ads from different websites. To address this issue, we should detect near duplicates in our database, considering job ads differ from one website to an another.

Manku et al. [10] showed that Simhash is practically useful for identifying near-duplicates in web documents belonging to a large page repository. Simhash is a fingerprinting technique where the fingerprints of near-duplicates differ in a small number of bit positions. Following Manku results, we apply the Simhash technique to hash the job ads then we calculate the hamming distance between the fingerprints.

We manually evaluate duplicated job ads in two different samples of 500 records then we compare it with our near duplicates results. The reported accuracy is 90% for both testing samples. Our original data contained more than 128.000 job ads, after deduplication, our data shrank to 55.264 job openings.

B. Offshore versus non offshore job ads

Since most collected job ads are submitted in a non-structured format, a research by keywords could miss some job ads. This is why we collected all job ads then we filter them using machine learning approach to automatically choose the offshore job ads. This approach consists of text categorization, a general inductive process that automatically
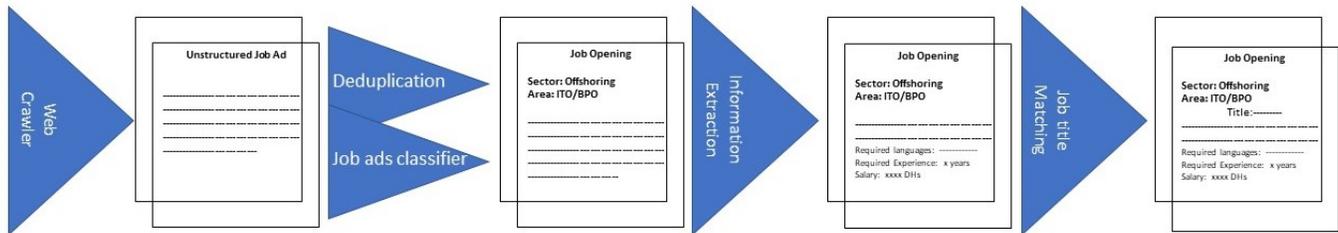
Fig. 1: Methodology

TABLE II: Classifiers Measures

Website	Precision	Recall	Fscore	Accuracy
www.avito.ma	87%	89%	88%	90%
www.jobs-ma.com	90%	91%	90%	91%
www.emploi.ma	90%	89%	90%	92%
www.rekrute.com	94%	95%	95%	95%
www.jobmaroc.com	95%	94%	94%	95%
www.embauchemaroc.com	90%	92%	90%	91%
www.m-job.ma	95%	94%	94%	95%
www.marocemploi.net	88%	81%	83%	89%
www.recruteonline.com	91%	90%	91%	90%

TABLE III: Distribution of offshore job ads

Area	Number of job ads
ITO	1514
BPO	26574

builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories [11].

To do this, we labeled a training dataset in order to feed the classifier then we tested on testing data. In the first, place we manually labeled a dataset containing all websites job ads, which gives an accuracy of 85%. Then we built a classifier for each website apart, which improves the accuracy up to 90%. A job ad is considered an offshore job ad if the job description is related to the offshore jobs and tasks, or the enterprise description characterize an offshore enterprise. We used several classifiers, the best accuracy was reported for support vector machine and naive bayes classifier (see table II).

Likewise, we built a classifier to automatically set ITO and BPO job ads apart. The accuracy is around 95%. This level of accuracy guarantees the reliability of the classifier (see table III). After applying the classifier, we remain with approximately 28.000 offshore job openings. Our analysis will be performed on these job openings.

C. Attribute extraction

Attribute extraction is performed using regular expressions. Some General attributes are already specified in some job ads. However, other attributes are unspecified such as experience, languages, salary and type of contract especially in non-structured job openings where we used regular expressions to extract them from the free text. In fact, we used dictionaries

(programming languages and speaking languages) coupled with regular expressions to detect the language required in the job ad. Once again using regular expressions and keywords we retrieve other features such as: salary, type of contract and the years of experience required. For the salary, we analyzed a range of job ads in order to set words that may indicate a remuneration (such as dirham, primes, remuneration etc.), then we associate these words with regular expressions to extract the salary, then we delete the outliers. We performed the same approach to extract experience.

D. Job title matching

Considering that job ads are submitted by different entities, they may employ different naming for the same occupation. Latent Semantic Indexing (LSI) can match job advertisement extracted from the Web with occupation description As shown in [12].

Latent Semantic Indexing (LSI) is a method that captures the hidden semantic relationships between documents by using techniques from linear algebra. Where Each document from the corpus is represented as vectors in a vector space. Each position in a vector represents the presence or not of the term in the document by, a positive value or a null value otherwise. Then these vectors are projected in a low-dimensional space obtained by singular value decomposition of the term-document matrix [13].

To increase the reliability of this technique we separate the ITO and BPO job ads and occupations. We took occupation description from national repositories [14].

V. RESULTS

A. Occupations

In the ITO, Analyst & Developer, Project manager alongside with Network manager are the most demanded occupation as shown in table IV, which means that ITO requires technical skills. On the other hand, BPO requires mostly soft skills such as languages and the sense of communication as depicted in table V, the call center agent is the most required occupation in the BPO. Consequently, human resources assistants are required in order to manage and hire the call center personnel.

TABLE IV: Most in demand occupations in ITO

Occupation	%of job openings
Analyst & Developer	13%
Project Manager	13%
Network Manager	10%
Database Administrator	9%
Web Designer	9%

TABLE V: Most in demand occupations in BPO

Occupation	%of job openings
Call center agent	60%
Human Resources assistant	29%
Client advisor	7%
Teleservice platform manager	1%
Accountant	0.35%

Fig. 2: Education level needed in job openings

B. Education Level

We can clearly see in fig. 2 that the ITO area requires a high education level. Conversely, the BPO area is lenient with qualifications which is mainly due to the fact that Call centers need mostly low skilled workers.

ITO requires high educated profiles such as developers, network and project managers and web designers. These jobs need technical competences which are taught in engineering schools, universities and vocational schools which pertain to the higher education. In fig. 3b, most of the job openings require an engineering degree especially for project managers.

Most job openings in BPO are call centers agents (see table V). Call centers train their chosen candidates, where their initial requirements are soft skills and foreign languages (see fig. 6). This is why BPO requires fewer years of higher education than ITO. The fact that a considerable fraction of job openings in the BPO does not require higher education, represent a good employment opportunity for the youth with no higher education. In fact, the HCP cites that more than 30% of youth is not engaged in work or education [15]. As we see in fig. 3a, call center agent and human resources assistants have no condition regarding higher education. Where the initial wages in the BPO, as depicted in fig. 7, starts from approximately 4.000 dirhams which is almost the double comparing to the minimal wage in Morocco (2.500 dirhams). Knowing that these wages can increase with supplementary hours.

C. Programming languages

Java is the leading programming language in the ITO area as we see in fig. 4. In the 2017 IEEE spectrum ranking, we find that Python is the top programming language and Java dropped to the third position where it was the first one in 2015 [16], [17]. In Databases SQL is the most required language in ITO job openings then come Oracle and MYSQL. Our results give us an insight into required programming languages in ITO, where they differ from international ranking. This difference confirms that Offshore sector in Morocco has its own specificities that should be examined closely. Although C language is widely taught in Morocco in IT education,
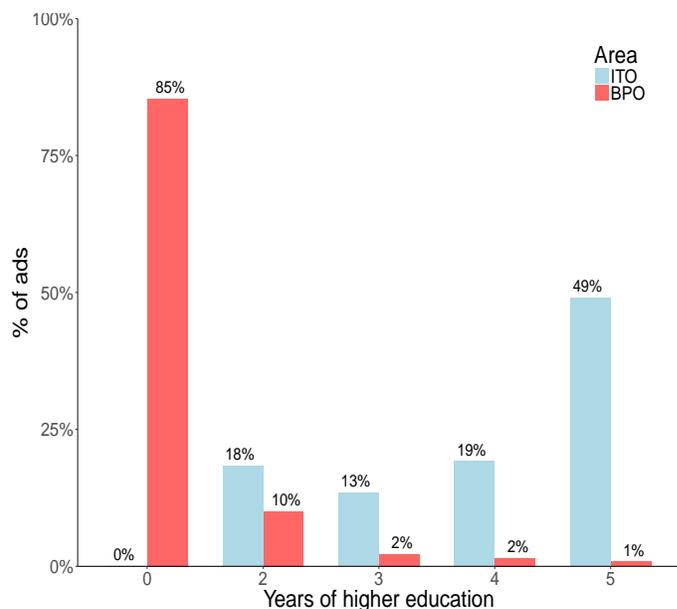
we can clearly see its absence from the top 10 programming languages in fig. 4. Conversely, Java language is barely taught, which can explain the mismatch. Emphasizing Java language in universities curricula can help youth to increase their employability.

In fact, the required programming ranking changes from occupation to another. In order to adapt university curricula for each occupation apart, fig. 5 gives an insight into programming languages for each occupation. For Analyst & Developer, JavaScript and PHP are the most required languages which are used in web-based applications. Project managers are more asked for Java language for application development. However, network managers should have been trained in operating system languages such as Linux also should have a knowledge of database languages. Project managers training should involve a training for programming languages with a focus on Java technology. Regarding network manager training, operating system languages (LINUX, UNIX and Shell) should be taught alongside with database languages.

D. Natural languages

Morocco follows the French system, English is the second language studied in high school, which explain the demand for French and English languages in call Centers as shown fig. 6. The Spanish language is also requested, the government is indeed making efforts to include the Spanish language in some training seen the closeness to Europe and specifically Spain. We see in fig. 7 that other foreign languages such as English, Spanish and Dutch are paying more than other languages which highlight the importance of learning a third foreign language. Learning foreign languages aside from French can attract other European companies in Morocco.

In ITO, French is the most required language followed by English. Morocco is the first destination for the French
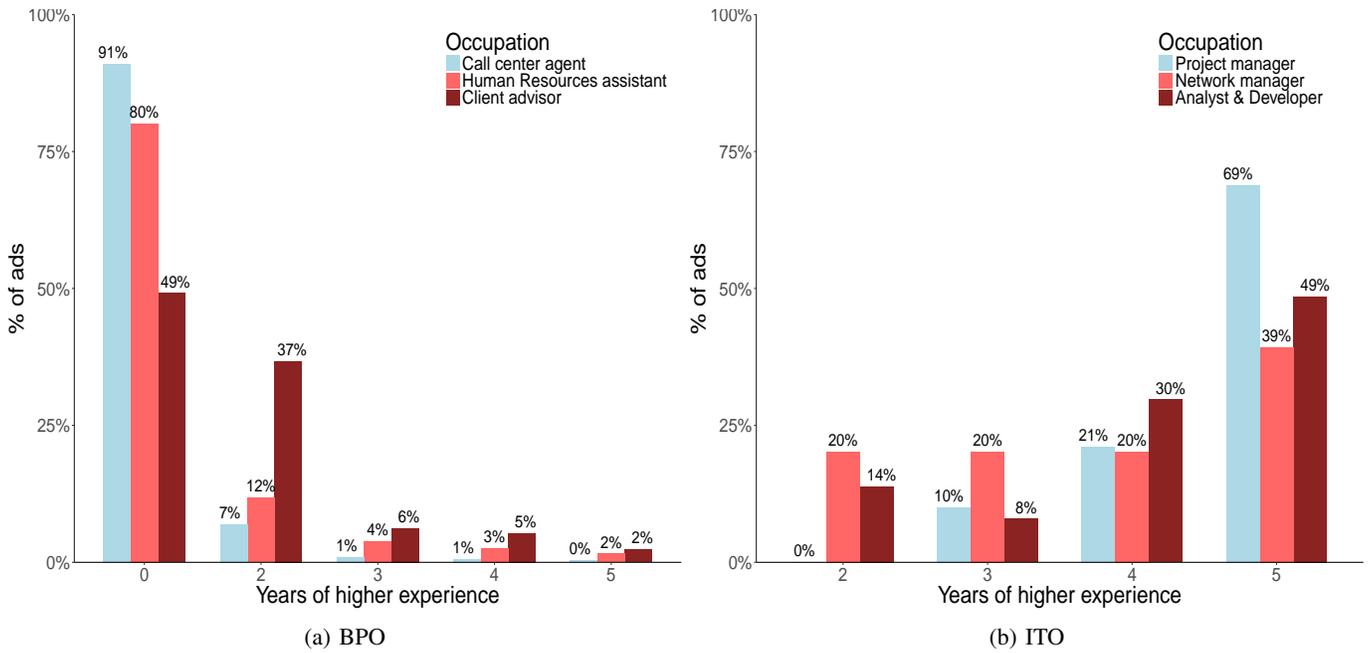
(a) BPO

(b) ITO

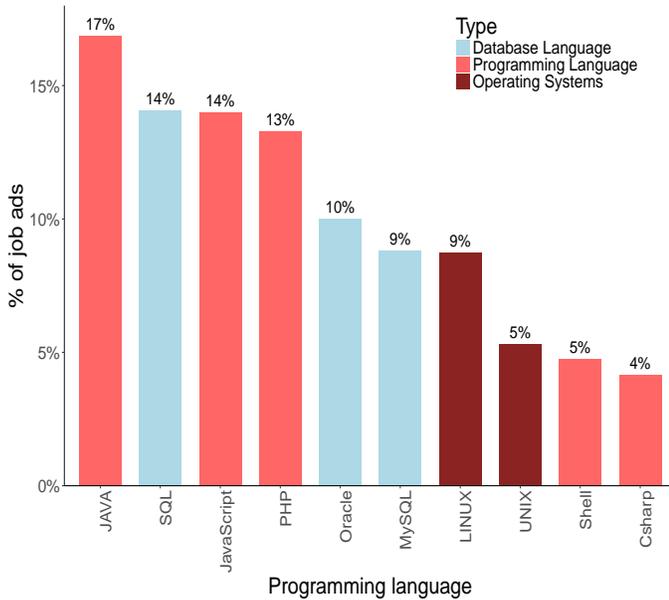Fig. 3: Education level vs. Occupation

Fig. 4: Top programming languages in the ITO

offshore which explain the dominance oh the French language [18]. However, programming languages require a proficiency in English since all programming languages are writing in English. Other European countries choose Morocco to outsource their activities, such as Spain, Germany, Belgium and Netherlands which explain the presence of other languages in fig. 6. Although they represent a small sample of the job openings, they may be an early indicator for the offshore job market trends.

Foreign languages are also required for project managers

as depicted in fig. 8b such as French, English and Spanish. Network managers are asked for a variety of foreign languages such as English, Spanish, Dutch, German and Portuguese languages as we see in fig. 8b. In the BPO, foreign languages are equally disturbed between the three main BPO occupations as shown in fig. 8a.

E. Contract types

The distribution of contract type in the other openings is quite normal as we see in fig. 9, 10, since the majority of the openings are permanent. The internship contract is quite in demand in the job openings, which can be explained by the efforts made by companies to prepare students for offshore especially in the ITO. On the other hand, the percentage of freelance is very low which is probably due to the incoherence of the legislation of this type of contract. A high rate of permanent contract indicates the stability and the health of the economy. Besides, it indicates the continuity of the offshore expand in Morocco. Some Job openings do not contain information about contract type, they are marked as unspecified. This problem occurs very often in BPO job openings rather than ITO job openings.

F. Experience

The salary rise is not statistically significant with the years of experience as we see in fig. 11, which can be explained by the weak presence of job openings requiring a high level of experience in online BPO job openings, see fig. 12a. Also, these job openings tend to omit the salary, which is specified afterwards during the job interviews.

Most of project managers, network managers and developers are asked for a minimum year of experience as depicted in fig. 12b, which highlights the fact that graduates should seek for

(a) Analyst & Developer

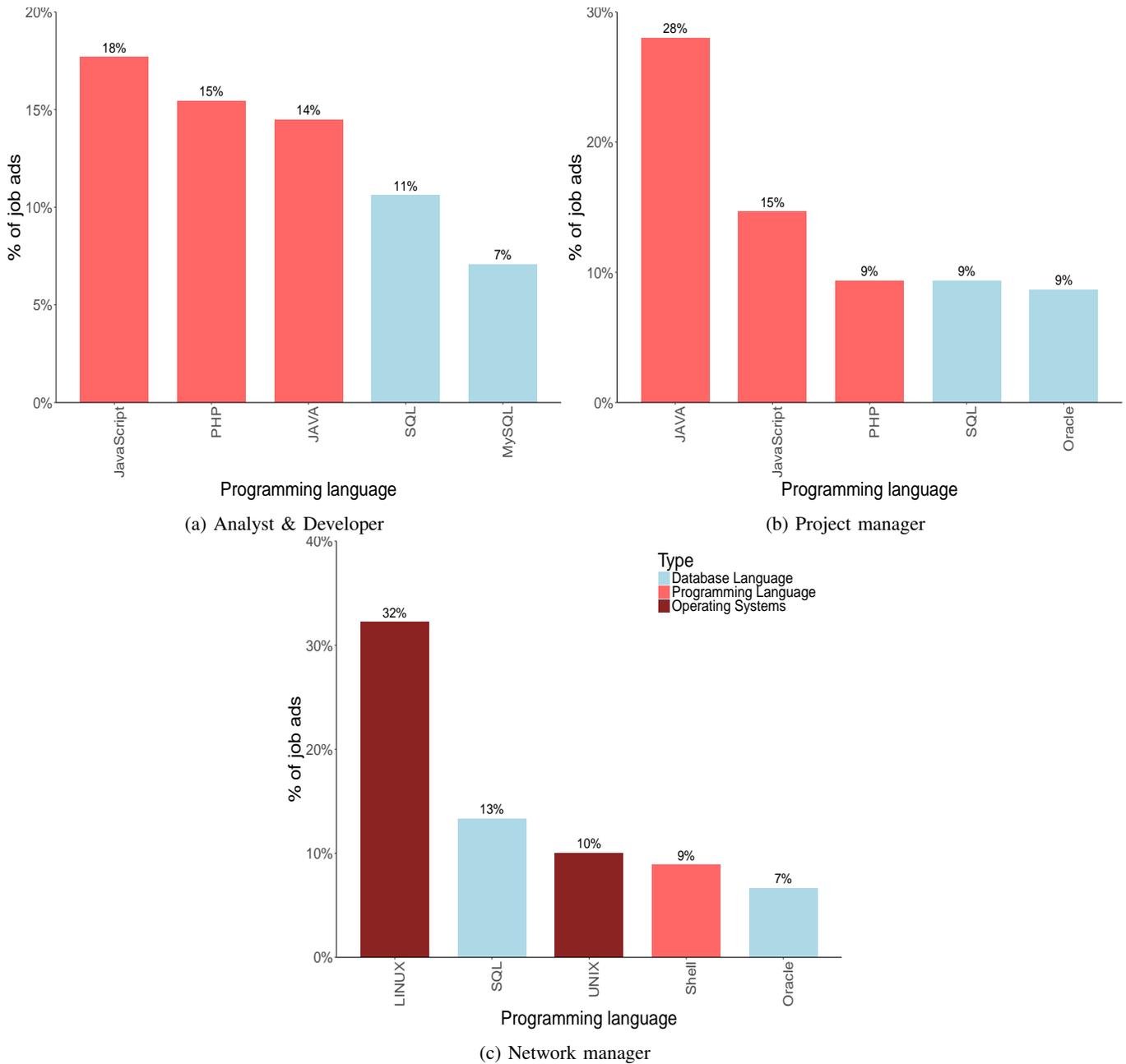(b) Project manager

(c) Network manager

Fig. 5: Programming languages vs. Occupation

internships in order to gain experience. In BPO, a considerable fraction of job openings, for the three top occupations, do not require a previous experience as we can see in fig. 12b, which confirm the role of the BPO in youth employment with no higher education nor a prior experience.

VI. LIMITATIONS AND FUTURE WORK

Our paper presents some limitations. These limitations reside in the fact that we cover only online job ads.

Due to the non-structured format of the job ads and their nature, not all job ads contain job specifications such as salary in IT job ads where companies do not specify the

remuneration, where candidates should discuss it during the job interview.

The results of this study provide a strong foundation for future work. One area of future work is to capture the trends and the emerging competencies in the offshore sector and to focus on required soft skills as we will continue collecting the job ads. Another area is to apply this methodology to the aerospace industry and the automotive sector since they are growing sectors and require a large workforce. While, we are in the process of informing job market stakeholders about our results in order to draw more attention to the skill mismatch in the offshore sector.
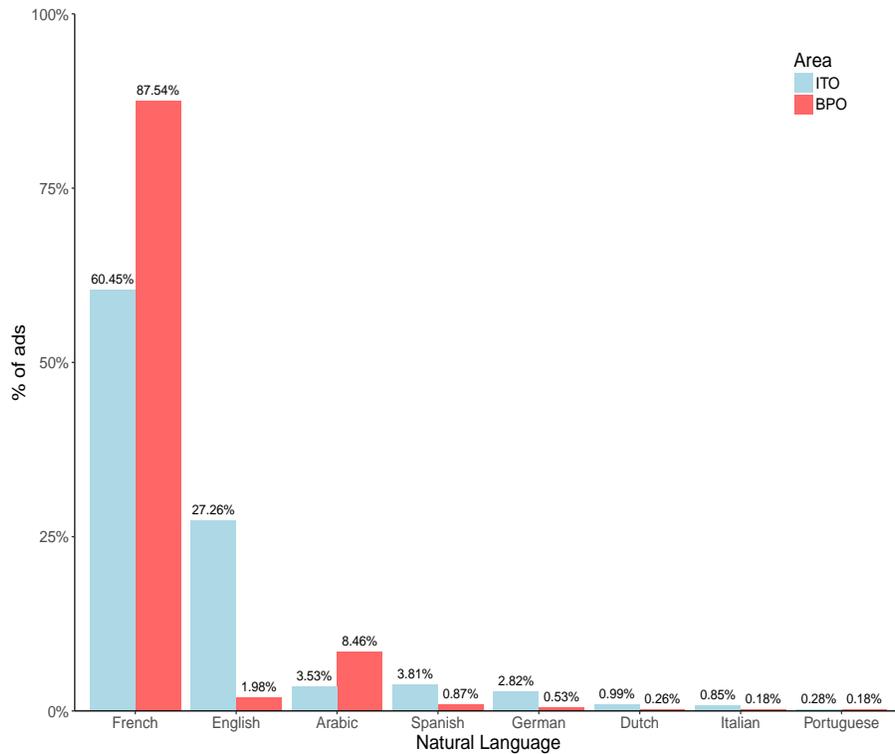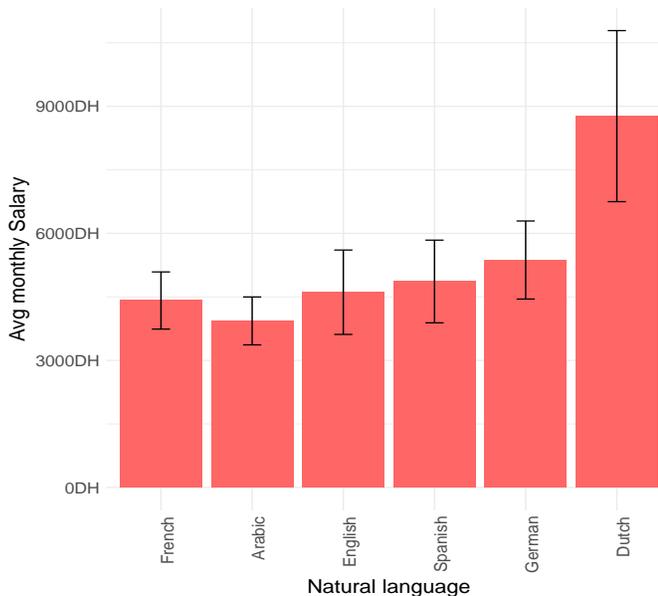
Fig. 6: Top natural languages

Fig. 7: Salary vs. Language in the BPO

VII. Conclusion

In this paper, we collect and analyze job ads from several websites, where ads have different formats: semi-structured and non-structured. We filter offshore job ads using text categorization approach, then we process these ads using several text mining approaches in order to draw an overview

of the offshore sector in Casablanca, the hub of the offshore in Morocco.

Our study reveals that advanced education is important in the ITO, conversely, the BPO area requires fewer years of education. In fact, Business Process Outsourcing gives job opportunities for youth with no qualifications nor experience, with a decent starting salary (double of the minimal wage). Also in our dependency analysis, we see closely the relation of the salary and languages where foreign languages aside from French and Arabic are paying more.

Our approach helps draw the big picture of the offshore job market needs. These findings can help universities to update their courses to meet the needs of the offshore sector which will increase the attractiveness of the country for foreign investment.

Our results indicate that French is the most needed language in offshore jobs, but that other foreign languages (English and Spanish) are also needed. In fact, we see that foreign languages are required in the Offshore which should urge the universities to focus more on the introduction of foreign languages in their curricula. In fact, foreign languages aside from French and Arabic are paying more in the Business Process Outsourcing. Besides, the top programming languages needed in Information Technology Outsourcing differ from programming languages taught in universities, where JavaScript and PHP are barely taught in most universities. They also differ from the international ranking of the top programming languages which confirms that the offshore needs should be examined closely by the Educators where they can use our results as an aid in
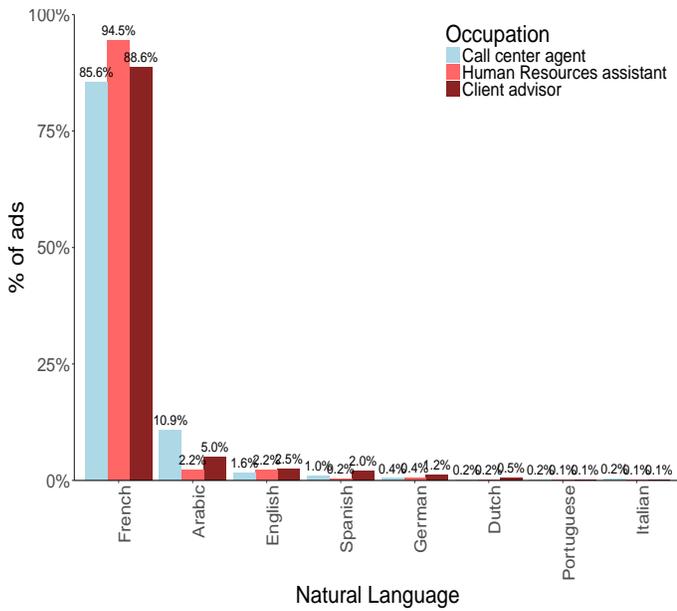
(a) BPO
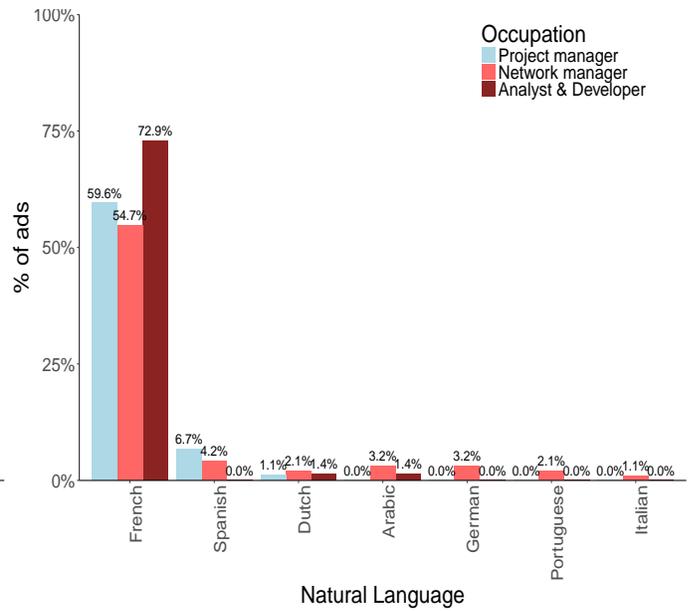
(b) ITO

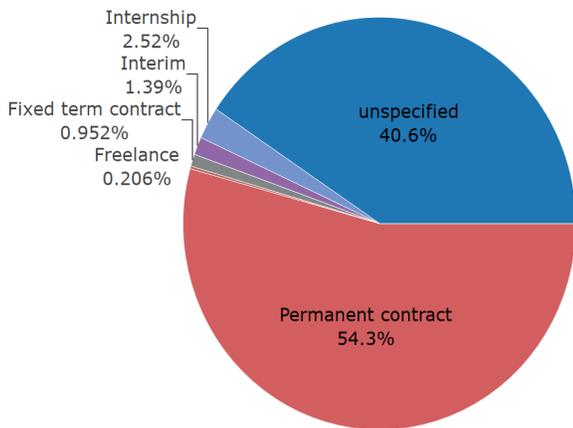Fig. 8: Natural languages vs. Occupation
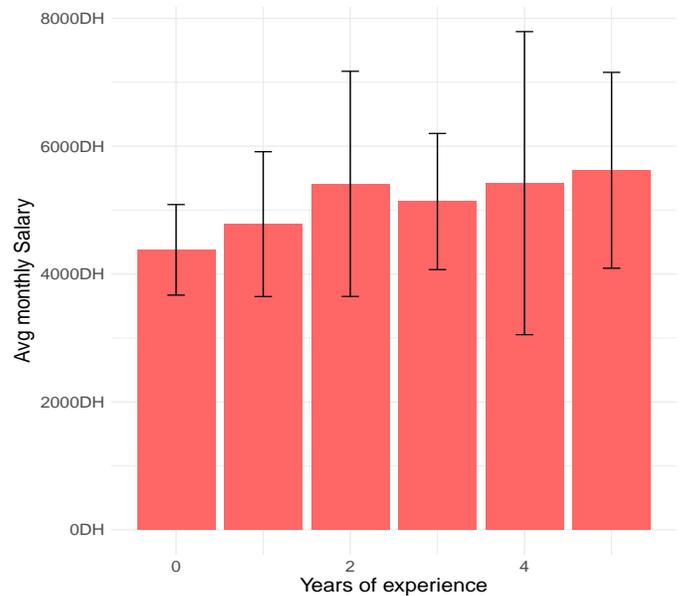

Fig. 9: Type of contract in the BPO
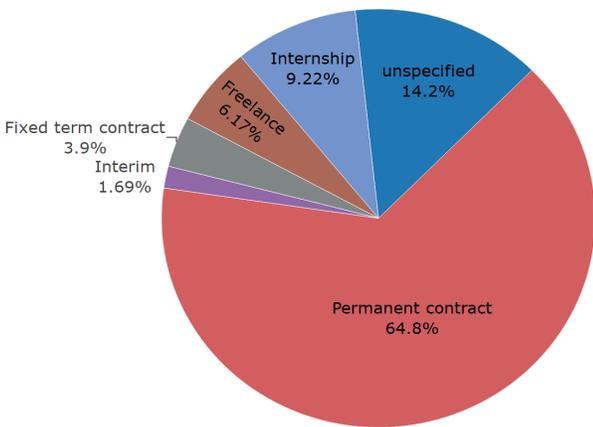

Fig. 11: Salary vs. Experience in the BPO

Fig. 10: Type of contract in the ITO

developing their curricula. In fact, these results could be used in educational curricula design, such as adding projects where the main focus will be on JavaScript and PHP, or by proposing additional courses for their undergraduates for programming languages and for foreign languages In fact on the basis of these results, our department is in the process of revisiting our curricula to improve the employability of our graduates.
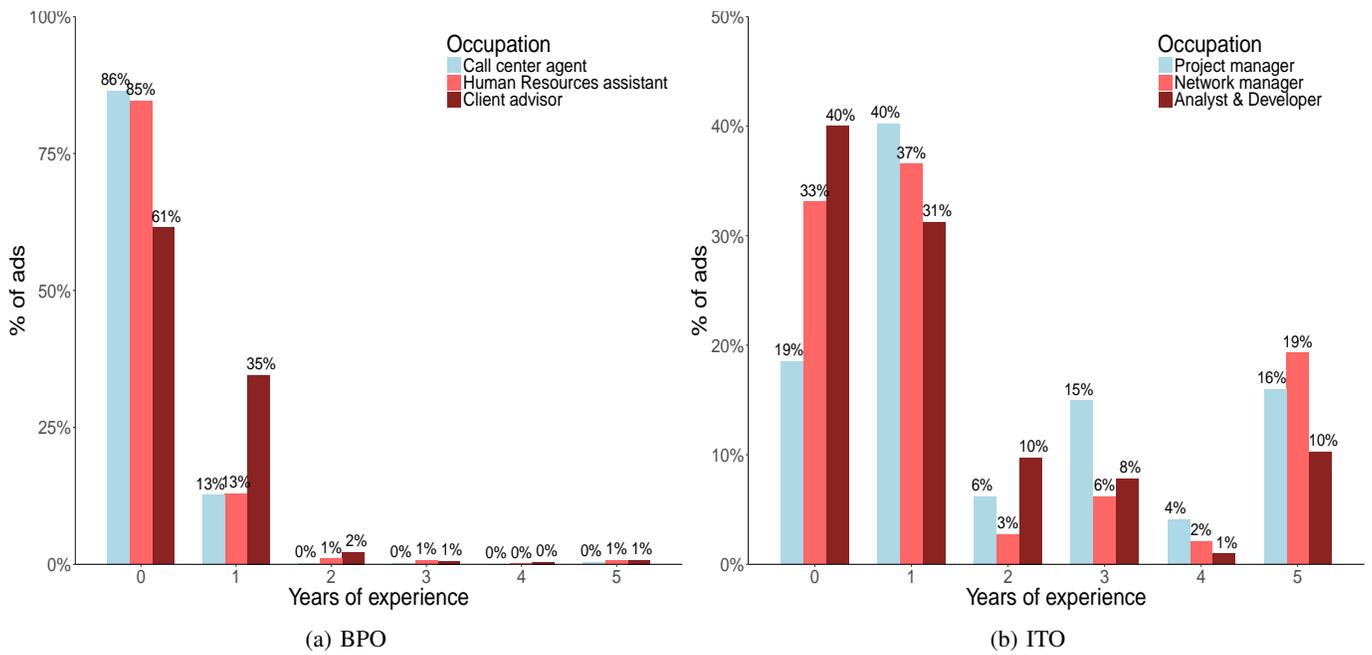
(a) BPO

(b) ITO

Fig. 12: Experience vs. Occupation

REFERENCES

[1] "Industrial Acceleration Plan 2014-2020 | Ministre de lIndustrie, du Commerce, de lInvestissement et de lEconomie Numrique." [Online]. Available: http://www.mcinet.gov.ma/~mcinetgov/en/content/industrial-acceleration-plan-2014-2020

[2] "OFFSHORING | Ministre de lIndustrie, du Commerce, de lInvestissement et de lEconomie Numrique." [Online]. Available: http://www.mcinet.gov.ma/~mcinetgov/fr/content/offshoring

[3] "Taux de chomage national selon le diplome."

[4] "Morocco: Selected Issues; IMF Country Report No. 16/36; November 30, 2015 - cr1636.pdf." [Online]. Available: https://www.imf.org/external/pubs/ft/scr/2016/cr1636.pdf

[5] H. Huang, L. Kvasny, K. D. Joshi, E. M. Trauth, and J. Mahar, "Synthesizing IT Job Skills Identified in Academic Studies, Practitioner Publications and Job Ads," in Proceedings of the Special Interest Group on Management Information System's 47th Annual Conference on Computer Personnel Research, ser. SIGMIS CPR '09. New York, NY, USA: ACM, 2009, pp. 121–128. [Online]. Available: http://doi.acm.org/10.1145/1542130.1542154

[6] C. Litecky, A. J. Igou, and A. Aken, "Skills in the Management Oriented IS and Enterprise System Job Markets," in Proceedings of the 50th Annual Conference on Computers and People Research, ser. SIGMIS-CPR '12. New York, NY, USA: ACM, 2012, pp. 35–44. [Online]. Available: http://doi.acm.org/10.1145/2214091.2214104

[7] T. Janotk, "Current labor market trends in the area of information technology in Slovakia," in 2015 13th International Conference on Emerging eLearning Technologies and Applications (ICETA), Nov. 2015, pp. 1–8.

[8] E. Malherbe and M. A. Aufaure, "Bridge the terminology gap between recruiters and candidates: A multilingual skills base built from social media and linked data," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Aug. 2016, pp. 583–590.

[9] R. Wilkinson, "Effective Retrieval of Structured Documents," in Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 311–317. [Online]. Available: http://dl.acm.org/citation.cfm?id=188490.188591

[10] G. S. Manku, A. Jain, and A. Das Sarma, "Detecting Near-duplicates for Web Crawling," in Proceedings of the 16th International Conference on World Wide Web, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 141–150. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242592

[11] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Comput. Surv., vol. 34, no. 1, pp. 1–47, Mar. 2002. [Online]. Available: http://doi.acm.org/10.1145/505282.505283

[12] I. Karakatsanis, W. AlKhader, F. MacCrory, A. Alibasic, M. A. Omar, Z. Aung, and W. L. Woon, "Data mining approach to monitoring the requirements of the job market: A case study," Information Systems, vol. 65, pp. 1–6, Apr. 2017. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0306437916304X

[13] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent Semantic Indexing: A Probabilistic Analysis," in Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, ser. PODS '98. New York, NY, USA: ACM, 1998, pp. 159–168. [Online]. Available: http://doi.acm.org/10.1145/275487.275505

[14] "REM/REC." [Online]. Available: http://www.dfp.gov.ma/component/content/article/307.html

[15] w. h. h. c. a. p. maroc, "Principaux enseignements sur la qualit de lemploi en 2016." [Online]. Available: http://www.hcp.ma/Principaux-enseignements-sur-la-qualite-de-l-emploi-en-2016_a1879.html

[16] F. Pashaee, "The 2015 Top Ten Programming Languages - IEEE Spectrum," Nov. 2015. [Online]. Available: https://www.linkedin.com/pulse/2015-top-ten-programming-languages-ieee-spectrum-farzin-pashaee

[17] S. Cass, "The 2017 Top Programming Languages," Jul. 2017. [Online]. Available: https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages

[18] "brochure offshoring 04-2017ok copie - brochure_offshoring-12052017.pdf." [Online]. Available: http://www.invest.gov.ma/upload/_ftp/documents/brochure_offshoring-12052017.pdf