2017 21st International Conference on Control Systems and Computer Science

How Adequate is your CV? Analyzing French CVs with *ReaderBench*

Gabriel Gutu, Mihai Dascalu, Stefan Trausan-Matu
Computer Science Department
University Politehnica of Bucharest
Bucharest, Romania
{gabriel.gutu, mihai.dascalu, stefan.trausan}@cs.pub.ro

Olivier Lepoivre
Domaine Digital
Groupe Randstad France
Saint Denis, France
olivier.lepoivre@randstad.fr

***Abstract*—This study is aimed at presenting a new *ReaderBench*-based tool built to support candidates in increasing the quality of their CV for a job opening. Both the visual quality and the textual content are considered while also providing an overview and corresponding feedback for the entire CV. The presented CV analysis tool uses advanced Natural Language Processing techniques to interpret and understand the content from written texts, while also considering their visual traits. The study was performed on a collection of more than 50 CVs that were manually annotated as positive or negative in terms of their visual and content-oriented aspects. A statistical analysis based on more than 400 textual indices was performed on the training corpora in order to extract the traits that define a good commercial CV. The results enabled us to build an online tool accessible on our website that provides recommendations for CVs written in French language.**

Keywords—CV analysis; text cohesion; semantic relatedness; textual complexity; Natural Language Processing

I. Introduction

Evaluating the quality of a *Curriculum Vitae* (CV) is a difficult and time consuming task when applying for a job position. Moreover, the procedure for choosing the most suitable CV for a position is subject to multiple criteria that can be used to differentiate among candidates. Unfortunately, it seems there is no gold standard or due to the wide range of different traits that Human Resource experts are interested in. By annotating a collection of more than 50 CVs with regards to their visual aspects, content, and overview classification, we identified several characteristics of interest. On one hand, a visually attractive CV may impress at a first glance, while a thorough examination may expose low-quality textual content. On the other hand, a qualitative CV in terms of textual content may be overshadowed by a messy design. Thus, we decided to build an automated system meant to evaluate CVs from both visual aspect and content, to support recruitment agencies in their manual endeavor of assessing commercial CVs.

The online tool that we built using our *ReaderBench* framework [1-3] is meant also to support individuals in creating better, more adequate CVs. Our tool examines their CVs and provides general statistics, warnings and suggestions meant to guide them in enhancing the quality of their CV, while referring to visual aspects and textual content. The online tool was built upon the results of two statistical analyses and the generated recommendations consider the most predictive indices for two criteria: visual aspects and the actual content of the CV.

In terms of structure, this study continues with a description of existent systems and their ability to process CVs, extract corresponding information and provide recommendations. The state of the art section is followed by a presentation of the *ReaderBench* framework which was used to perform the in-depth Natural Language Processing (NLP) analyses. This section also describes the components integrated within the streamlined process of evaluating CVs. The collection of documents used in this study is presented in section IV, together with details regarding the undergone refinements. Afterwards, the online tool is presented, followed by results and conclusions.

II. Existing Systems

There are numerous available systems (see Table I) that enable users to perform shallow automated CV analyses. Table I presents an overview of the tested services, their benefits, as well as their limitations. All the presented tools provide only basic analyses regarding an uploaded CV. Some applications display links to additional services that enable the online creation of your CVs after registration, while others direct the user towards a consultant.

Since none of the previous tools exhibited characteristics of interest for the envisioned in-depth analyses, we decided to create our own tool integrated with the already available *ReaderBench* framework [1-3], which will be presented in the next section.

III. The ReaderBench Framework

ReaderBench is an open-source framework designed to extract valuable data and information from texts using advanced NLP techniques including a wide range of textual complexity indices, semantic relatedness based on semantic models, as well as sentiments or valences from textual units. *ReaderBench* uses a NLP processing pipeline, which consists of a series of processes meant to clean undesired words and transform them for better contextualization. In other words, raw texts extracted from CVs require lexical adjustments in order to enhance follow-up semantical analyses that in return provide more adequate results [4].

2379-0482/17 $31.00 © 2017 IEEE
DOI 10.1109/CSCS.2017.85

TABLE I. DESCRIPTION OF EXISTING SYSTEMS

Name and URL	Description
CV Word Checker http://www.cvwordchecker.com	The tool provides an overview of the textual content of a CV by identifying good words, bad words, bad phrases, pronouns and passive voice using predefined lists. The tool suggests the usage of at least 10 good words, while words belonging to the other categories should be removed. The input field allows only textual input which may bring errors in terms of textual structure like delimitation of paragraphs and sentences, particularly when using a CV with a multi-column design.
Resume-Check https://www.resume-check.com	The online application identifies 20 of the most common problems within a CV (e.g., resume length, references, experience, education, spacing, capitalization, filler words, spelling errors) and provides a brief overview of them. A "Fix this" button is displayed for each identified problem that redirects users to a CV builder tool or to a dedicated page where they can contact experts. The tool is meant as a promotional service for commercial services.
RezScore http://rezscore.com	Users can score a CV and access a detailed analysis showing scores like brevity, impact and depth. Additionally, a list of matched jobs and a map of skills are displayed. The users may also request a human review of the CV.
Jobscan https://www.jobscan.co	The online scanner allows users to check whether a CV matches a specified job opening. The tool provides a score for the position and extracts a list of hard skills, soft skills and keywords of the uploaded CV.
Resunate https://www.resunate.com	The tool works similar to Jobscan – the users can enter a job description and a match score between the CV and the desired position is computed.
The *Enhancv* https://enhancv.com	The online application allows users to create a CV using templates with modern, attractive designs and shows suggestions while doing it. The system does not perform an a posteriory assessment, but displays recommendations for already-built CVs.
Hackmyresume https://www.npmjs.com/ package/hackmyresume	The tool generates a CV in many formats by using specific JSON files, but it is meant only for programmers or to be used as a library for third party applications.

We continue now with the description of the *ReaderBench* components used to create our CV analysis tool. First, we introduce the semantic models that express text relatedness. Second, we describe the lists of words used to determine sentiment valences, as well as specific categories of words. Third, we detail visual and content-oriented indices, followed by the presentation of third party libraries.

A. Semantic Models

Regarding semantics, *ReaderBench* integrates WordNet as ontology [5], and both Latent Semantic Analysis (LSA) [6] and Latent Dirichlet Allocation (LDA) [7] as semantic models. *WordNet* is a lexical database for English that groups words into *synsets*, which are sets of synonyms [5]. There are several implementations of WordNet available in multiple languages out of which WOLF (Wordnet Libre du Français, Free French Wordnet) (http://alpage.inria.fr/~sagot/wolf-en.html), the French version of the lexicalized database, was used for this study [8]. The internal representation of relations established between words allow the usage of specific algorithms to compute similarity scores between words. Numerous semantic distances based on WordNet have been developed, out of which *ReaderBench* integrates the following three distances that showed the best results in previous studies [2]: Wu-Palmer's, Leacock & Chodorow's and the path length distance [9].

LSA is a NLP technique used to determine relations of semantic similarity between units of texts based on a term-document matrix which stores the number of co-occurrences of words in the corresponding documents [6]. LSA uses a bag-of-words approach, which means that words are treated as a set, disregarding their order or relations among them within a document. The proximity in terms of relevance between two concepts is determined through cosine similarity [6]. The vector space is obtained after applying a singular value decomposition (SVD) on the normalized term-document matrix. For the experiment presented here, LSA was trained with the French *Le Monde* corpus (http://lsa.colorado.edu/spaces.html) and we used log-entropy as normalization of the term-document matrix.

LDA is a semantic model that tries to surpass some problems of LSA. It actuates a process of generation of topics using a probabilistic procedure [7]. LDA also disregards word order and the similarity between two concepts is computed as the inverse of the Jensen-Shannon dissimilarity [4] among their topic distributions. For this experiment, LDA was also trained on the French *Le Monde* corpus.

B. Dedicated Word Lists

ReaderBench integrates several word lists, which are usually of two types: 1) lists of word categories that express similar meanings in terms of lexical types, sociological or psychological aspects, and 2) lists of valences or sentiments having a score set for each word within a specific interval. *ReaderBench* integrates two word lists for French: *Affective norms for French words* (FAN) [10] and *Linguistic Inquiry and Word Count* (LIWC) [11].

Affective Norms for French Words is the French version of *Affective Norms for English Words* (ANEW) [12], a list of words developed for scientific research at education institutions. ANEW contains 1,034 words annotated with scores for valence, arousal and dominance. The scores of valence gathered for each word allowed us to classify them into positive, negative and neutral. Since valence scores varied between -3.8 and 3.8, words in the [-1; 1] range were considered neutral for not inducing a strong sentiment; positive and negative words were set considering the positive, respectively negative remaining ranges.

Linguistic Inquiry and Word Count (http://liwc.wpengine.com) is a classification of more than 4,500 most common words or stems of a language. The words were classified by experts into 64 categories that cover linguistic (parts of speech), paralinguistic, psychological (cognition, social affective, etc.), and personal aspects. LIWC has been published in various languages of which the French version was used for this experiment [13]. Of particular interest were the following lists of categories: positive emotions, negative emotions, affective processes, anger, sadness, anxiety and aggression,

which outline the characteristics of an individual and their way of expression [14]. Specific categories may be followed by employers, where they are interested in the presence or absence of certain keywords. As an example, one employer may find it more acceptable to have more *Positive Emotions* words than *Negative Emotions*.

C. Visual and Textual Complexity Indices

While relating to the indices considered for subsequent analyses, we started by considering statistics regarding the structure of the text, i.e., the number of pages, paragraphs, sentences, words and content words. *Visual aspects* covered statistics like the number of images contained within the CV and the number of colors; both were normalized to the number of pages. Font statistics was another visual aspect considered; it included the number of font types, basic font types and font sizes used in the texts. Font sizes were relevant while relating to the number of different sizes, normalized by the number of pages. Statistics regarding the usage of Bold, Italic, and both Bold and Italic characters were performed, as well, in terms of visual aspects. Thus, the total number of corresponding characters was computed and normalized to the total number of characters.

Words' valences were determined (such as positive, negative or neutral words) using the *Valence* FAN scores. The number of words contained in categories of the LIWC list were calculated. Textual complexity indices available for French language were also computed and cover the following categories: surface, lexical, syntax, semantics, and discourse structure. The detailed presentation of the indices is available in previous work [2, 15].

D. Integrated Third Pary Libraries

Our *ReaderBench* framework relies on additional third-party libraries to accomplish different operations. The *Stanford Core NLP* library [16] provides valuable support in the NLP-related processing of the input texts. With regards to the training process of new semantic models, *Apache Mahout* [17] was used to perform the decomposition of the SVD matrix required for LSA, while *Mallet* [18] allowed us to infer new topics for LDA. Apache PDFBox (https://pdfbox.apache.org) supports the process of content extraction from pdfs, both for textual content and for visual statistics (e.g., types of fonts, the sizes of texts, colors or images). D3.js (https://d3js.org) is another incorporated library that embodies a series of graph models used to generate the concept map.

IV. The Collection of CVs

The collection that was used for our experiment consisted of 52 CVs written in French language, which were selected from a wider sample of more than 100 CVs. The CVs were provided in PDF or Microsoft Word .doc format, but they were all converted to PDF in order to simplify the process of extraction of text by using a single library. The main criterion considered for the selection of CVs was the necessity of having editable text. Scanned documents were disregarded because of the necessity of an Optical character recognition (OCR) library, which would introduce additional noise due to the misinterpreting of texts.

The CVs were annotated as positive or negative in terms of their visual and textual content-oriented aspects (see Table II). The majority of CVs belonged to males (48). Regarding age, only 32 of the CVs contained either the year of birth, or the age of the applicant specified in clear (M = 41 yrs., SD = 11 yrs., min = 21 yrs., max = 59 yrs.). The current study does not differentiate among socio-demographical characteristics of the candidates because of the relative small number of CVs. Nevertheless, these characteristics might be considered for future researches that account for demographic groups particularities.

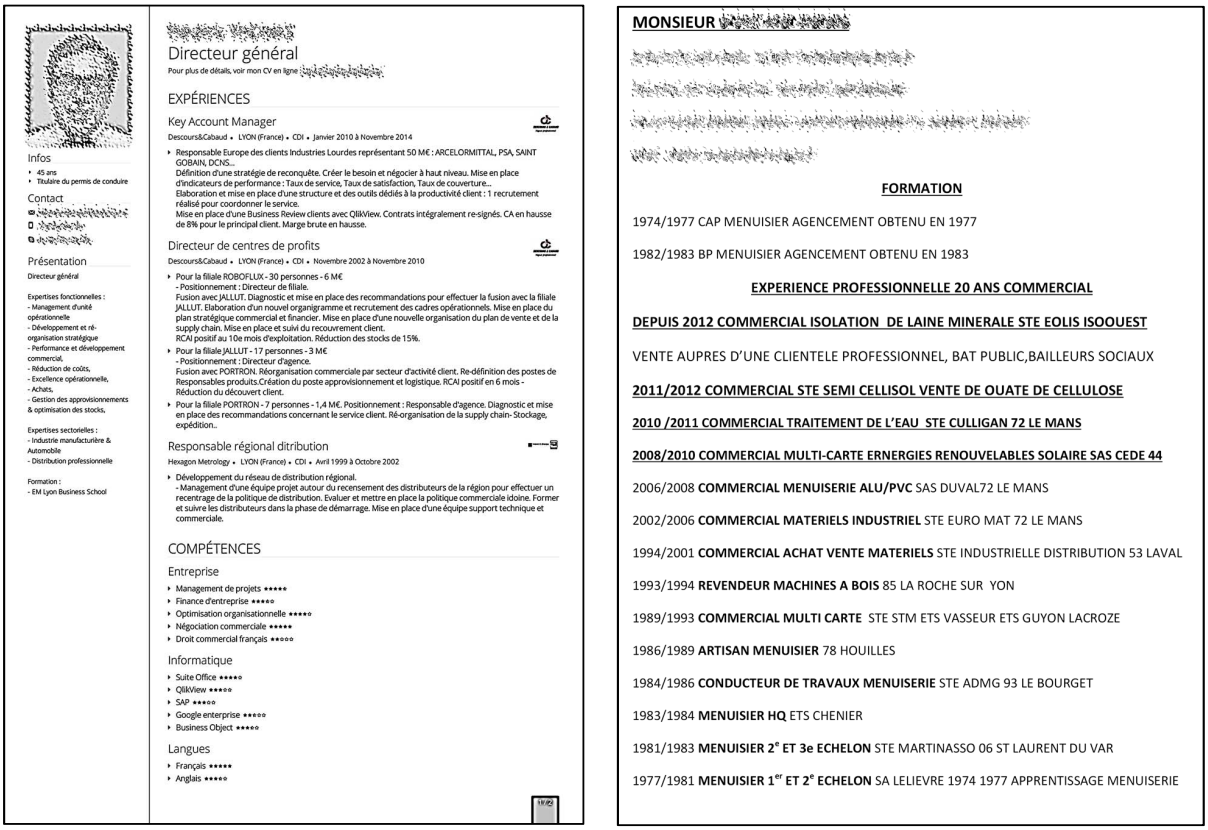
TABLE II. General corpus descriptives (N = 52).

Criterion	# Positive	# Negative
Visual aspects	31	21
Content-oriented	25	27

Figure 1.a shows an example of CV classified as positive in terms of both visual and textual aspects that exhibits the following traits: a good organization of the content, the presence of colors and logos of the companies, a picture of the candidate (which was blurred for this study), the presence of contact information, the delimitation of experience in a dedicated area that makes them easier to be read, as well as the existence of commercial figures. A CV marked as negative in terms of both visual and content is shown in Figure 1.b since it was too short and it lacked valuable commercial information.

Observations made on our collection showed that CVs that were classified as positive in terms of visual aspects were aerated, contained a profile photo, colors or logos of companies for whom the applicant worked for. Positive CVs were also easy to read, condensed and facilitated the conceptualization of the most important keywords. Positive CVs contained, also, well distributed content according to the information that the candidate wished to highlight, which may refer to the distribution of text in columns or well-delimited paragraphs. In contrast, aspects that were observed when classifying CVs as negative included: lack of readability, absence of spacing between paragraphs or sections, very few information or too much confusing information that produces false expectations, and small text size and bad structure (for example, interval dates that are displayed on the right instead of the left side of the CV).

In terms of textual classification, positive CVs contained details regarding: 1) the location of previous companies for which the candidate worked for or the description of responsibilities for previous positions, 2) the existence of commercial figures like company's turnover, and 3) the usage of words appearing in specific vocabularies related to the working domain. CVs classified as negative for textual content lacked details regarding the candidate's experience or interval dates when they worked for the company, absence of commercial figures, spelling mistakes or lack of readability.



(a) (b)

Fig. 1. Side-by-side comparison of CVs being classified as (a) positive versus (b) negative. Some details were blurred for anonymisation.

V. The Online Tool

Our tool allows users to receive recommendations regarding their CV written in French language and is available on *ReaderBench*'s framework website [19], on the *Demo* section at http://readerbench.com/demo/cv. As some CVs may contain special characters and could produce errors, we suggest using the Europass format for testing purpose, available at http://europass.cedefop.europa.eu/documents/curriculum-vitae/examples. The required input parameters are described in Table III, while the provided output results are presented in Table IV, with corresponding visual displays shown in Figure 2. The tool is meant for both regular users who wishes to improve the overall quality of their CV, as well as employers who can set their own list of keywords to be sought for.

TABLE III. Input Parameters of the Online Tool

Field	Description
Basic options	
CV file	A CV file in the PDF format that should contain selectable text. A maximum size of 25 MB is permitted for each file to be processed. A recommendation to minimize the file is provided if the size exceeds 5 MB, while a suggestion to reduce the quality of the images for files larger than 1 MB, but smaller than 5 MB, is shown.
Specific keywords	A list of keywords separated by commas (spaces between keywords are ignored) to be searched for within the CV.
Words to ignore	A list of words that need to be excluded from the CV analysis because they may induce noise. Words like the name of months or days may be included in this list.
Language	A select menu of available languages French and English languages are currently supported.
Advanced options	
LSA corpus	A select menu of pretrained *Latent Semantic Analysis* corpora for the chosen language. Two options are available for French: *Le Monde* (http://lsa.colorado.edu/spaces.html) and *Text Enfants* [20].
LDA corpus	A select menu of pretrained *Latent Dirichlet Allocation* corpora for the chosen language. The same corpora for LSA were used for LDA.
Part of speech tagging	Option to enable or disable the automatic identification of parts of speech. The usage of this setting provides better results as the tagger tries to identify what is each word's corresponding part of speech.
Compute dialogism	Option to enable or disable the automatic identification of voices, i.e. semantic chains that span throughout the discourse [21]. As CVs mostly do not contain linked or narrative texts in general, it is not that important to enable this option.
Semantic similarity threshold for concept map	A number field that accepts values between 0 and 1 as minimum threshold for semantic similarity scores between concepts from the concept map. Each link between keywords extracted from the input text below this threshold will not be rendered in the concept map.

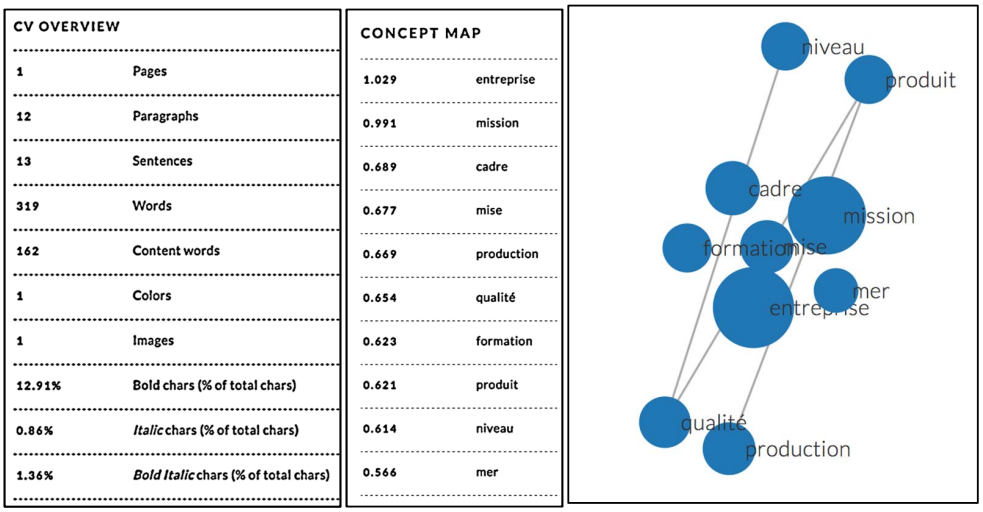


Fig. 2. An example of output results for a CV.

TABLE IV. OUTPUT RESULT SECTIONS OF THE ONLINE TOOL

Section	Description
CV Overview	Presents general statistics related to the visual aspects of the CV's structure such as the number of pages, paragraphs, sentences, words and content words (i.e. words that are nouns, verbs, adjectives or adverbs used for conceptualizing the input text), the number of colors and images, as well as the percentage of bold, italic and both bold and italic characters. This section presents recommendations based on visual or textual indices with abnormal values. For example, if the CV contains more than two pages, the user receives a recommendation to stick only to 1 page.
Social Network Accounts	Presents a list of extracted social media network accounts, if present, or recommendations to include them, if absent. Currently, the tool recognizes LinkedIn and Viadeo profiles.
Keywords Relevance	Comprises of two subsections: *Keyword Overlap* and *Relevance*. The first subsection displays which specific keywords exist in the CV with the corresponding number of occurrences and the cohesion score with the entire textual contents of the CV. The second subsection shows the overall cohesion score between the list of keywords as a whole and the textual contets of the CV.
Concept Map	A concept map showing the most relevant 10 extracted keywords and corresponding semantic links is generated. The *Semantic similarity threshold* input field is considered here for pruning the generated graph. The size of each node is proportional to the concept's relevance score.
Positive Words	Displays a list of positive words extracted using the ANEW *Valence* scores.
Negative Words	Displays similar information as the above, but for the negative words determined using the ANEW *Valence* scores.
Linguistic Inquiry and Word Count	Displays the list of words found in each LIWC category; empty categories are ignored.
Textual Complexity	Shows indices related to textual complexity measures grouped by category.

VI. RESULTS

Our experiment was conducted using the French *Le Monde* corpora due to its adequacy and better language coverage for both LSA and LDA semantic models; the part of speech tagger and the dialogism options were both enabled. The semantic similarity threshold for concept map was set to 0.3. Regarding the uploaded files, for our experiment the average size of a CV was 0.19MB (SD = 0.13MB, min = 0.01MB, max = 0.58MB). All files larger than 0.5MB contained images.

Statistical analyses were performed to investigate both the visual and content characteristics of CVs in terms of the previously presented variable indices computed with the *ReaderBench* framework. A total of more than 400 indices were integrated. First, all indices were checked for normality and those that demonstrated non-normality were removed. Second, multicollinearity was assessed as pair-wise correlations ($r > .70$); if indices demonstrated multicollinearity, the index with the strongest effect in the model was retained for subsequent analyses.

Afterwards, two multivariate analyses of variance (MANOVA) [22] were conducted to examine the effect of each index in terms of the considered criterion. For all the variables that were considered in the subsequent analyses, Levene's test of equality of error variances was not significant ($p > .05$). With regards to *visuals* aspects, the indices from Table V present (upper part), in descending order of effect size, the visual and surface indices that were significantly different. Thus, based on the given CV collection, appealing CVs use more, but not too many, font types and avoid including barely readable text reflected in small font size.

Table V (lower part) presents, in descending order of effect size, the indices that were significantly different between positive and negative CVs for content-centered characteristics. First, CVs labeled as positive contain more adverbs and more words expressing positive emotions than negative CVs. Second, the average number of words per sentence labeled in LIWC as expressing achievements, leisure or friendship relations shows a more positive CV. Third, a higher standard deviation of unique

words per sentences and a higher entropy of words encountered in positive CV are indicative of a more diverse vocabulary. Forth, positive CVs are more cohesive by having a higher document cohesion flow score computed using the path length similarity and the maximum value criteria [23]. Fifth, the average number of syntactic dependencies per sentence, both determiner and multi-word expression, denotes a positive CV.

Two stepwise Discriminant Function Analyses (DFAs) were performed to predict the visual, respectively, the content characteristics of CVs. The first DFA retained one variables as significant predictor (*Simple font types*) and removed the remaining variables as non-significant predictors. The resulting DFA model was capable to significantly differentiate between CVs while relating to visual aspect, Wilks' $\lambda = .871$, $\chi^2(df = 1) = 6.808$, $p = .009$. The DFA correctly allocated 33 (18 + 15) of the 52 CVs from our collection, resulting in an accuracy of 63.5% (the chance level for this analysis is 50%). For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 27 (12 + 15) of the 52 CVs for an accuracy of 51.9%, denoting clear inconsistencies in the manual evaluation of the CVs. The measure of agreement between the CV's rating while considering its visual design and that assigned by the model produced a weighted Cohen's Kappa of .310, which demonstrates fair agreement.

The second stepwise DFA, centered on predicting the quality of a CV's content, retained one variables as significant predictor (*Adverbs LIWC*) and removed the remaining variables as non-significant predictors. The resulting DFA model was capable to significantly differentiate between CVs while addressing content, Wilks' $\lambda = 0.794$, $\chi^2(df = 1) = 11.413$, $p = .001$. The DFA correctly allocated 35 (20 + 15) of the 52 CVs from our collection, resulting in an accuracy of 67.3% (the chance level for this analysis is 50%). For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 35 (20 + 15) of the 52 CVs for an accuracy of 67.3%, thus surpassing the results of the visual analysis. The measure of agreement between the CV's rating in terms of content quality and that assigned by the model produced a weighted Cohen's Kappa of .352, demonstrating fair agreement.

TABLE V. TESTS OF BETWEEN- SUBJECTS EFFECTS FOR SIGNIFICANTLY DIFFERENT INDICES PREDICTIVE OF THE VISUAL ASPECT.

Index	*M (SD) positive*	*M (SD) negative*	*F*	*p*	Partial η^2
Visual aspects					
Simple font types	3.58 (1.34)	2.43 (1.72)	7.373	.009	.129
Minimum font size	8.51 (3.19)	6.38 (3.65)	4.711	.035	.086
Structural aspects					
Number of adverbs (LIWC)	11.22 (4.73)	6.92 (3.79)	12.966	.001	.206
Number of words labeled with positive emotions (LIWC)	9.11 (3.62)	6.68 (2.85)	7.155	.010	.125
Average words per sentence labeled with achievement (LIWC)	0.37 (0.17)	0.26 (0.13)	7.003	.011	.123
Sentece standard deviation in terms of unique words	5.00 (2.35)	3.52 (1.60)	6.909	.011	.121
Word entropy	4.74 (0.28)	4.57 (0.27)	5.447	.024	.098
Document flow average cohesion using path length similarity and maximum value criteria	0.44 (0.05)	0.40 (0.06)	5.014	.030	.091
Average number of syntactic dependencies per sentence (multi-word expression)	1.36 (0.57)	1.04 (0.45)	4.816	.033	.088
Average words per sentence labeled with leisure (LIWC)	0.23 (0.10)	0.17 (0.08)	4.741	.034	.087
Average number of syntactic dependencies per sentence (determiner)	1.80 (1.04)	1.28 (0.64)	4.688	.035	.086
Number of words labeled with inclusion (LIWC)	1.30 (0.87)	0.76 (0.93)	4.645	.036	.085
Average words per sentence labeled with friends (LIWC)	0.12 (0.07)	0.08 (0.06)	4.613	.037	.084

VII. CONCLUSIONS

Our online tool allows users to test the quality of their CVs and provides them with general statistics and recommendations. To our knowing, there is no other tool freely available that performs similar in-depth CV analyses, especially in French language. Moreover, our automated analysis of CVs provides benefits for both regular users who need to understand what should they improve to increase their chances of being accepted for a job position, as well as employers requiring a bulk analysis of multiple CVs.

The performed discriminant analyses showed that simple font types are significant for determining the visual aspects of a CV. Regarding the contents of the CV, the number of adverbs extracted from LIWC was the most predictive index. Overall, the visual classification had an accuracy of 63.5%, while the textual content classification exhibited an accuracy of 67.3%.

The results provided by the two statistical analyses will be further used in researches to determine the quality of a CV by combining the two criteria (visual aspects and textual content) into a global score.

Our envisioned enhancements cover the extension of our French CV collection and the definition of formulas to predict the quality of a CV. Characteristics like age, location or gender will be considered into future researches to determine particularities of demographic groups. Another potential research direction targets the creation of a collection of representative English CVs, thus extending this study for English language as well.

ACKNOWLEDGMENT

This work was funded by the 644187 EC H2020 RAGE (Realising and Applied Gaming Eco-System) project - http://www.rageproject.eu/project and by University Politehnica of Bucharest through the "Excellence Research Grants" Program UPB–GEX 12/26.09.2016. We would also like to thank Sandrine Rey Peiro for her support in creating the CV collection used within this study.

REFERENCES

[1] M. Dascalu, P. Dessus, S. Trausan-Matu, M. Bianco, and A. Nardy, "ReaderBench, an environment for analyzing text complexity and reading strategies," in *16th Int. Conf. on Artificial Intelligence in Education (AIED 2013)*, Memphis, USA, 2013, pp. 379–388.

[2] M. Dascalu, *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence* vol. 534. Cham, Switzerland: Springer, 2014.

[3] M. Dascalu, L. L. Stavarache, P. Dessus, S. Trausan-Matu, D. S. McNamara, and M. Bianco, "ReaderBench: An Integrated Cohesion-Centered Framework," in *10th European Conf. on Technology Enhanced Learning*, Toledo, Spain, 2015, pp. 505–508.

[4] C. D. Manning and H. Schütze, *Foundations of statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

[5] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM,* vol. 38, pp. 39–41, 1995.

[6] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to Latent Semantic Analysis," *Discourse Processes,* vol. 25, pp. 259–284, 1998.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993–1022, 2003.

[8] B. Sagot, "WordNet Libre du Francais (WOLF)," ed. Paris: INRIA, 2008.

[9] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics,* vol. 32, pp. 13–47, 2006.

[10] C. Monnier and A. Syssau, "Affective norms for french words (FAN)," *Behavior Research Methods,* vol. 46, pp. 1128-1137, 2014.

[11] J. W. Pennebaker, R. J. Booth, and M. E. Francis, "Linguistic inquiry and word count: LIWC [Computer software]," in *Austin, TX: liwc. net*, ed. Austin, TX: University of Texas, 2007.

[12] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings," The Center for Research in Psychophysiology, University of Florida, Gainesville, FL1999.

[13] A. Piolat, R. J. Booth, C. K. Chung, M. Davids, and J. W. Pennebaker, "La version française du LIWC : modalités de construction et exemples d'application," *Psychologie française,* vol. 56, pp. 145–159, 2011.

[14] Y. R. Tausczik, & Pennebaker, J. W., "The psychological meaning of words: LIWC and computerized text analysis methods," *Journal of language and social psychology,* vol. 29, pp. 24-54, 2010.

[15] M. Dascalu, P. Dessus, M. Bianco, S. Trausan-Matu, and A. Nardy, "Mining texts, learner productions and strategies with ReaderBench," in *Educational Data Mining: Applications and Trends*, A. Peña-Ayala, Ed., ed Cham, Switzerland: Springer, 2014, pp. 345–377.

[16] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, MA, 2014, pp. 55–60.

[17] S. Owen, R. Anil, T. Dunning, and E. Friedman, *Mahout in Action*. Greenwich, CT: Manning Publications Co., 2011.

[18] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," ed. Amherst, MA: University of Massachusetts Amherst, 2002.

[19] G. Gutu, M. Dascalu, S. Trausan-Matu, and P. Dessus, "ReaderBench goes Online: A Comprehension-Centered Framework for Educational Purposes," in *Romanian Conference on Human-Computer Interaction (RoCHI 2016)*, Iasi, Romania, 2016, pp. 95–102.

[20] G. Denhière, B. Lemaire, C. Bellissens, and S. Jhean-Larose, "A semantic space for modeling children's semantic memory," in *Handbook of Latent Semantic Analysis* Mahwah: Erlbaum, 2007, pp. 143–165.

[21] M. M. Bakhtin, *The dialogic imagination: Four essays*. Austin and London: The University of Texas Press, 1981.

[22] G. D. Garson, *Multivariate GLM, MANOVA, and MANCOVA*. Asheboro, NC: Statistical Associates Publishing, 2015.

[23] S. A. Crossley, M. Dascalu, S. Trausan-Matu, L. Allen, and D. S. McNamara, "Document Cohesion Flow: Striving towards Coherence," in *38th Annual Meeting of the Cognitive Science Society*, Philadelphia, PA, 2016, pp. 764–769.