Classification of Web Job Advertisements: A Case Study

Flora Amato[1], Roberto Boselli[23], Mirko Cesarini[23], Fabio Mercorio[23], Mario Mezzanzanica[23], Vincenzo Moscato[1], Fabio Persia[1], and Antonio Picariello[1]

[1] Dept. of Computer Science and Systems, University of Naples Federico II, Italy
[2] Dept. of Statistics and Quantitative Methods, Univerisity of Milan-Bicocca, Italy
[3] CRISP Research Centre, Univerisity of Milan-Bicocca, Italy

Discussion Paper

Abstract. This work is concerned with classifying Web job advertisements against a standard classification system of occupations, by applying and comparing different text classification techniques. As a first step, we evaluated the classification algorithms using a *hit/not-hit* approach, that is either the prediction is correct or not compared to a gold classification provided by domain experts. Then, we built a distance function on top of the *affinity* relationship between occupations provided by the classification system. Both the classification scores we computed and the affinity distance employed have allowed a more finely grained evaluation of the classified outcomes, providing to authors useful insights towards the improvement of the classification process.

Keywords: Text Classification, Knowledge Management, Machine Learning

1 Introduction and Contribution

One of the drive of change in the labour market consists in the increasing use of the Web - by both employers and job seekers - for advertising demand and supply, and this enables new ways for recruitment (e.g., *social recruitment* and *e-recruitment*). On the other hand, organizations and governments have defined a number of national and international skills/occupations classifiers, that aim to classify and standardize labour *occupations*, *skills* and *competences* over several countries and languages. In such a scenario, the research activity we describe here goes towards two directions.

First, the **reconciliation** of Web job offers over an international standard occupation classifier - rather than a proprietary one - gives to labour market analysts and policy makers a *lingua franca* useful for studying and understanding the labour market dynamics over several countries, overcoming the linguistic boundaries. To this end, we applied several (and different) text classification techniques to classify a real dataset of Web job offers onto to the ISTAT classifier, namely CP2011 [1]. We evaluated the effectiveness of each approach by comparing the results against a gold classification.

Second, the **evaluation** of the classification algorithms effectiveness by looking at the *similarities* between the occupations classified. To this end, we exploited the *affinity* relation that the ISTAT occupation classifier defines for some occupations. The rationale of this relationship is to express that a professional profile (e.g. *programmer*) is akin to another one, such as *software developer*. As a consequence, this affinity relationship would allow us to employ a metric distance between a pair of occupations, thus the distance between two occupation becomes the shortest path between them if exists, otherwise these occupations are not related at all. Thanks to this distance measure we can perform a fine-grained comparison between the classification algorithms effectiveness as we detail in Sec.4.1, and these information would help us in improving and planning the further research steps.

In Sec. 2 we provide a literature review while in Sec. 3 we briefly introduce the dataset and techniques we used. Then, the experimental results are provided in Sec. 4, while the concluding remarks and future works are outlined in Sec. 5.

2 Related Work

In the last decade, the huge availability of text information has led to a strong interest in automatic text extraction and processing technology for extracting task-relevant information [1]. In such a scenario, Information Extraction (IE) techniques automatically extract structured information from unstructured and/or semi-structured documents, exploiting different kinds of text analysis, mostly related to Natural Language Processing (NLP) methodologies and to cross-disciplinary perspectives including Statistical and Computational Linguistics, see, e.g., [2]. Moreover, IE techniques may be associated with text mining [3] and semantic technologies activities in order to detect relevant concepts from text data, for information indexing, classification and retrieval[4, 5] aims, as well as long term preservation issues [6, 7]. Focusing on the the labour market domain, the extraction of meaningful information from unstructured texts has been mainly devoted to support the e-recruitment process attempting to automate the *resume* management for matching candidate profiles with job descriptions. To give a few examples, the work [8] proposes a system aiming to screen candidate profiles for jobs, by extracting various pieces of information from the unstructured resumes through the use of probabilistic information extraction techniques as Conditional Random Fields. Similarly, the "mandatory communications" system realised by the Italian Ministry of Labour and Welfare has been used for studying the Italian Labour Market dynamics performing both data quality [9] and knowledge discovery activities [10]. Differently, in [11] a cascaded Information Extraction model based on SVM for mining resumes is used whilst [12] uses

[1] The Italian standard classifier for Occupations is based on the logic of the ISCO (International Standard Classification of Occupations), that has been built to be cross-linkable with the latter.

structural relevance models to identify job descriptions and resumes vocabulary. In [13] the authors develop a job recommender system to dynamically update the job applicant profiles by analysing their historical information and behaviours. Poch et. al [14] aim to match the appropriate candidates to a job offer, and for this task they use supervised classifiers to suggest a ranked list of job offers to job seekers. Among the methods, they use Latent Dirichlet Allocation (LDA here on) to cluster similar job offers. Job clustering is accomplished by authors in terms of candidate classification for one or more classes of jobs according to a model learnt from the matching information and not to a well-established classifier.

Although all these approaches are relevant and effective, they differ from our purposes as we aim at processing *job-offers* rather than resumes, and this requires to deal with shorter texts that present a high degree of heterogeneity. Here, the joint use of *different* techniques would be beneficial in evaluating the effectiveness of these approaches in our application domain.

3 Applied Techniques

Here we shortly describe the main techniques we used to classify a job offers, as we have done in our previous work [15].

We considered about $40,000$ job vacancies scraped from 12 Web sources and a subset of 412 job offers selected for being a representative sample. Then, each job vacancy has been manually labelled by domain experts at CRISP Research Centre using the qualification codes outlined in the CP2011 classifier, by looking at both offer titles and full descriptions to assign labels. This sample dataset will be used as a *gold benchmark* to evaluate the classification technique outcomes.

Furthermore, a common text preprocessing pipeline was used before applying any of the approaches applied that includes: tokenization, lower case reduction, html special characters substitution, stop words removal, misleading words elimination, and numbers elimination. The word stemming was performed using the Italian stemmer provided by the NLTK framework version 3.0a3 [16]. Finally, in this experimental phase only job titles have been considered as input of the classification process.

Explicit Rule based approach. A commercial tool has been used to classify texts using a rule based approach. The user has defined rules for classifying texts focusing on taxonomies modelling the terminology of a certain domain. The rules building process is driven by a twofold goal: (i) to identify the relevant terms used in vacancy Web ads, and (ii) to relate them to the occupation codes used in the CP2011 classifier.

Machine Learning approach. Each job offer title is turned into a vector of word occurrences according to a bag of words approach. Then, the whole set of job offer titles was turned into a matrix of 412×542 elements whereas each line represents a job offer title, and the columns represents the features (i.e., the word count of the different stemmed words).

Then, two machine learning classifiers were used to perform the text classification purposes: the LinearSVC (an implementation of Support Vector Machine Classification using a linear kernel) and the Perceptron classifier, both built using the Scikit-learn framework [17]. A grid search of the classifier parameters maximizing the classification accuracy was performed on both classifiers.

LDA based approach. The technique only leverages titles of Web job offers as document set and is based on two different steps: *feature extraction* and *classification.* The goal of the feature extraction task is to derive for each ISTAT job category a particular data structure, named *Weighted Word Pairs* (WWP) and containing the most relevant pairs of *lexical items* (i.e. single word, a part of a word or a chain of words) together with the probabilistic dependencies characterizing titles of job vacancies [18–20].

On the other hand, the classification procedure is based on the matching between terms derived from a job vacancy title and the set of pairs related to different ISTAT categories, according to a distance-based approach.

In a first stage, all possible pairs of relevant terms (without repetitions) from texts related to the titles of job vacancies are extracted exploiting a classical *Natural Language Processing* (NLP) pipeline. Such pairs are then compared with the WWPs of all job categories using as similarity the *Levenshtein* metric.

For each category, the number of *hits* (positive matchings) is determined and each single hit is weighted by the dependency probability of the related pairs in the WWP structure. The category having the highest score is finally selected as *winner* and chosen for the classification.

4 Experimentation

The 412 sampled job offers were classified against 62 qualification codes, unfortunately 42 of them occurs less than 5 time in the sample. Table 1(a) shows the precision, recall, f-score, and accuracy for all the techniques, as reported in [15]. We can comment the following.

Table 1. Text classification techniques scores. Within parenthesis we show the same values computed by excluding the vacancies corresponding to infrequent occupation codes

		LDA	**Rules**	**Linear SVC**	**Percept.**
(a)	Accuracy	.5 (.51)	.444 (.469)	**.556 (.633)**	.483 (.543)
	Avg. Precision	**.507 (.587)**	.353 (.432)	.259 (.576)	.284 (.503)
	Avg. Recall	**.502** (.538)	.354 (.432)	.272 (**.576**)	.254 (.503)
	Avg. FScore	**.471** (.532)	.328 (.462)	.26 (**.607**)	.263 (.564)
	Precision Std.Dev.	.41 (.305)	.378 (.328)	**.34 (.239)**	.364 (.243)
	Recall Std. Dev.	.391 (.274)	.348 (.224)	.358 (.224)	**.332 (.195)**
	FScore Std. Dev.	.366 (.253)	**.326** (.257)	.337 (.216)	.336 (**.201**)
(b)	Avg. shortest-path length	1.50(1.29)	2.52(2.07)	1.6(0.94)	1.9(1.49)

- The machine learning approach - based on the use of a *Linear SVC* classifier - reaches good performances when an ample number of training samples are available, while precision and recall values drop considerably when considering all the categories. This is reflected into the LinearSVC data showed in Tab. 1: a gap exists between precision computed (1) considering all the occupation codes, and (2) excluding the infrequent ones, the same is for recall;
- The *rule-based* approach has a similar gap (although smaller) on precision values between (1) and (2), the same for recall. The rationale is that domain experts paid more attention to some specific occupation codes (by the way, the more frequent ones) during the rules developing process. Furthermore, the cases matching no rule generate a certain number of miss-classifications which also negatively affect the results;
- The statistical approach - based on *LDA* - has satisfying results respect to all the categories (we have an overall precision of 50%), and - as expected - such results are less or not affected by the infrequent occupation codes at all. Here, the incorrect classifications are related to the presence of categories having a similar lexicon and a different number of training samples.

The researchers at CRISP Research Centre have compared the information content of titles with respect to the vacancy full descriptions, finding that about 30% of the offer titles do not carry enough information to identify the occupation. This, in turn, motivates us in including the descriptions of job vacancies within the classification process, leveraging a deep semantic analysis of the related texts.

Furthermore the classification performed over the complete dataset which revealed a distribution of job occupations very close to the gold classification distribution, as we shown in [15], and this reveals that, from a statistical perspective, the job offers in the gold classification are representative of the observed population (i.e., the whole dataset). This result is not surprising as the job offers sample has been selected using statistical sampling techniques by researcher at CRISP.

In addition, we exploited a well-suited multidimensional visualisation technique to investigate the classification results, namely the *parallel-coordinates*, here omitted due to the space restrictions. Nonetheless, we invite the reader to refer to [15] for details while the online demo has been made publicly available[2].

4.1 Exploiting the Affinity Metric

As we introduced in Sec.1, the ISTAT classifier defines an *affinity* property between a pair of occupations. This relationship allows expressing that two different occupations are closely related with respect to their occupational profiles, while remaining distinctly categorised within the classification system. To give a few examples, the ISTAT classification system declares that *specialists in computer science and mathematics* (code 211) are affine to *engineers* (code 221) that, in turn, are affine to *directors and managers* (code 122).

[2] `http://goo.gl/6qC5Vj`

The idea here is to exploit these relationships between occupations to have a fine-grained evaluation of how far is the occupation code returned by a classification algorithm with respect to the correct one provided by the gold classification benchmark. This, as a result, makes us able to perform a detailed evaluation of the classification algorithm effectivenesses.

To this end, we build a graph G with unary edge weights where the nodes set is composed by occupation codes $O = \{o_1, o_2, \ldots, o_n\}$. The edges set E is built by creating an edge e_{ij} between two occupation codes $o_i, o_j \in O$ if and only if the affinity relationship is defined between them, that is $e_{ij} \in E \longleftrightarrow \mathit{affinity}(o_i, o_j)$. No loops over the same node are allowed. Furthermore, a dummy node o_0 with no outgoing edges has been included to handle items not classified by the *rule-based* algorithm.

Given such a kind of graph, we compute the *distance* between two occupations o_1, o_2 as the shortest-path between them. Notice that for an occupations pair a path could not exist (i.e., either there is no affinity between these occupations or the occupation has not been classified). Although in theory this latter distance should be infinity, for analysis purposes we set it to the diameter of the graph, that is 13.

For the sake of clarity, we can formalise our *distance* function as follows.

$$\forall o_1, o_2 \in G distance(o_1, o_2) = \begin{cases} 0 & \text{if } o_1 = o_2 \\ diameter & \text{if } \nexists \text{ path between } o_1, o_2 \\ |ShortestPath(o_1, o_2)| & \text{otherwise} \end{cases}$$

Result Comments. Table 1(b) reveals that the LDA approach has the lowest average shortest-path length. This means that LDA tends to classify a job offer over an occupation code that is distant averagely 1.5 from the correct classification, while *rules-based* approach here has the worst performance, mainly due to unclassified job offers that automatically assigns to it the maximum distance. Furthermore, if we neglect the infrequent job offers (the ones occurring less than 5 time in the sample) all the distance values improve and, here, the *linear SVC* reaches an average distance less than 1. In other words, the SVC classifier averagely classifies over an occupation code that is *affine* with the right one.

This trend is also confirmed by the plot given in Fig. 1, where the average shortest path lengths have been grouped by occupation codes. The lower graph shows the frequency of each occupation code within the gold benchmark, while the upper graph figures out the average shortest path lengths for each occupation code. As one might imagine, for supervised learning approaches the highest the number of job offers with the same code, the lowest the average distance from the gold benchmark. Differently, *rule-based* algorithm performs averagely well even on infrequent job offers.

5 Conclusions and Expected Achievements

In this paper we have compared several techniques to classify vacancies against an occupation classifier used by the Italian National Institue of Statistics (IS-

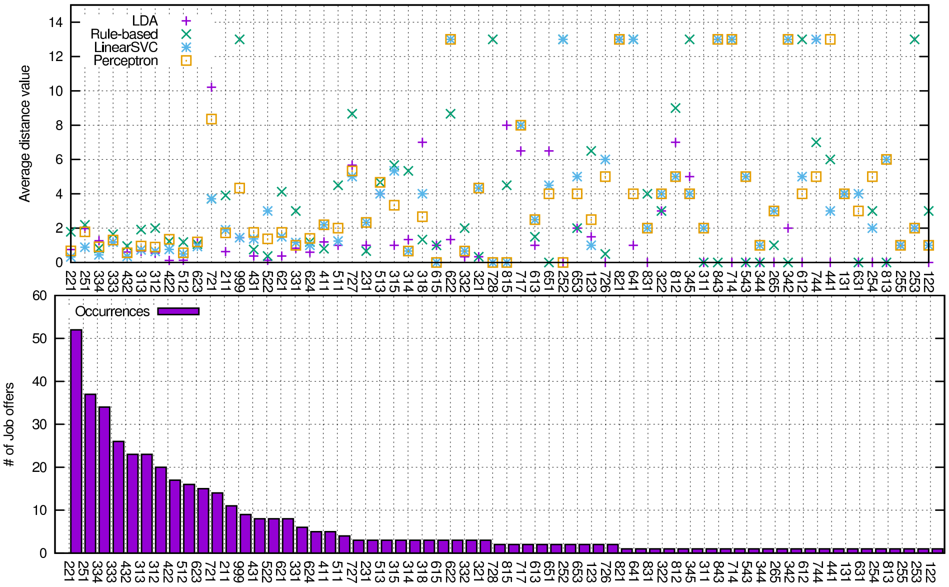

Fig. 1. Average Distances for each classification algorithm

TAT). The classification techniques were firstly evaluated using a hit/not-hit approach (i.e. either the occupation code was correctly predicted or not with respect to the gold benchmark), however to improve the evaluation results we built a distance function on top of the *affinity* information given by the ISTAT classifier. This distance function enables a more finely grained evaluation on the classification algorithm outcomes. All these results are currently under investigation as they guide our research towards the creation of a unified framework for classifying Web job offers.

To this end, as a future work, we scheduled the following research steps. First, to consider the full descriptions of job vacancies and to tackle the increased complexity of longer (and noisier) texts through computational linguistic approaches. Second, to use classification techniques for extracting other relevant information from the vacancy texts (in addition to occupation codes) e.g., the required skills, contract types, business sectors, education levels, etc.

References

1. Ralph Grishman. Information extraction: Capabilities and challenges, 2012.
2. Christopher S. Butler. *Statistics in Linguistics*. Blackwell, Oxford, 1985.
3. Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer, 2012.
4. Flora Amato, Angelo Chianese, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperli. Snops: a smart environment for cultural heritage applications. In *Proceedings of the twelfth international workshop on Web information and data management*, pages 49–56. ACM, 2012.

5. Flora Amato, Antonino Mazzeo, Vincenzo Moscato, and Antonio Picariello. Exploiting cloud technologies and context information for recommending touristic paths. In *Intelligent Distributed Computing VII*, pages 281–287. Springer, 2014.
6. F. Amato, A. Mazzeo, A. Penta, and A. Picariello. Knowledge representation and management for e-government documents. volume 280, pages 31–40, 2008.
7. Flora Amato, Antonino Mazzeo, Vincenzo Moscato, and Antonio Picariello. Semantic management of multimedia documents for e-government activity. 2009.
8. Amit Singh, Catherine Rose, Karthik Visweswariah, Vijil Chenthamarakshan, and Nandakishore Kambhatla. Prospect: a system for screening candidates for recruitment. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 659–668. ACM, 2010.
9. Roberto Boselli, Mario Mezzanzanica, Mirko Cesarini, and Fabio Mercorio. A policy-based cleansing and integration framework for labour and helthcare data. In *Knowledge Discovery and Data Mining, LNCS 8401*, pages 141–168. Springer, 2014.
10. Mario Mezzanzanica, Roberto Boselli, Mirko Cesarini, and Fabio Mercorio. A model-based evaluation of data quality activities in KDD. *Information Processing & Management*, 51(2):144–166, 2015.
11. Kun Yu, Gang Guan, and Ming Zhou. Resume information extraction with cascaded hybrid model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 499–506. Association for Computational Linguistics, 2005.
12. Xing Yi, James Allan, and W Bruce Croft. Matching resumes and jobs based on relevance models. In *30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 809–810. ACM, 2007.
13. Wenxing Hong, Siting Zheng, and Huan Wang. Dynamic user profile-based job recommender system. In *Computer Science & Education (ICCSE), 2013 8th International Conference on*, pages 1499–1503. IEEE, 2013.
14. Marc Poch, Núria Bel, Sergio Espeja, and Felipe Navıo. Ranking job offers for candidates: learning hidden knowledge from big data. In *Language Resources and Evaluation Conference*, 2014.
15. Flora Amato, Roberto Boselli, Mirko Cesarini, Fabio Mercorio, Mario Mezzanzanica, Vincenzo Moscato, Fabio Persia, and Antonio Picariello. Challenge: Processing web texts for classifying job offers. In *Semantic Computing (ICSC), 2015 IEEE International Conference on*, pages 460–463, Feb 2015.
16. Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.
17. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
18. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.
19. Francesco Colace, Massimo De Santo, Luca Greco, and Paolo Napoletano. Improving text retrieval accuracy by using a minimal relevance feedback. In *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, volume 348 of *Communications in Computer and Information Science*, pages 126–140. Springer, 2013.
20. Francesco Colace, Massimo De Santo, Luca Greco, and Paolo Napoletano. Text classification using a few labeled examples. *Computers in Human Behavior*, 30(0):689 – 697, 2014.