# Analysing Textual Content of Educational Web-Pages for Discovering Features useful for Classification Purposes

Vladimir Estivill-Castro, Matteo Lombardi and Alessandro Marani

School of Information and Communication Technology
Nathan Campus, Brisbane, 4111, QLD, Australia
Email: v.estivill-castro@griffith.edu.au
matteo.lombardi@griffithuni.edu.au alessandro.marani@griffithuni.edu.au

***Abstract*—Studies in Information Retrieval and Technology Enhanced Learning have not been able yet to propose reliable support to students and teachers when seeking educational resources on the Web. The driving force of web-search has been to match the topic of a query with the topic of documents. This paper involves Natural Language Learning approaches for an in-depth analysis of the common traits among educational web-pages. We analyzed the textual content of resources coming from educational websites and a survey among instructors. We computed more than 100 attributes and tested their significance for classification against web-pages from non-educational sources. Our analysis selected a set of 53 attributes. The results of a classification task prove that our traits allow for highly accurate filtering of resources with educational purposes. Moreover, the reliability of the proposed methodology is statistically verified.**

***Keywords*–Learning Objects; Internet based systems; Navigational aspects for on-line learning; Recommender Systems.**

## I. Introduction

The Internet is among the most popular places students and teachers explore educational resources to support their educational tasks [1]. However, search engines like Google and other web-based recommender systems still struggle in suggesting web-pages matching to a specific purpose of interest, for example, *education* [2]. Automatically identifying web-content suitable for education is one of the most challenging objectives because it requires extraordinary attention.

Studies in Information Retrieval (IR) and Technology Enhanced Learning (TEL) have proposed several solutions to support teaching and learning needs of instructors and pupils within an enclosed platform [3] [4]. However, those research efforts have not been able yet to recommend a reliable tool that can leverage the potentially infinite amount of pedagogical resources hosted on the Internet for helping users during their educational tasks. As a result, after receiving recommendations from existing search engines, students and teachers must spend additional time and effort to filter only web-resources useful for education. Personalization has improved web-search by identifying what topics users prefer, and some progress has been achieved in deducing the purpose of the search (e.g., the user is about to book a trip) for tailored advertising [5]; however, this is a very different use of recommendation. Instead, we focus here on identifying documents with a purpose in the sense of being of value for a learning objective.

Our exploration covers more than 2,300 web-pages obtained from the Seminarsonly website [6], and other sources human instructors [7] identified as relevant for teaching. We incorporate semantic technologies when processing natural language to we elicit more than 100 features computed directly from the text of web-resources. We analyze our features to discover which of these become attributes that permit a clear distinction between resources suitable for education and those not suitable. The resulting feature set is evaluated performing a binary classification of items in our dataset. We built such dataset labeling the aforementioned educational web-pages as “relevant for education”. We labeled as “not relevant for education” pages crawled from the former DMOZ Web directory, currently known as Curlie [8]. Our evaluation covers learning with several representatives of classification algorithms. We apply Student’s t-test to strengthen the validity of our feature set. In particular, we tested the accuracy distribution across the results of a 30-fold cross validation when using all the selected traits, and when reducing the feature space utilizing Principal Component Analysis (PCA) and Support Vector Machine (SVM). The t-test confirms that all the features are essential for achieving the best accuracy in our filtering task for each classifier.

Section II which describes our data-set. Section III describes or features. Our evaluation methods appears in Section IV. Section V reports experimental results. Section VI contrast our work and Section VII provides conclusions.

## II. Data analysis

We involved Semantic Web techniques and organized the information into semantic entities extracted from the textual content of web-pages, where a *semantic entity* is an instance of a DBpedia [9] resource that groups a collection of properties. Semantic entities can be associated with one or more consecutive words. Following other contributions [10] [11] [12], we use the Dandelion API [13] for deducing all the semantic entities in text. In our case, we simplify the analysis of complex and articulated texts by considering the semantic entities extracted from them as their representation. We suggest that, when the text is an educational resource, semantic entities contain the most distinctive pieces of information about what the content, concepts, knowledge and skills educators deliver through the text. Hence, we expect that a set of entities will represent the entire text reflecting the same knowledge content without losing any proper traits.

### A. The dataset

Our goal is to extract features from web-pages and test their validity to recognize whether or not a web-page is suitable for educational purposes. Hence, the items in our dataset are web-pages with two possible values for the class: TRUE, when a resource has been declared relevant for teaching some concepts, or FALSE when the page does not contain educational content. Our dataset consists of more than 2,300 educational web-pages we extracted from two different sources. The first source is the Seminarsonly website, which hosts content about Computer Science, Mechanical, Civic and Electrical Engineering, as well as Chemical and Biomedical sciences among others. The second source of educational material is a subset of web-pages ranked by instructors during a survey [7]. The survey's first phase automatically used queries by an intelligent system against a search engine with names of educational concepts and courses. The second phase exposed groups of 10 retrieved pages to instructors who judged the suitability of the web-page as a learning-object suitable for teaching. In particular, whether the page could support the learning of the concepts of the query in the originator course. The instructors used a 5-point Likert scale to rank how likely would they use that web-page for teaching a concept. When web-pages are highly ranked uniformly by judges, it is certain that the page is suitable for being used in an educational context. For that reason, in this analysis, a web-page is labelled as TRUE ("relevant for education") only when it collected 3 points (relevant) or more (where the maximum is 5 points —- Strongly relevant) in the survey. Other pages from the survey are discarded. On the other hand, we obtain the web-pages classified as FALSE ("non-relevant for teaching") by the crawling of URLs contained into the DMOZ open directory. In particular, we included pages coming from all the 15 categories represented in DMOZ, resulting in more than 3,200 web-pages. We consider those web-resources not suitable for teaching. In total, our dataset consists of around 5,600 labelled web-pages, according to their usability in educational contexts.

### B. Extraction of Semantic Entities

We exploit DBpedia entities extracted from web-pages for deducing information about the content of a whole page. For each extracted entity, Dandelion also reports a confidence value for that association. The higher the confidence, the more reliable the link between the part of the text and the entity. The tool also allows to select a threshold of minimum confidence for the extraction, avoiding to retrieve poorly related entities. Hence, the higher the confidence threshold, the higher the effectiveness of the extraction process but, on the other hand, the number of entities extracted tends to decrease when the threshold is high. DBpedia also offers the *type* of an entity: places, companies and personal names. When no match is found, Dandelion assigns the type *Concept* to the entity (refer to Figure 1).

## III. Feature elicitation process

We analyze four parts of each web-page separately: the *Title*, the *Body*, the *Links* and the *Highlights*. We extract the last two from the body itself of the page. In particular, the *Title* is extracted from the title tag and the *Body* element from the body tag. Then, inside the *Body* tag, the text between the anchor $< a >$ tags is concatenated and labeled as the *Links*, while we obtain the *Highlights* by merging the text between the tags $< h1 >$, $< h2 >$, $< h3 >$, $< b >$ and $< strong >$ . In this way, we separate all the four elements of a web-page, allowing for a thorough analysis of the page itself.

We apply the same approach to all the four parts of a web-page. In the end, we may find a feature that is significant for classification purposes when considering a specific part of the page (e.g., the *Links*), while the same feature could be discarded for a different part (for instance, the *Title*). For that reason, we run the Dandelion API Entity Extraction tool on all the resources in our dataset, considering one part of a web-page at a time, so that the entities will also have a label that indicates the part of a page from which they originated.

The following sections present the groups of features extracted from our resources. For each group, we selected the semantic entities according to four different thresholds for the confidence: the default 0.6, then 0.7, 0.8 and finally 0.9.

### A. Lexical features

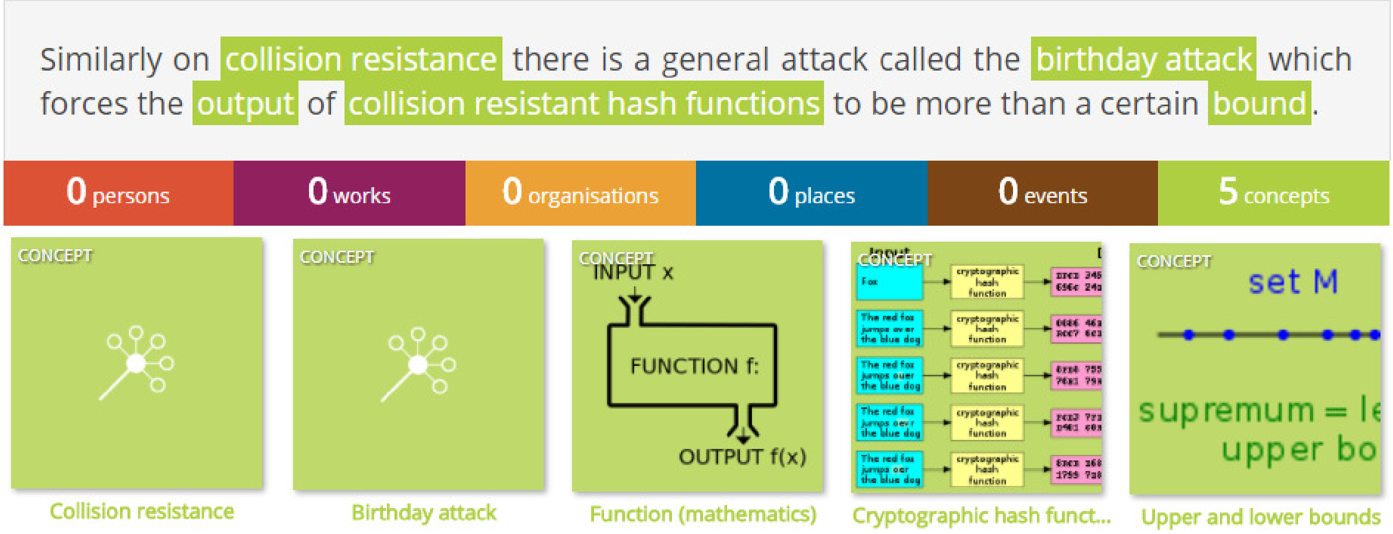
We base the first group of features on NLP for discovering characteristics and quantity of the terms used in a web-page. In particular, the following attributes exploit the complexity of the words, as well as the number of semantic entities and concepts relative to the length of a text.

1. The Complex_Words_Ratio $= \frac{\text{\# complex words}}{\text{\# words}}$ is is the ratio of the number of complex words on the total number of words (i.e., the length) in a text. We used the Fathom API [14] for deducing the quantity of complex words: words composed by three or more syllables.
2. Feature Number_entities is the total # of entities of any type extracted from a text.
3. Entities_By_Words $= \frac{\text{\# entities}}{\text{\# words}}$ is the number of entities extracted from a text, with respect to the total number of words. This feature measures how many words it is necessary to read for finding a semantic entity.
4. Concepts_By_Words $= \frac{\text{\# concept entities}}{\text{\# words}}$ is a feature similar to the Entities_By_Words, but considering only the concept-type entities. The idea is to have an insight into how many words it is necessary to read for finding a concept.
5. Concepts_By_Entities $= \frac{\text{\# concepts}}{\text{\# entities}}$ reports the fraction of entities that are also concepts, with respect to the total number of entities found in a text.

### B. Features based on Semantic Density

Researchers in TEL refer to *Semantic Density* (SD) as the quantity of topics presented by a resource with respect to a characteristic of the resource itself. For instance, the IEEE Learning Object Metadata schema defines the Semantic Density of a resource as the ratio of the number of concepts taught on the length of the resource (commonly measured in minutes or hours). As a result, a resource yields high SD when many topics are squeezed in a short time frame.

We consider the different entities in a text as topics delivered by a resource. Then, we measure two different SD values for a text: one value concerning the number of words, and the other related to the reading time (similarly to the aforementioned IEEE standard). For an even more comprehensive analysis of the text, we also take into account only the concept entities. In the end, we compute SD of a web-page using four attributes.



Similarly on collision resistance there is a general attack called the birthday attack which forces the output of collision resistant hash functions to be more than a certain bound.

| 0 persons | 0 works | 0 organisations | 0 places | 0 events | 5 concepts |
|---|---|---|---|---|---|

| CONCEPT | CONCEPT | CONCEPT | CONCEPT | CONCEPT |
|---|---|---|---|---|
| Collision resistance | Birthday attack | Function (mathematics) | Cryptographic hash funct... | Upper and lower bounds |

INPUT x
FUNCTION f:
OUTPUT f(x)
cryptographic hash function
Fox
The red fox jumps over the blue dog
set M
supremum = le
upper bo

Figure 1. Entities in a text found by Dandelion API.

**1.** SD_By_Words $= \frac{|\text{Entities}|}{\#\ \text{words}}$ measures how many distinct entities Dandelion extracted from the text (i.e., the set of discussed topics), with respect to the number of words. When two texts have similar quantities of words, the one with more distinct entities is the denser.

**2.** Similarly to the previous feature, SD_By_ReadingTime $= \frac{|\text{Entities}|}{\text{reading time}}$ is now measured in relation to the reading time of the text. In this case, the text is denser when the reading time is low, and the number of distinct entities (i.e., topics) is high.

**3.** SD_Concepts_By_Words $= \frac{|\text{Concepts}|}{\#\ \text{words}}$ considers only distinct concept entities, with respect to the number of words. In educational texts, concept-type entities are more frequent than other types. Hence, the concept-based SD is expected to hold significant information for educational classification.

**4.** SD_Concepts_By_ReadingTime $= \frac{|\text{Concepts}|}{\text{reading time}}$ measures the quantity of concepts taught by a text according to the expected reading time. As an example, let us consider two texts where Dandelion extracted the same amount of distinct concepts. In that case, the text which requires less reading time presents concepts in a more condensed way, so it holds higher SD than its counterpart. In essence, less time is spent for other entities (i.e., non-concepts) that are not likely to be used in educational resources, while important concepts receive more attention.

### C. *Selection of the most promising features*

At this stage, nine groups of numerical features represent each web-page. In our dataset, the content of a single item is split across four web-elements. Furthermore, for each element of a page, entities are extracted at four different thresholds, except for the Complex_Words_Ratio group, which leverages only natural language text so it does not require semantic entities extraction. Since the first four attributes in the count are those that involve the ratio of complex words, and we include one feature for each element of the page, we have

$$\#\ \text{potential\_features} = 4 + 8 * 4 * 4 = 132 \text{ features.}$$

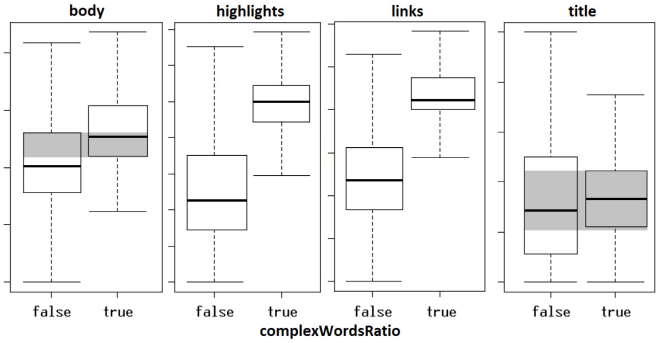


Figure 2. The distribution of the four features in the Complex_Words_Ratio group, according to the class.

However, some of those features may not be useful to discriminate between a resource relevant for education and one not suitable for that purpose. We use a parallel coordinate visualization [15] of groups of features and we select only the traits where a visual distinction is clear among the web-pages in our dataset. Our filtering process is performed according to the distribution of the values of each feature, and we now explain it in the following paragraphs. The criterion for selecting or discarding a feature is that there is no overlap between the most frequent values of the TRUE and FALSE distributions, namely, the values from the first quartile (Q1) to the third quartile (Q3) in a box plot representation. In the interest of saving space, we discuss only the first two groups of features and show the box plots for their distributions. The first group is **Complex_Words_Ratio**. Figure 2 illustrates that the *Highlights* and the *Links* distributions overlap between classes only across the quartiles Q1 and Q3. The area in gray highlights that most of the values from first to third quartile are in common for the *Body* and *Title* elements, while *Highlights* and *Links* are able to separate TRUE and FALSE items with high accuracy. But the *Body* and *Title* distributions display significant commonality for the their most frequent values. Hence, the two features

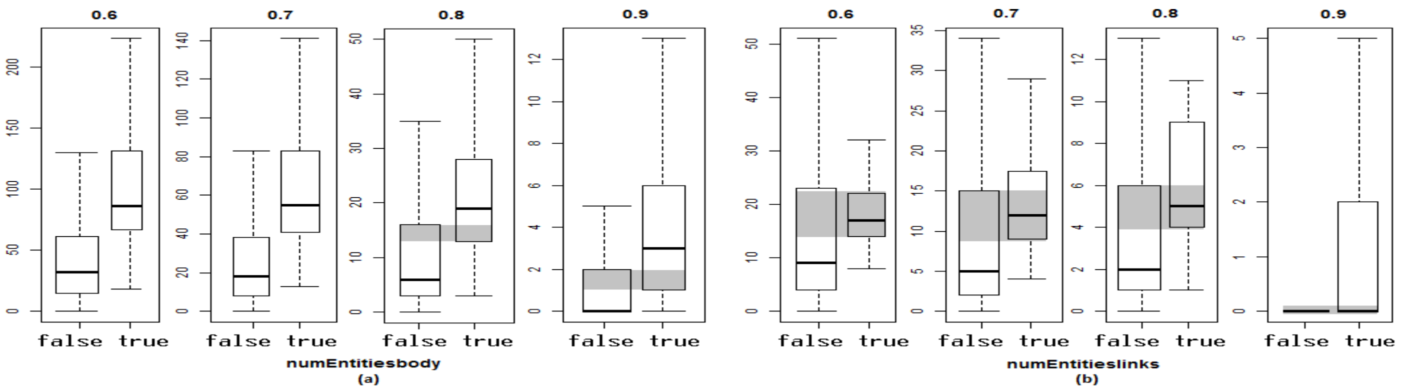


Figure 3. Analysis of TRUE and FALSE items distributions for features in the **Number_entities** group, extracted from *Body* (a) and *Links* (b) elements of a web-page. The gray areas indicate an overlap.

TABLE I. THE 53 ATTRIBUTES SELECTED FOR THE OVERALL FEATURE SET, DENOTED BY A ⋆ SYMBOL.

| **Group** | *Body* | | | *Links* | | | *Highlights* | | | *Title* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | .6 | .7 | .8 | .6 | .7 | .8 | .6 | .7 | .8 | .6 | .7 | .8 |
| **Complex_Words_Ratio** | | | | | ⋆ | | | ⋆ | | | | |
| **Number_entities** | ⋆ | ⋆ | | | | | | | | ⋆ | ⋆ | |
| **Entities_By_Words** | ⋆ | ⋆ | | ⋆ | ⋆ | ⋆ | | | | | | |
| **Concepts_By_Words** | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | | | | |
| **Concepts_By_Entities** | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | | | | | | |
| **SD_By_Words** | | | | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | | | |
| **SD_By_ReadingTime** | | | | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | | | |
| **SD_Concepts _By_Words** | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | | | | |
| **SD_Concepts _By_ReadingTime** | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | ⋆ | | | | |

selected for this group are Complex_Words_Ratio_*Links* and Complex_Words_Ratio_*Highlights*, while the others are discarded. Moving to the **Number_entities** group, there are 16 possible combinations amongst 4 threshold values and 4 elements of the web-page. Here we show 8 of those potential attributes. The first four from the left (Figure 3a) are about the count of entities found in the *Body* considering the four values of confidence thresholds, while the others (Figure 3b) consider just entities found among the *Links*. Only 2 out of 8 attributes are useful for classification. They are Number_entities_*Body*_0.6 and Number_entities_*Body*_0.7, because all the other distributions overlap between TRUE and FALSE items. Interestingly, when threshold is 0.9, the number of entities dramatically decreases in both educational and non-educational web-pages. Especially among the latter group, there are only form 0 to 2 entities in the *Body*, and none in the *Links*. Since all the features computed at threshold 0.9 experience the same decrease, in order to have a fair comparison, we discard them. The remaining 8 traits for this group are computed taking into account the *Highlights* and *Title* elements. In the first case, all the distributions overlap so none of the attributes is selected. About *Title*, distributions of entities at threshold 0.6 and 0.7 do not overlap so they are selected, while raising the threshold to 0.8 the two distributions overlap. We apply the same methodology to the other groups. What features are selected as discriminators by the above analysis is summarized in Table I. Note that group Complex_Words_Ratio does not require entity extraction, therefore, it has only one attribute per page element.

## IV. Methodology for Evaluation

In the evaluation phase, we aim to see whether or not the 53 proposed attributes allow state-of-the-art classifiers to achieve high accuracy in recognizing the web-pages labeled as relevant for education in our dataset. In order to achieve that goal, we applied popular feature selection algorithms to our set of traits, and then we compared the accuracy on the same set of classifiers. The rationale behind our choice is that some features may be discarded by generic algorithms as not useful or redundant, or combined to obtain a new set of attributes. However, in case the overall accuracy decreases applying feature selection methods, we can conclude that the proposed features allow classifiers to yield higher performance in an educational task, thus, all 53 traits are important when filtering web-pages in such field. The algorithms for feature selection chosen as baselines in this work are Principal Component Analysis - PCA [16] and Support Vector Machine - SVM [17].

### A. Classifiers and evaluation measure

In order to produce a comprehensive evaluation across all types of machine-learning algorithms for classification, we used state-of-the-art classifiers belonging to four families, namely Bayesian, Rule-based, Function-based, and Tree-based classifiers, for a total of eight algorithms. From the first family, we chose **Bayesian Network** built with hill-climbing method [18]. The three rule-based methods involved are **Decision Table** [19], **Repeated Incremental Pruning to Produce Error Reduction - RIPPER** [20] and **Partial decision list - PART** [21]. From the function-based classifiers we selected **Logistic** [22] and **Sequential Minimal Optimization - SMO** [23]. Finally, as tree-based classifiers, we opted for **J48**, which builds a pruned C4.5 decision tree [24], and the popular **RandomForest** algorithm [25]. We used the default implementation and parameters provided by WEKA for all classification methods. We recorded the performance of the classifiers on a 30-fold Cross Validation according to their **Average Precision (AP)**, which is the mean of the **Precision (P)** in a classification task across all the 30 folds:

$$P(f) = \frac{\text{\# correctly classified items}}{\text{\# items}} \quad , \quad AP = \sum_{f \in folds} \frac{P(f)}{\text{\# folds}}.$$

where $f$ is the i-th fold, and # folds is 30 in this study. We present our results in the next section as percentage values.

In addition, we aim to strengthen our claim performing a statistical analysis of our feature set against those generated by PCA and SVM, comparing the distribution of P in all the folds using the Student's paired T-test. The null hypothesis $h_0$ to be investigated is:

$$h_0 = \text{The chosen feature set does not influence P.}$$

While the alternative hypothesis $h_1$ is:

$$h_1 = \text{P is higher when using all 53 features.}$$

If $h_0$ is significantly rejected and $h_1$ confirmed, we demonstrate the actual validity of all the attributes proposed in this work. To verify at least a 95% of such significance, we look for values of $p$<0.05 in our T-tests. We ran PCA, SVM and the classifiers using the WEKA 8.3.2 Java library with default parameters. The entire evaluation is performed on a Windows 10 machine, with Intel i7-6700 octa-core processor @ 3.4GHz and 32GB of RAM.

## V. Results

As described in Section IV, we applied two state-of-the-art feature selection algorithms, PCA and SVM, to build two sets of attributes we will use as baselines throughout our evaluation. To achieve a more comprehensive comparison, we created those two sets differently. The first one, called *PCA*, is obtained running PCA on our dataset. The number of resulting components, in this case, is fourteen. The second set of traits comes from SVM, a method for ranking features. We selected the ten most valuable attributes according to the SVM algorithm, forming the *Top10-SVM* feature set.

Figure 4 shows the AP measured when running different classifiers using the two aforementioned baselines, and our 53 attributes. We call our feature set *AllFeatures*. In every test performed, the proposed set *AllFeatures* allows classifiers to obtain the highest precision in average on the 30 folds of the cross-validation testing. However, we also performed statistical testing to verify if we can reject the null hypothesis $h_0$ (namely, "there is no evidence that the chosen feature set influences the precision of a classifier") and accept the alternative $h_1$. In particular, since we have two baselines, two alternative hypotheses will be verified:
$h_1^{PCA}$ = *"A classifier achieves higher precision when considering all features than the ones by PCA"*
$h_1^{SVM}$ = *"A classifier achieves higher precision when considering all features than the ones by SVM"*.

Table II reports the results of the Student's T-test performed in our evaluation. We verified a significance of at least 95% for our hypotheses considering each classifier. We reached higher statistical significance, around 99% ($p$-value<0.01) for $h_1^{PCA}$ on the majority of the classifiers. Only **BayesNet** has a slightly higher $p$-value (0.01359). However, it is still lower than 0.05. When testing our 53 features against those labeled most important by SVM, also $h_1^{SVM}$ is accepted with 99% or more significance on all the algorithms but one. Indeed, the $p$-value when using **DecisionTable** is 0.01688, yet smaller than the required threshold of 0.05.

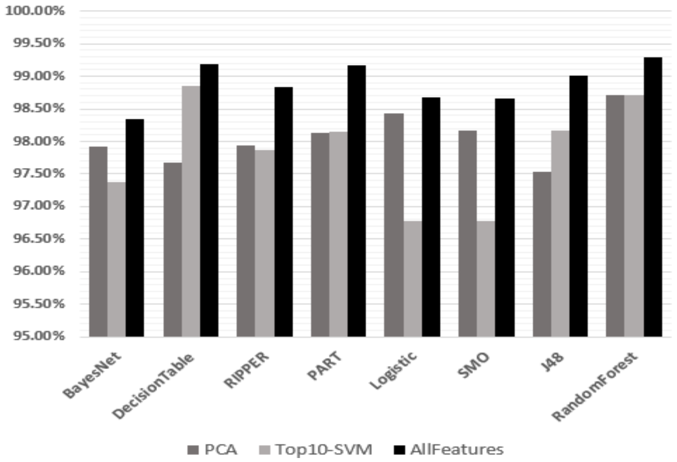


Figure 4. The average precision (AP) computed for each classifier when using the different feature sets analyzed in our evaluation process.

TABLE II. STUDENT'S T-TEST RESULTS FOR EACH CLASSIFIER. THE DESIRED $p$-VALUE <0.05 IS INDICATED WITH "*", A $p$-VALUE <0.01 IS LABELED WITH "**".

| **Classifier** | *AllFeatures* vs. *PCA* | | *AllFeatures* vs. *Top10-SVM* | |
|---|---|---|---|---|
| | T | $p$-value | T | $p$-value |
| **BayesNet** | 2.3266 | 0.01359 * | 7.3054 | 2.39E-08 ** |
| **DecisionTable** | 6.5606 | 1.73E-07 ** | 2.2284 | 0.01688 * |
| **RIPPER** | 5.0055 | 1.25E-05 ** | 4.8125 | 2.14E-05 ** |
| **PART** | 5.2519 | 6.30E-06 ** | 5.2318 | 6.66E-06 ** |
| **Logistic** | 2.5343 | 0.008463 ** | 10.15 | 2.35E-11 ** |
| **SMO** | 4.0649 | 0.0001677 ** | 9.6948 | 6.64E-11 ** |
| **J48** | 7.6944 | 8.73E-09 ** | 4.4585 | 5.69E-05 ** |
| **RandomForest** | 4.2105 | 0.0001126 ** | 4.3679 | 7.31E-05 ** |

## VI. Related work

Extraction and selection of attributes from a text is a popular research topic [26] [27]. Recently proposed approaches are also based on alternative methods from other research fields. For instance, [28] applied a technique for encoding signals called Wavelet Packet Transform for web-page analysis. Also deep learning methods like Convolutional Recurrent Neural Network [29] have been applied for the classification of relations in texts. To elicit features useful for filtering educational web-resources, our approach leverages techniques for analysing texts coming from the Knowledge Management, Information Retrieval and the Semantic Web communities. In the field of Education, [12] used semantic entities from DBpedia to describe and enrich texts coming from the Coursera [30]. platform.

Additional criteria have been suggested when dealing with content from the Web: Several studies shown how latent information can be found analysing both text and structure of web-pages. [31] suggested a methodology for deducing the category of a web-page considering the loading time of different objects like images, CSS theme, Javascript code and Flash content. However, only a group of 6 categories can be deducted, and educational-related ones are not part of it. Also, [32] proposed a more general approach which takes into account the fields of web-pages such as title, body and anchor text (i.e., the text used to embody a URL) for evaluating datasets of web-pages. [33] demonstrated that links in a web-page are important for automatic classification; thus these authors exploited links for deducing pages of academic institutions. However, their work is about identifying pages useful for extracting the internal organization of an Institute, rather than educational resources delivered in educational coursework.

## VII. Conclusions

We examined a dataset of more than 5,600 web-pages with the goal of identifying the purpose of a web-page (suitability as an educational resource). This is a very different task than recognizing the subject matter nor the topic of a web-page. We attack this problem by seeking what features can be extracted from web-pages and their content. We proposed and identified those useful for classifying online resources for the purpose of education. We incorporated techniques from both natural language processing and semantic analysis for the definition of an initial set of 132 potential predictors. Then, the most promising traits are the output of an in-depth feature selection process which results in a set of 53 characteristics extracted from four sections of a web-page (see Table I). We evaluated the validity of our proposed features on the binary classification task that discriminates whether the purpose of the web-page is educational. In particular, we performed a 30-fold cross-validation test on our dataset using several state-of-the-art classifiers of many types and learning models. As baselines, we used feature selection algorithms for reducing the number of attributes according to two general approaches: PCA and SVM. We demonstrated that the average precision (AP) across the folds is higher when using our suggested 53 features. Furthermore, results of Student's T-test strengthen our proposal with all test repetitions achieving $p$-value $< 0.05$, and many lower than 0.01. This statistical significance at very high levels for all classifiers confirms the features are informative and effective in providing discrimination capacity to classifiers across several families. We expect our work to facilitate retrieval and recommendation of web resources suitable for specific purposes, especially for helping students and teachers in educational tasks.

## References

[1] H. Drachsler, K. Verbert, O. C. Santos, and N. Manouselis, "Panorama of recommender systems to support learning," in Recommender systems handbook. Springer, 2015, pp. 421–451.

[2] M. Lombardi and A. Marani, "A comparative framework to evaluate recommender systems in technology enhanced learning: a case study," in Advances in Artificial Intelligence and Its Applications. Springer, 2015, pp. 155–170.

[3] C. Limongelli, M. Lombardi, A. Marani, F. Sciarrone, and M. Temperini, "A recommendation module to help teachers build courses through the moodle learning management system," New Review of Hypermedia and Multimedia, vol. 22, no. 1–2, 2015, pp. 58–82.

[4] S. Sergis and D. Sampson, "Learning object recommendations for teachers based on elicited ict competence profiles," Learning Technologies, IEEE Transactions on, 2015, pp. 67–80.

[5] J. Arora, S. Agrawal, P. Goyal, and S. Pathak, "Extracting entities of interest from comparative product reviews," in Conf. on Information and Knowledge Management. ACM, 2017, pp. 1975–1978.

[6] Seminarsonly, http://www.seminarsonly.com/.

[7] A. Marani, "WebEduRank: an educational ranking principle of web pages for teaching," Ph.D. dissertation, Griffith University, 2018.

[8] Curlie, https://curlie.org/.

[9] DBpedia, http://wiki.dbpedia.org/.

[10] D. Taibi, R. Rogers, I. Marenzi, W. Nejdl, Q. A. I. Ahmad, and G. Fulantelli, "Search as research practices on the web: The sar-web platform for cross-language engine results analysis," in 8th ACM Conf. on Web Science, ser. WebSci '16. New York, NY, USA: ACM, 2016, pp. 367–369.

[11] M. Brambilla, S. Ceri, E. Della Valle, R. Volonterio, and F. X. Acero Salazar, "Extracting emerging knowledge from social media," in 26th Int. Conf. on World Wide Web, 2017, pp. 795–804.

[12] C. Limongelli, M. Lombardi, A. Marani, and D. Taibi, "Enrichment of the dataset of joint educational entities with the web of data," in Advanced Learning Technologies (ICALT), IEEE 17th Int. Conf. on. IEEE, 2017, pp. 528–529.

[13] Dandelion, https://dandelion.eu/.

[14] Fathom, search.cpan.org/dist/Lingua-EN-Fathom/lib/Lingua/EN/Fathom.pm.

[15] A. Inselberg, Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Berlin, Heidelberg: Springer-Verlag, 2009.

[16] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and intelligent laboratory systems, vol. 2, no. 1-3, 1987, pp. 37–52.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine learning, vol. 46, no. 1, 2002, pp. 389–422.

[18] G. F. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," Machine learning, vol. 9, no. 4, 1992, pp. 309–347.

[19] R. Kohavi, "The power of decision tables," in 8th European Conf. on Machine Learning, ser. ECML'95. Springer, 1995, pp. 174–189.

[20] W. W. Cohen, "Fast effective rule induction," in 12th Int. Conf. on Machine Learning, A. Prieditis and S. J. Russell, Eds. Morgan Kaufmann, July 9th-12th 1995, pp. 115–123.

[21] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," in 15ht Int. Conf. on Machine Learning, ser. ICML '98. San Francisco, CA, USA: Morgan Kaufmann, 1998, pp. 144–151.

[22] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," Applied statistics, 1992, pp. 191–201.

[23] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in Advances in Kernel Methods - Support Vector Learning, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, January 1998, pp. 185–208.

[24] J. R. Quinlan, "C 4.5: Programs for machine learning," The Morgan Kaufmann Series in Machine Learning, San Mateo, CA, 1993.

[25] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, 2001, pp. 5–32.

[26] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in ICML, vol. 97, 1997, pp. 412–420.

[27] M. J. Paul, "Feature selection as causal inference: Experiments with text classification," in 21st Conf. on Computational Natural Language Learning (CoNLL), 2017, pp. 163–172.

[28] A. Mahajan, S. Roy, and S. Jat, "Feature selection for short text classification using wavelet packet transform," in 19th Conf. on Computational Natural Language Learning, 2015, pp. 321–326.

[29] D. Raj, S. Sahu, and A. Anand, "Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text," in 21st Conf. on Computational Natural Language Learning (CoNLL), 2017, pp. 311–321.

[30] Coursera, https://www.coursera.org/.

[31] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Characterizing web page complexity and its impact," IEEE/ACM Trans. on Networking, vol. 22, no. 3, 2014, pp. 943–956.

[32] S. Robertson, H. Zaragoza, and M. Taylor, "Simple bm25 extension to multiple weighted fields," in 13th ACM Int. Conf. on Information and knowledge management. ACM, 2004, pp. 42–49.

[33] P. Kenekayoro, K. Buckley, and M. Thelwall, "Automatic classification of academic web page types," Scientometrics, vol. 101, no. 2, 2014, pp. 1015–1026.