DISCOVERING SIMILAR ORGANIZATIONAL SOCIAL MEDIA STRATEGIES USING CLASSIFICATION AND CLUSTERING

Alvaro Figueira[1], Luciana Oliveira[2]

[1]CRACS / INESC TEC & University of Porto (Portugal)
[2]CICE / ISCAP & Polytechnic of Porto (Portugal)

Abstract

Organisations have been striving to account for the resources they've been allocating to Social Media integration and management, essentially because this integration has been occurring without a previously designed content strategy, which will foster the desired fan engagement.

In order to establish a comparison of social media strategies between HEIs, we developed a seven category model, encompassing the fundamental communication areas of focus for higher education service providers. Then, we performed a classification of these HEI posts in Facebook, according to our model. For this step, we used six of the most promising, and prominent, classifiers to obtain a predicted category for each post. Combining all posts from each HEI according to the model we get the HEI's editorial strategy. By clustering the overall social media strategies and corresponding response rate we discover the sector's monitoring HEI and, through a benchmarking process, we retrieve useful inputs for the design of social media strategies for HEI.

Keywords: Social media strategy, Automatic classification and clustering, Higher Education Institutions.

1 INTRODUCTION

The adoption of social media channels by Higher Education Institutions (HEI) has not been accompanied by a carefully planned and strategically designed process. The adoption of social media channels has mainly been sustained by HEI joining the trend to adopt social media environments, aiming at mediatisation, but lacking alignment with organizational goals and the adoption of efficient monitoring and benchmarking methods, which can address the need to measure the efficiency of communication and conversation on social media.

Bearing in mind that a social media strategy needs to be aligned with and framed in the overall organizational strategic management goals, the structure of its communication strategy should provide a clear indication of the HEI's positioning. However, the right question relies on: how can one attempt to measure the efficiency of social media approach that has not been strategically designed and is the result of a set of unarticulated and situational messages? Thus, on a first stage it is not relevant, neither possible, to determine the social media strategies' level of efficiency, which organizations constantly seek. However, determining the organizational positioning of HEI current strategies will allow to combine monitoring and benchmarking methods to foster the identification opportunities and threats, which can serve as inputs for the internal evaluation of social media strategies', for the necessary strategic readjustments and a subsequent efficiency measurement.

In order to optimize their social media strategy, organisations should first seek a comprehensive analysis of previous content/activity's purpose on social media and a global sector perspective, so that they can establish a set of goals to be achieved. Built upon these insights it is then possible to evaluate one organisation's overall performance, to determine its social media strategy signature and to introduce valuable improvements to it. Ultimately, organisations should be able to acquire knowledge on their positioning inside their sector, and to identify best practices from leading agents' strategies.

In order to accomplish this, we begin by presenting which are the key relevant editorial areas for the HPPES (Higher Public Polytechnic Education Sector) on social media. On section three, we explain how the retrieval and classification of social media messages was conducted for the entire sector, in order to obtain a sector overview of relevant editorial areas and each agent's individual aggregation of areas. This sector analysis is then thorough detailed allowing us to report on the main aggregations of editorial areas that are being conducted by HPPEI, by comparing and clustering 43 vectors (43 agents) in a 7-dimensional vector space model (7 editorial areas). The achievements of the

implementation of this methodology, in terms of efficiency and on acquired knowledge depth, are discussed in section five.

2 THE PROPOSED EDITORIAL MODEL

In order to establish a comparison of social media strategies between HEI's, we developed a seven category model, encompassing the fundamental communication areas of focus for higher education service providers: Education; Research; Society; Identity; Administration; Relationship and Information. In order to apply the methodology, we propose, if the identification of the editorial areas for a specific agent/market is not yet set, is incomplete or is only implicit, a small sample of social media messages should be run through a communication professional, so an efficient and comprehensive editorial model can be built.

2.1 The editorial areas

Specifically concerning the HPPES, some editorial areas are straight forward and some were added after a manual classification of a small sample of messages.

In any case, the previously identified guiding principles for the design of a social media editorial model (Table 1) where applied and tuned to the HPPES and we considered the following: (a) the heavy HEI's mission towards society and the great diversity of organizational stakeholders (students, faculty, staff, employers, partners, research centres, etc.); (b) the specifics of the educational service (a co-produced service); (c) a multi-channel wide holistic approach to communication management; (d) the need to balance between organizations' institutional and transactional needs in order to ensure their competiveness and financial survival; and (e) the dialogical nature that is intrinsically linked to social media environments.

Table 1. Editorial model for HPPES [1]

Education	Research	Society	Identity	Administration	Relationship	Information
- Promotes higher education courses (educational offer) - Promotes complementary training (internal or external)	- Informs on and / or calls for participation in: congresses, seminars and other scientific meetings - Promotes / informs on internal and external research results / awards - Promotes / informs on internal and external publications (papers, articles, books, proceedings, etc.)	- Promotes / informs on organizational partnerships and contracts and patents - Promotes employability, streaming placement offers and career opportunities - Promotes / informs on knowledge / technology transfer - Promotes other organizations' initiatives / performance (partners and other relevant stakeholders) - Promotes demonstrations, exhibitions and showcases, conducted by students or faculty	- Institutional events (celebrations, awards and tributes, graduation ceremonies, etc.) - Students, faculty and staff honorable mentions - Institutional promotion, advertising (identity, image, reputation) - Corporate Social Responsibility initiatives - Institutional clipping - Participation / representation in fairs and exhibitions	- Informs on deadlines and administrative processes - Informs on procedures and admissions - Promotes and informs on support services (goals, contacts, working hours, etc.)	- Fosters conversation - Requires opinions - Introduces current internal, external, societal or academic issues propelling audience involvement - Boosts emotional connection between organization and publics (greetings, humor, sympathy, motivation, etc.)	- Streams external relevant information, news and regulations related to academic areas, political and societal issues (economic and social impact) - Informs on recreational and cultural initiates with no particular connection to schools' scientific areas (concerts, sports events, etc.)

3 SOCIAL MEDIA COMMUNICATION ANALYSIS

The analysis presented in this paper encompasses one full academic year of messages posted on Facebook, by 43 agents of the HPPES. These 43 agents were selected among the sector's total set of 94 schools. The set of criteria for agent selection included as prerequisites: the agent is a provider of educational services and manages its own Facebook page before September, the 1st 2013.

3.1 Social network data retrieval

According to the defined criteria, research consisted of retrieving and classifying all messages posted by HPPEI on Facebook. We used two methods: an in-house made system, specially built for the purpose using the available Facebook API and a third-party software for collecting information from Social Networks. From an initial list of the relevant agent Page Id's, the two systems accessed the posts retrieving the following fields:

List 1. Fields collected from Facebook posts.

1) PostId	2) Message	3) Link	4) Name
5) Description	6) Caption	7) #Likes	8) #Comments
9) #Shares			

The two systems retrieved the same number of posts (15.104), during the entire school year, which consolidated our confidence about the validity of the returning set.

In this step we permed the classification of the 15.104 posts according to our editorial model, listed in the previous section. Clearly, this was a time demanding task to be done by hand and impossible to be undertaken on the-fly if done exclusively by humans. In another work [1] we proposed an automatic method to categorize the social media posts, based on the conjunction of several of the most recent and promising algorithms for text categorization.

For this step we used the six of algorithms:

Support Vector Machines. Linear SVMs are a machine learning algorithm [2] based on a geometric method that tries to separate two classes through a hyperplane, picking the one that maximizes the margin between the two classes. More recently, this method was evolved [3] to deal with a multiple number of classes. We used the Multi-class SVM lib for this analysis.

Random Forests. RFs were created [4] to overcome the overfitting effect of the decision trees. Within this method multiple decision trees are created during training time, and the mode of the resulting class is the presented output.

LogiBoost [5]. This algorithm belongs to a larger category of boosting algorithms which comprehend AdaBoost, LPBoost and some others, all based on a common framework called AnyBoost [6]. Generically, the boosting algorithms try to reduce variance and pre-training effects in supervised learning by re-weighting a set of classifiers according to the rule: weak classifiers should gain weight and strong classifiers should lose weight. The LogiBoost is implemented in several regression and classification packages. We used the one implemented in "caTools" for R.

K-Nearest Neighbours [7]. Although being one of the simplest, machine learning algorithm, it is still very useful because of it wide range of applicability. The algorithm relies on the previous classification of the neighbors to each training data, classifying according to the majority up to the defined k elements. The training data is presented in a vector space model and all trained examples are vectors in that multidimensional space.

MultiLayer Perceptrons. The "perceptron" is an algorithm, in Machine Learning theory, that is able to classify an input vector using a linear prediction function, which combines a set of computed weights to the vector parameters (Freund, 1999). When it is needed to solve non-linear problems we need more than a layer of perceptrons. Typically, multi-layer perceptrons (MLP), use sigmoide function as an activation function.

Deep Neural Networks. This type of algorithms [8] are based on the concept of pre-training a multi-layered feedforward neural network, one layer at a time, treating each layer as an unsupervised restricted Boltzmann machine, and then using supervised backpropagation for fine-tuning the neural net. Deep learning algorithms are based on an underlying assumption that observed data is generated by the interactions of a multitude of different factors on different levels. Deep learning assumes that

these factors can be organized into multiple different levels of abstraction. Therefore, varying the number of layers and of layer sizes can provide the needed amounts of abstraction [9].

All the algorithms were used through public and open source libraries ("caret" and "h2o"), available for the R programming language.

3.2 The training process

We started the training process of the algorithms by creating a golden standard. This set was created by manually annotating 350 random posts according to the proposed editorial model. In this sample set there were posts without text, which were considered by us to be included in a special category which we labelled as 0 (zero). This manual classification produced a total coverage of the 7 + 1 categories.

As a second step, we retrain the classifiers using a bigger set, of 512 random posts, which were once more, manually classified. The new set, similarly to the first one, had a full coverage of the seven categories plus one, for blank posts, labelled as zero. We illustrate in Figure 1 the comparison of the two manual classifications according to the number of labels produced. In Figure 1 the categorization of the 350 posts is represented by a dashed line and the 512 posts by a solid line. Categories are distributed along the x axis.

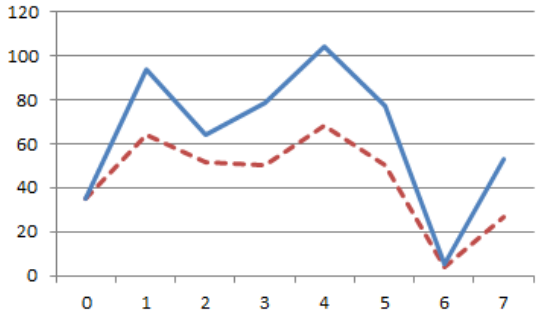

Figure 1. Label count for the 350 and the 512 posts.

It is easy to see that category 6 is still problematic due to its reduced number of posts. Apart from that consideration, we may also observe a tendency for a proportional increase in the number of posts in the remaining categories, when expanding the analysis from 350 to 512 posts.

We then computed the respective accuracy of the automatic classifications using each of the employed algorithms. For this, we used a "confusion matrix" based on the number of false positives, false negatives, true positives, and true negatives. Comparing the accuracy for the two samples, every method had an accuracy increase in the larger set.

However, the global average for this metric, in the new training, was an improvement of only 3%, over the 6 techniques. Therefore, we didn't continue to manually tag bigger sets.

3.3 Enhancing the classification by using more variables

Maintaining our view to classify the posts relying only of the post itself and on the features associated with each post, we augmented List 1 by presenting more information to the classifiers. We refer to information that was already retrieved during post collection, but that was not used in classification:

List 2. Extra fields collected from Facebook posts.

1) From	2) Type	3) Name
4) Date	5) Status	6) Story
7) Hour	8) Link	

Therefore, we had now 17 variables for each post, possibly some of them with no values.

Our approach was to gather all the text in all the 17 fields, together with a "link explosion". We mean by "link explosion" the separation of each term in an URL that is joined to another term by a slash, by punctuation signs, or by the protocol's name.

4 SOCIAL MEDIA STRATEGIES IN THE SECTOR

We define a "social media communication strategy" as the combined, and normalized, intensity in each of the editorial areas previously mentioned. It represents the effort of the HEI in each editorial area of our model, which is normalized by the total number of posts for that HEI. The 43 agents included in the study are presented by codenames (A3 to U2) in Table 2.

In this section we present the intensity of each editorial area per agent, determine the overall editorial sector tendency and reveal the main social media strategies being pursued in the sector.

4.1 Discovering similar social media strategies

Referring to data presented on Table 2, each HEI can be represented by a vector of the form: HEI_x = (#C1$_x$, #C2$_x$, #C3$_x$, #C4$_x$, #C5$_x$, #C6$_x$, #C7$_x$) such that C*i* = normalized number of posts in editorial area *i*, $\forall x \in$ {all the considered HEI}.

Our motivation is that, over a sufficiently large enough period of time (12 months), the number of posts of one agent in each editorial area reflects the HEI social media communication effort for the 7-areas editorial model. Hence, the vector described above is a *signature* for the social media strategy of each HEI, which ultimately can be compared among different HEI.

We compare the different strategies by comparing the 43 vectors in a 7-dimensional vector space model. The comparison is made using a distance metric adaptable to the model. Although several metrics are available for this, it appears to us that the Euclidean distance is perfectly suited to the job because it is straight forward to compute and it does not suffer from the "dimensionality problem". In our case, as we are working on an only 7-dimensional vector space model, dimensionality is not a problem. Along this process, as the "distances" between each signature, and every other, are being computed, we are getting their similarity, i.e., the similarity of the social media strategy between HEI. The rationale behind is: the smaller the distance between HEI, the closest the strategy would be among them. Taking this model, a step further, we group these signatures into clusters of similarity. Hence, having the signatures in a vector space model, we can use a clustering algorithm, which we describe in the next section.

4.2 Grouping similar social media strategies

In this study we used the k-means algorithm [10] to create the clusters of HEI signatures. K-means has the advantage of being easily found implemented in most data-mining tools, and packages. The algorithm works by, iteratively, grouping the vectors into k groups, in such a way that the geometric centroids of these groups move towards a space-position which diminishes the total distance between each group-centroid, and all the vectors in that grouping. The process stops when the centroids do not move from the previous position.

We recall that in our analysis we collected posts from 43 HEIs. During the study we represented each HEI by its codename. We begin by building a table formed by the 43 HEIs and their respective posting, normalized intensity in each area of our model, as listed in Table 2.

Table 2. Intensity in posting, per editorial area.

HEI	Education	Research	Society	Identity	Administration	Relationship	Information
A3	4%	11%	72%	9%	3%	0%	0%
A4	6%	9%	3%	28%	4%	0%	48%
B4	37%	7%	11%	39%	3%	0%	3%
D3	39%	32%	0%	11%	16%	0%	3%
E3	20%	28%	0%	33%	18%	0%	3%
E5	20%	15%	12%	36%	16%	0%	1%
E6	18%	13%	7%	30%	18%	0%	14%
F1	9%	2%	2%	80%	6%	0%	0%
F2	15%	32%	24%	21%	8%	0%	0%
F3	10%	26%	16%	22%	26%	0%	0%
F4	27%	1%	18%	18%	12%	1%	23%
F5	36%	1%	1%	55%	2%	0%	5%
F6	50%	1%	15%	28%	5%	0%	0%
G4	37%	5%	6%	45%	5%	0%	2%
G5	25%	16%	4%	35%	7%	0%	13%
H1	12%	10%	2%	72%	4%	0%	0%
H2	18%	25%	6%	29%	20%	0%	3%

H4	41%	10%	9%	32%	3%	0%	5%
H5	53%	7%	6%	26%	8%	0%	1%
H6	15%	22%	5%	26%	31%	0%	0%
H7	30%	20%	3%	35%	12%	0%	0%
I1	26%	8%	6%	43%	14%	0%	3%
I2	11%	45%	3%	31%	9%	0%	1%
J1	32%	11%	16%	26%	16%	0%	0%
J2	48%	8%	8%	31%	2%	0%	3%
J4	22%	8%	6%	39%	26%	0%	0%
J5	14%	24%	14%	29%	17%	0%	2%
K1	25%	34%	7%	13%	18%	0%	2%
K4	49%	2%	5%	28%	11%	0%	5%
K5	36%	9%	11%	30%	6%	0%	9%
L1	48%	18%	5%	15%	13%	0%	0%
L2	20%	3%	11%	46%	20%	0%	0%
L3	45%	45%	0%	9%	0%	0%	0%
N3	3%	44%	30%	12%	11%	0%	0%
N5	40%	10%	10%	20%	20%	0%	0%
P2	21%	16%	12%	14%	9%	0%	28%
P4	31%	6%	33%	12%	17%	0%	0%
P5	22%	12%	12%	29%	20%	0%	6%
P6	23%	5%	13%	42%	15%	0%	3%
P7	54%	2%	11%	21%	12%	0%	0%
R1	30%	13%	12%	28%	13%	0%	3%
U1	8%	44%	12%	9%	18%	0%	8%
U2	23%	22%	10%	9%	8%	1%	28%
	25%	14%	11%	32%	10%	0%	8%

As we can see, each line of the table may represent a vector in a 7-dimensional space model.

4.3 Finding the optimal number of groupings

While with k-means we do have a 'process' to group HEIs according to their social media editorial similarities into *k* groups, we still don't know the optimal value of that *k*. This value is important because, although the k-means algorithm is guaranteed to converge, it only converges to a local (and not a global) minimum. As *k* can be any integer between 0 and 43, there are still many possibilities to explore. In this situation, we try to determine the value of *k*, based on several clustering runs, and picking the one with the best performance vectors (as centroids). Amongst multiple clustering runs, the low *average-within-centroid* distance yields better clusters, because it indicates how cohesive are the clusters [11].

For this analysis we used the R programming language together with three R-packages: the 'stats' (default); the 'cluster', for accessing the k-means analysis, and; the 'fpc', for k-means with estimating k, and initializations. We started by trying clusterings of 2 up to 25 groups. We measure the success with two methods: the Average Silhouette, and the Calinski Harabasz index [10]. Since the Calinski Harabasz index returned always k-1 as the ideal number of clusters for this set, we prefer to use the Average Silhouette. We experimented 100 runs with maximum of 100 iterations per run. The resulted value was seven.

Finally, we clustered the 43 vectors using k-means, using k=7. As a result, we got a clustering with 2 clusters of a single element, one with 16 elements (the biggest one), followed by a cluster with 11elements. The full dimensions of the clusters are illustrated in Figure 2.

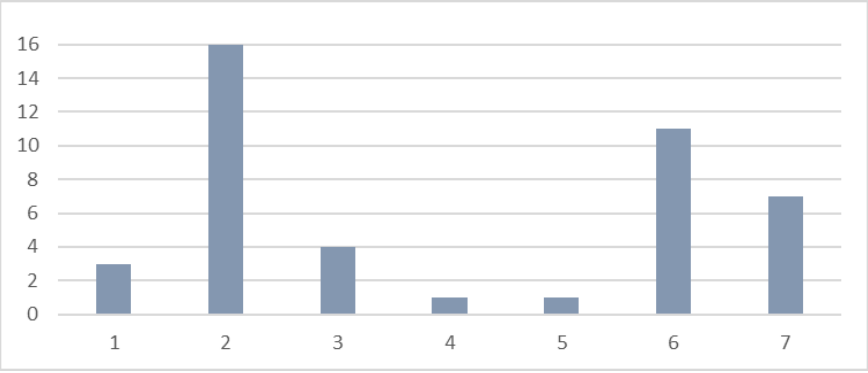

Figure 2. Number of HEI's in each cluster.

Therefore, we can group HEI using this clustering result. We will have 3 HEIs in cluster 1; 16 HEI's in cluster 2; and so on. The k-means algorithm also labels the HEI according to the assigned cluster. Hence, it is possible to distinguish social media strategies by stating that the social media strategy generically adopted by any HEI that belongs to one cluster is sufficiently different from the social media strategy adopted by any HEI that belongs to any other cluster.

4.4 Discovering similar and dissimilar social media strategies

HEI social media strategies can be compared along the seven axis defined in our model. This process has been partially described in [1]. However, using the clustering process, described in the previous sections, we are now equipped with an association property which allows us to:

a) Identify the signature of the group of HEI that belong to the same cluster;

b) Compare similar strategies to understand the result of small strategy variations;

c) Compare different strategies (i.e. from HEI in different clusters) in order to understand the result of bigger variations in the editorial strategy.

Clearly, from Figure 3, we can see that there are centralized strategies (clusters 1, 4 and 6); decentralized strategies (cluster 3), and; hybrid strategies (the remaining clusters: 2, 5 and 7). Based on Figure 2 and Figure 3, we can also state that the most generalized strategy (clusters 2 and 6) tend to favour editorial areas 4 (Identity) and 1 (Education). I.e., the sector tendency favours these two editorial areas.

We can also describe the "average strategy" for each cluster using its centroid (which is computed by the algorithm), and use it to establish comparisons between every centroid, instead of comparing each HEI strategy individually. As a result, it is possible to reduce the amount and complexity of comparisons between agents in order to establish a well sustained benchmarking rational in the sector, in this case, reducing from 903 comparisons (43 HEI) to only 21 (7 clusters).

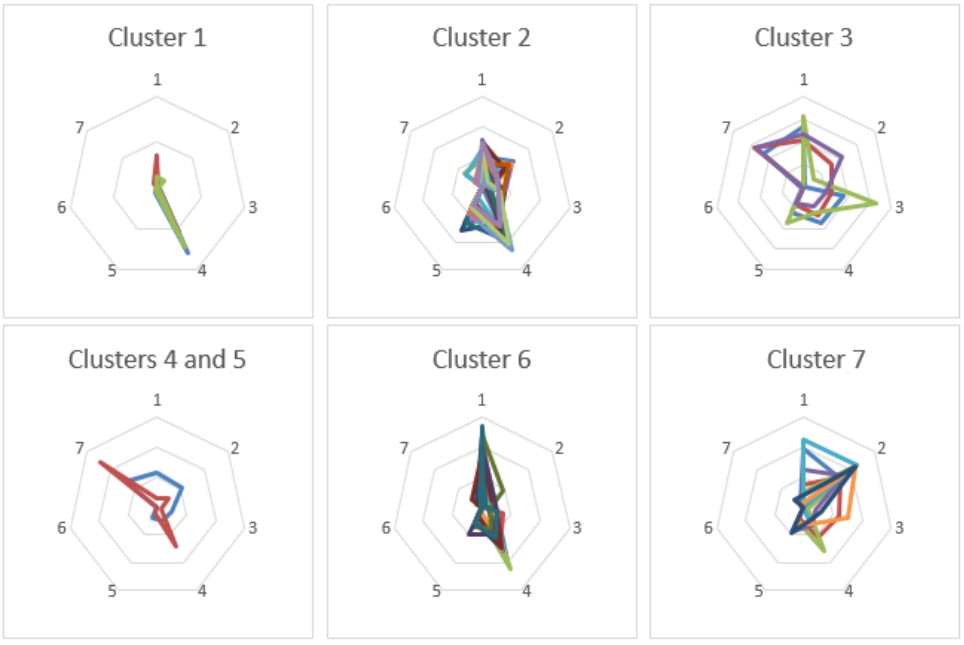

Figure 3. Similar strategies, by cluster, represented in our model.

As a footnote we must also address the situation of editorial line 6 ("Relationship") that had very few posts. Actually, only two of the 43 HEI's had posts in this editorial area. This results in two conclusions: a) contrary to what we may observe in Facebook pages of high-ranked HEI around the globe, in Portugal there is no tradition and/or motivation to post in the area; b) the relationship area does not influence the overall results, but no specific results for that area should also be taken into account.

5 CONCLUSIONS

In this article we presented a base model for the analysis of social media communication by clustering similar editorial strategies. The model is based on a previous categorization of the messages posted in social networks, according to a 7-category model, referenced, and described, in the previous sections, and, in more detail, in previous research [1]. Posts from an entire year were collected from Facebook; a small amount of them was used to build the editorial model, and the remaining were used for training. In our study, the classification accuracy did not increase with samples bigger than 3% of the total data. We used, for categorization, six of the most well-known algorithms to perform this kind of task. We described how, during training, we improved the classifying accuracy by using all the features associated with each post, and how we fine-tuned the categorization parameters. With this enhancement, the resulting accuracy of the method increased from 51% up to 68%. After the ensemble of 6 algorithms was trained, the whole sample of data was run through it.

By clustering the overall social media strategies, we discovered that there are 7 distinct groupings which characterize the whole set of 43 HEI. Clearly, we can find centralized, decentralized and hybrid strategies. There is a tendency for all strategies favour editorial area 4 (Identity), except when the decentralized tendency is strong, in which case area 1 (Education) is more prominent. In all situations editorial area 6 has been marginalized. The clustering process has also as a consequence the reduction of the number of comparisons needed for the benchmarking process, which ultimately allows an easier return of inputs for the design of social media strategies for HEI, in general.

ACKNOWLEDGMENTS

This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project «Reminds/ UTAP-ICDT/EEI-CTP/0022/2014»

REFERENCES

[1] Luciana Oliveira, Álvaro Figueira, Benchmarking Analysis of Social Media Strategies in the Higher Education Sector, Procedia Computer Science, Volume 64, 2015, Pages 779-786, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.08.628.

[2] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". Machine Learning 20 (3): 273.

[3] Koby Crammer and Yoram Singer. 2002. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* 2 (March 2002), 265-292.

[4] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1): 5–32.

[5] Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. Annals of Statistics 28(2): 337–407.

[6] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean (2000); Boosting Algorithms as Gradient Descent, in S. A. Solla, T. K. Leen, and K.-R. Muller, editors, Advances in Neural Information Processing Systems 12, pp. 512-518, MIT Press

[7] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician 46 (3): 175–185.

[8] Ronan Collobert and Jason Weston. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (ICML '08). ACM, New York, NY, USA, 160-167.

[9] Bengio, Y.; Courville, A.; Vincent, P. (2013). "Representation Learning: A Review and New Perspectives". IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8): 1798–1828

[10] Calinski, R. B., and Harabasz, J. (1974) A Dendrite Method for Cluster Analysis, *Communications in Statistics*, 3, 1-27.

[11] Kaufman, L. and Rousseeuw, P.J. (1990). "Finding Groups in Data: An Introduction to Cluster Analysis". Wiley, New York.