Improvement of massive open online courses by text mining of students' emails: a case study

Diego Buenaño-Fernández
Facultad de Ingeniería de Ciencias Agropecuarias
Universidad de Las Américas
Quito
170125
Ecuador
diego.buenano@udla.edu.ec

Sergio Luján-Mora
Departament of Software and Computing Systems
University of Alicante
Alicante
Spain
sergio.lujan@ua.es

W. Villegas-Ch
Facultad de Ingeniería de Ciencias Agropecuarias
Universidad de Las Américas
Quito
170125
Ecuador
william.villegas@udla.edu.ec

ABSTRACT

In recent years, the constant increase in the number of online courses has led to radical changes in the education sector. These new online learning environments present a series of challenges that are difficult to manage using traditional methods. The challenges relate to the level of commitment and motivation shown by students on this type of course. Several articles have been identified from the analysed literature related to the application of text or opinion mining techniques for the analysis of comments made in social networks. In the educational field, articles related to the topic that focus on the analysis of opinion have been identified based on entries included in discussion forums for online courses. Many publications are geared towards solutions in the English language, and the nature of linguistic analysis of this type of study makes it necessary to adapt them for languages other than English. In this paper, we explore the opinion mining through text mining in emails from Massive Open Online Courses (MOOC). The opinion mining expressed in emails is a complex task due to the thematic disparity of emails, their size and the depth of linguistic analysis required. The purpose of this study is to analyse students opinions about their courses, their instructors, and the main tools used on the course. The research focus on the calculation and analysis of the frequency of terms, the analysis of concordances, groupings and n-grams. The case study used in this paper is a MOOC on the topic of web development with more than 40,000 enrolled students.

TEEM 2017, October 18–20, 2017, Cádiz, Spain
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5386-1/17/10…$15.00
https://doi.org/10.1145/3144826.3145393

CCS Concepts

• Computing methodologies~Information extraction • Computing methodologies~Supervised learning • Information systems~Sentiment analysis

Keywords

Opinion mining; Massive Open Online Course; MOOC; Supervised learning; Text mining.

ACM Reference format:

Diego Buenaño-Fernández, Sergio Luján-Mora and W. Villegas-Ch. 2017. Improvement of massive open online courses by text mining of students' emails: a case study. In *Proceedings of 5th International Conference on Technological Ecosystems for Enhancing Multiculturality, Cádiz, Spain, October 2017 (TEEM 2017), 7 pages.*
DOI: https://doi.org/10.1145/3144826.3145393

1: INTRODUCTION

Globalization and the proliferation of Massive Open Online Courses (MOOC) has radically altered the model of education. New technology in this field offers the opportunity to increase the availability of courses to a far greater audience than that provided in the traditional setting. However, the implementation has significant challenges that must be overcome to allow students to take full advantage of them [1]. The flexibility of the platforms in which MOOC operate, and the wealth of learning resources they provide, allows for the inclusion of large numbers of students across a greater geographical base. This interaction between students and systems produces large-scale learning behaviour data and leaves traces of the educational process on systems that are useful for analysis [2].

Given the large volume of emails that are usually generated in a MOOC, it is useful to develop methods that are oriented to the automatic processing of texts with an acceptable reliability [3], and which should become valuable tools to support the educational process.

Interactions between students and MOOC can be explored using text mining techniques, with the aim of improving learning and personalizing the students' experience [4]. This can be met by (1) predicting the level of popularity of the courses; (2) obtaining feedback on the content of the courses so that the tutors can analyse and improve their teaching strategies and (3) obtaining feedback on the platform support so that administrators can improve user experiences.

In this article, it is proposed to apply text mining techniques to analyse opinion from the emails processed in a MOOC. Taking into account that the MOOCs have a particularity that distinguishes them from other type of courses and it is that the target students can become too heterogeneous.

When processing natural language, the analysis of sentiment or mining of opinions refers to the discipline that identifies and classifies fragments of text that contain opinion, either emotional or subjective [5].

In online educational environments it is important to monitor students' attitudes, and a positive attitude towards a course has proven to be an indicator of motivation in students. In contrast, previous studies have shown that expressions of frustration, loneliness, and boredom are associated with less meaningful learning and problematic behaviours [6]. In previous works [6] [7] [8], some objectives related to the application of text mining in MOOC were identified, among the most important are:

- Determine the factors that influence students to complete the course.
- Collect private comments about the course, to help plan a new edition of it.
- Use the sentiment and subjectivity of users to predict commitment to the course.

The supervised classification of web texts has traditionally been used for the categorisation of long texts, typically extracted from forums or blogs of online courses. Tasks such as determining if a document refers to an entity, whether it expresses an opinion or not, or detects what the subject matter is in the message are some of the tasks in which efforts have been invested by the academy [3]. The supervised algorithms, allow evaluation of reliability, and for this reason are the most used in the analysis of sentiment [3].

The rest of this work is organized as follows. In section 2, a description of works related to the proposed topic is given. Section 3 describes the proposed method, which is divided into two major phases: the first phase is oriented to the process of data collection and pre-processing and the second to the definition of the model for the opinion mining. Section 4 describes the results obtained from the process. In addition, there is a discussion about the lessons learned based on the results of our case study. Finally, in section 5, the pertinent conclusions and a general outline for future works related to the topic are presented.

2: RELATED WORK

Most of the works mentioned are linked to the challenges and opportunities associated with the implementation of text mining techniques on unstructured information generated in MOOC. A recognized concept in this area is that of “Affective Computing” which in [9] is defined as “The study and development of Artificial Intelligence oriented to the design of systems and devices that can recognize, interpret and process human emotions, through the analysis of texts”. In this paper, the authors discussed the opportunities and challenges of using opinion mining in e-learning environments, such as an “Affective Computing” application. In [10], a data modelling approach is proposed from a perspective of learning analytics, oriented to the most relevant indicators that guarantee the success of a MOOC. This work proposes an application framework and an intelligent system design scheme that recognizes emotions and mines topics, with the objective of analysing learning and personalized learning in a MOOC. The research proposed in [7], focuses on a study to identify the impact of the analysis of the sentiment expressed in students' opinions about MOOC. The proposed research also identifies the impact of this analysis of students' sentiments and opinions of MOOC.

One issue to consider when talking about MOOC is the level of drop-out that occurs in these contexts, so the work proposed by [11] focuses on the pedagogy used and details of the learning mechanisms that apply in these environments. It also makes recommendations for MOOC developers to improve student engagement.

3: METHOD

The present paper proposes the working methodology illustrated in Figure 1. The emails used for analysis, are those collected by the instructor of the online course on “Web Development” hosted on Google platform “Activate” in collaboration with the University of Alicante (Spain). In the course privacy notice, the student accepts the terms in which it is specified that the data generated will be used to manage their participation in the program.

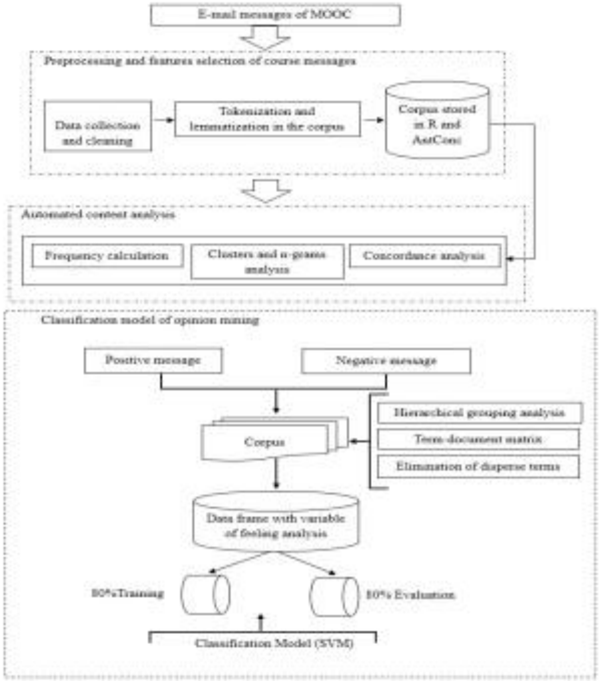

E-mail messages of MOOC
Preprocessing and features selection of course messages
Data collection and cleaning
Tokenization and lemmatization in the corpus
Corpus stored in R and AntConc
Automated content analysis
Frequency calculation
Clusters and n-grams analysis
Concordance analysis
Classification model of opinion mining
Positive message
Negative message
Hierarchical grouping analysis
Corpus
Term-document matrix
Elimination of disperse terms
Data frame with variable of feeling analysis
80%Training
80% Evaluation
Classification Model (SVM)

Figure 1: Methodology for the process of opinion mining in emails in a MOOC.

To perform the bulk download of files from the instructor's email account, the tool "Save Emails and Attachments" was used. This is a backup and archiving add-on tool in Gmail that downloads email messages and attachments from Gmail to Google Drive. Using this tool, it was possible to download all emails corresponding to the period 2015-2017, in PDF format. In this phase, it was necessary to manually delete some irrelevant information, as well as repeated samples, emails that did not contain any text, or emails that were not oriented to the following requirements: (1) to obtain feedback on the content of the courses so that the tutors can analyse and improve their teaching strategies and (2) to obtain feedback on platform support for administrators to enhance users' experiences.

Once an important amount of unstructured textual data (data that does not have an identifiable internal structure) was collected and following the proposed methodology, the next step was the pre-processing of such data. For this we used the tools AntConc and R. AntConc is a free software package for linguistic analysis of texts, available for Windows, MacOS and Linux operating systems that allows you to work with text or PDF files, whereas, R is a free software environment for the management of computer statistics that compiles data and works on a wide variety of UNIX, Windows and MacOS platforms. A corpus was loaded from the text files worked in the previous phase in both computer tools.

At this stage, it was important to verify the character encoding system where the text documents were stored; for the present study, we used UTF-8 encoding. For the tokenisation phase the elimination of numbers, punctuation marks, blanks and empty words (stopwords) was performed. In the case of R, several functions included in the ReadMe library (rm) were used. A final step in the pre-processing was the corpus's lemmatization. In this process, each word of the corpus (token) was related to its canonical form or motto, this allowed us to treat a greater number of words of homogeneous form.

The first data analysis performed after the pre-processing stage was automated content analysis, understood as a statistical-analytical exercise whose objective is to obtain numerical information from a set of textual data [12]. The AntConc tool was used to calculate frequencies, clusters and n-grams, and finally the concordance analysis.

According to the proposed methodology, the next phase is the structuring of the classification model for the analysis of opinion. For this purpose, the process of manual classification of the collected information was initially performed, so that in the first instance, the computer learns the different rules established to classify emails according to the polarity of sentiment. These rules include words or word combinations associated with positive or negative messages.

This article presents a classification model used in machine learning called Support Vector Machine (SVM). The purpose of this model is to identify a hyperplane of (n-1) dimensions, which divides two classes of entities located in a multidimensional space as shown in Figure 2. Each point in the graph corresponds to a text message and each message is placed in the multidimensional space according to the words it has [13]. To test the model, a set of training messages and a set of evaluation data were defined. Based on information from the revised literature [12], the database was divided as follows: 80% of data for training and 20% of data for evaluation of the model. At the training stage, the data consists of pairs of predictors and target values. For each predictor value a label with a target value is used. If the algorithm can predict a categorical value for a target attribute, it is called the classification function.

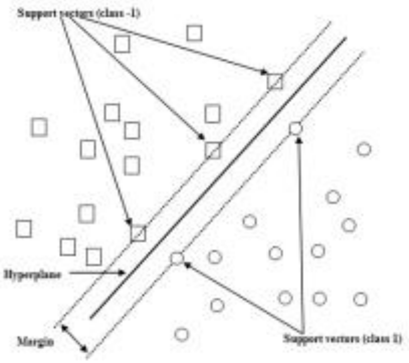


Figure 2: Representation of the hyperplane with the SVM.

There are several methods to determine the effectiveness of the SVM-based model; however, those based on accuracy and recall are the most common in this field. The accuracy method is based on the number of evaluations of a given class that are correctly classified over the total number of evaluations of that class. The recall method is based on the number of total revisions that are correctly classified from a given class [12].

To represent the frequency of occurrence of linguistic units in a text, a matrix Mmxn was constructed. In the reviewed literature, there are generally three types of matrices, which analyse similarities of documents: term-document matrices, word-context matrices and pair-pattern matrices [13]. In the present work and prior to executing the algorithm, it was necessary to generate the term-document matrix, using the function of R DocumentTermMatrix(). Based on this matrix, it was possible to identify the occurrences of the indexed terms in the document collection. In the matrix, each row represents a document and each column a term, so that the count becomes simple, however with this method the order in which the terms appear is ignored, that is to say that the linguistic structure of the text is unknown [13]. Generally, most of the elements of this matrix are 0, so it is a scattered matrix.

Next, a hierarchical grouping analysis was performed to identify groups of related words, based on the distances between them. The dispersed terms of the matrix were removed to preserve only the most frequent words and thus obtain more interpretable results from the cluster. This process was performed using the RemoveSparseTerm () function of R. The sparse argument can assume values between 0 and 1, and represents the dispersion of the words that we want to keep. If it is set very high (about 1), many words are conserved, almost all, because it is indicated that it is necessary to preserve terms even though they are widely dispersed. The opposite happens if this value is set very low (about 0). The value to be chosen will depend on the type of documents to be evaluated, for this work some tests were carried out to find the balance between dispersion and number of terms, the chosen value was 0.95. With this procedure, we make sure that those words mentioned least in the emails are removed.

Once this was done, it was necessary to convert the data into an object on which subsequent operations can be performed, i.e. an array or a dataframe. The chosen classification algorithm was run on this object.

4: RESULTS AND DISCUSSION

The data used in this study was obtained from an initial set of 1400 emails received by the course instructor. The data source was manually debugged by converting the PDF files into text format containing only the information corresponding to the body of the email. In addition, emails that were not related to the course and some files that did not contain any information were eliminated. Another group of emails that were not considered were those that only had programming code and no important opinions to consider. About 30% of files were discarded leaving a final total of 950 emails to analyse. For this group of files, it was verified that the type of text encoding was UTF-8.

The first type of analysis was the frequency analysis, which involves counting the words of a corpus and displaying them in an ordered list from the most frequent to the least frequent. To perform this analysis, we used the AntConc tool and the numerical results of this can be seen in Table 1. The number of tokens represents the set of characters separated by a blank space, while the number of word types represents the repeated tokens. This shows there are 7890 different shapes throughout the text. The relationship between these two figures is called the ratio words/forms and is usually calculated according to the following formula:

$$ratio\ types/tokens = \frac{\text{Number of word yTypes}}{\text{Number of tokens}}$$

Table 1: Initial description of frequency analysis

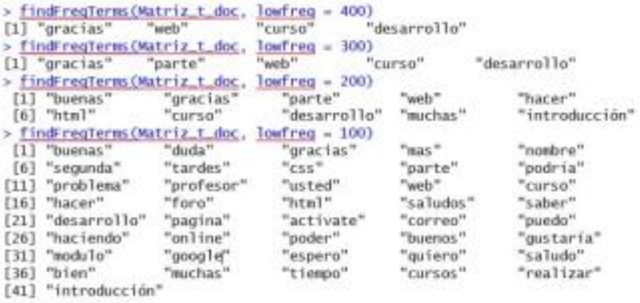
Description	Results
Number of documents	950
Number of tokens	82,174
Number of word types	7,890
Ratio types/ tokens	0.096

From these global statistics on the composition of the corpus we can calculate some indicators, and although they are not that significant in themselves, they must be considered, since they can influence the outcome of other mathematical operations that are dependent on the size or composition of the corpus. Before making this calculation and to obtain better results it was necessary to eliminate all the words that lack lexical meaning, by eliminating empty words or stopwords. In this group, all words in closed grammatical categories are considered. For this process, we worked with a text file containing a list of approximately 620 empty words, the same that were loaded in the AntConc tool in parallel with the generated corpus. The data that was obtained is shown in Table 2. It is observed that the number of tokens decreased by 58% with respect to the first report and that the result of ratio types/tokens is closer to 1 than in the previous calculation. This indicates that the lexical richness has been improved after pre-processing.

Table 2: Description of frequency analysis data after tokenisation of the corpus

Description	Results
Number of documents	950
Number of tokens	34,763
Number of word types	7,387
Ratio types/tokens	0.212

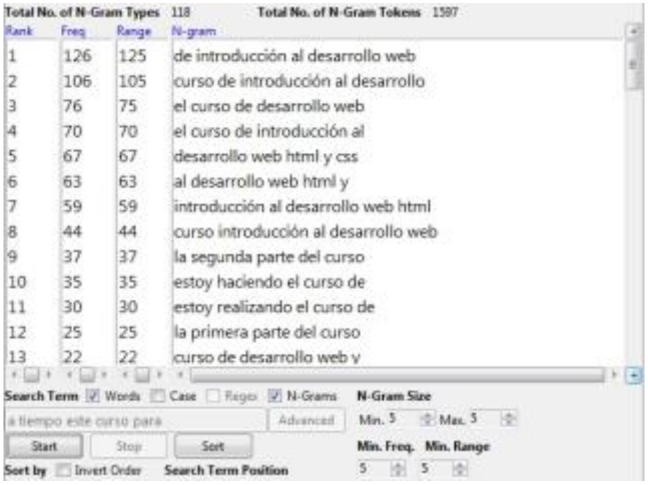
After performing the tokenisation and stemming of the corpus, we analysed the words that appear most frequently. Figure 3 shows a summary of frequencies taken from the R tool after applying the findFreqTerms () function. In this list of words are terms directly related to the name and subject of the course itself, such as: web, course, introduction and development. However, if it look at the words at the bottom of the list in Figure 3, allows can find words with greater semantic meaning for the analysis of opinion, for example, thanks, good, greetings, problem and doubt. Whereas it is true, these terms alone do not say much, they do give us clues to develop the study. For this reason, the next task was the semantic analysis of how words are related in the corpus, that is, how often do two or more words appear together. These sets of words that can appear in a text, in a certain consecutive order, can be placed in an analysis of n-grams, where n is the number of words [14].



```
> findFreqTerms(Matriz_t_doc, lowfreq = 400)
[1] "gracias"    "web"        "curso"      "desarrollo"
> findFreqTerms(Matriz_t_doc, lowfreq = 300)
[1] "gracias"    "parte"      "web"        "curso"      "desarrollo"
> findFreqTerms(Matriz_t_doc, lowfreq = 200)
 [1] "buenas"       "gracias"      "parte"        "web"          "hacer"
 [6] "html"         "curso"        "desarrollo"   "muchas"       "introducción"
> findFreqTerms(Matriz_t_doc, lowfreq = 100)
 [1] "buenas"       "duda"         "gracias"      "mas"          "nombre"
 [6] "segunda"      "tardes"       "css"          "parte"        "podría"
[11] "problema"     "profesor"     "usted"        "web"          "curso"
[16] "hacer"        "foro"         "html"         "saludos"      "saber"
[21] "desarrollo"   "pagina"       "activate"     "correo"       "puedo"
[26] "haciendo"     "online"       "poder"        "buenos"       "gustaría"
[31] "modulo"       "google"       "espero"       "quiero"       "saludo"
[36] "bien"         "muchas"       "tiempo"       "cursos"       "realizar"
[41] "introducción"
```

Figure 3: Frequency of terms in the analysed corpus.

To calculate n-grams it is not necessary to eliminate the stopwords, since this is a way to capture some structural information on how words are related without the need to apply syntactic analysis techniques [15]. Figure 4 shows the ranking of the first n-grams of size 5, which confirms the above in the analysis of frequencies per word, i.e., n-grams with a higher frequency of occurrence are related to the description of the course.



Total No. of N-Gram Types 118 Total No. of N-Gram Tokens 1597

Rank	Freq	Range	N-gram
1	126	125	de introducción al desarrollo web
2	106	105	curso de introducción al desarrollo
3	76	75	el curso de desarrollo web
4	70	70	el curso de introducción al
5	67	67	desarrollo web html y css
6	63	63	al desarrollo web html y
7	59	59	introducción al desarrollo web html
8	44	44	curso introducción al desarrollo web
9	37	37	la segunda parte del curso
10	35	35	estoy haciendo el curso de
11	30	30	estoy realizando el curso de
12	25	25	la primera parte del curso
13	22	22	curso de desarrollo web y

Search Term ☑ Words □ Case □ Regex ☑ N-Grams

N-Gram Size Min. 5 Max. 5

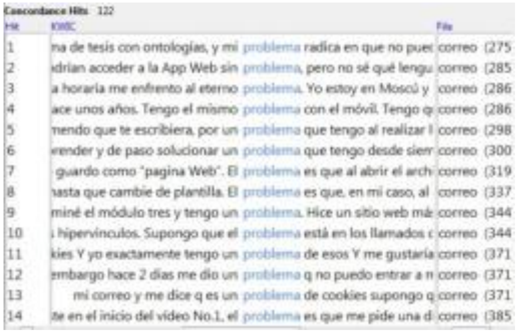
a tiempo este curso para

Min. Freq. 5 Min. Range 5

Sort by □ Invert Order Search Term Position

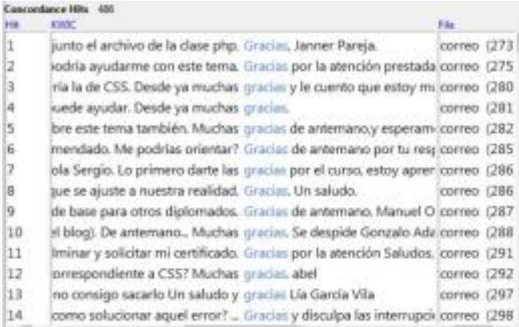
Figure 4: Frequency of n-grams in the analysed corpus.

Finally, in this phase, a concordance analysis was performed using the AntConc tool, which makes the analysis of linguistic patterns that appear with a certain frequency in the corpus possible and reflects their real behaviour within the context. The most common match type is keyword in context. The concordances are instruments considered indispensable in the study of the lexical patterns and for this reason, it is key to the investigation of a corpus [16]. In Figure 3, some words were presented that after initial frequency analysis, had been considered for the analysis of opinion. In Figure 5 the concordance analysis table generated in the tool for two words can be observed. This analysis was used as a complement to the frequency analysis performed initially. Programs that show the matches of a text are called Key Word In Context (KWIC).



Concordance Hits 122

Hit	KWIC	File
1	na de tesis con ontologías, y mi problema radica en que no pued	correo (275
2	drian acceder a la App Web sin problema, pero no sé qué lengu	correo (285
3	a horaria me enfrento al eterno problema. Yo estoy en Moscú y	correo (286
4	ace unos años. Tengo el mismo problema con el móvil. Tengo q	correo (286
5	nendo que te escribiera, por un problema que tengo al realizar l	correo (298
6	render y de paso solucionar un problema que tengo desde siem	correo (300
7	guardo como "pagina Web". El problema es que al abrir el arch	correo (319
8	hasta que cambie de plantilla. El problema es que, en mi caso, al	correo (337
9	miné el módulo tres y tengo un problema. Hice un sitio web más	correo (344
10	hipervínculos. Supongo que el problema está en los llamados c	correo (344
11	kies Y yo exactamente tengo un problema de esos Y me gustaría	correo (371
12	embargo hace 2 dias me dio un problema q no puedo entrar a m	correo (371
13	mi correo y me dice q es un problema de cookies supongo q	correo (371
14	te en el inicio del video No.1, el problema es que me pide una d	correo (385

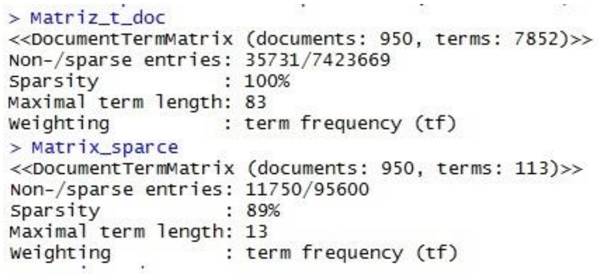


Concordance Hits 486

Hit	KWIC	File
1	junto el archivo de la clase php. Gracias, Janner Pareja.	correo (273
2	podria ayudarme con este tema. Gracias por la atención prestada	correo (275
3	ría la de CSS. Desde ya muchas gracias y le cuento que estoy mı	correo (280
4	uede ayudar. Desde ya muchas gracias.	correo (281
5	bre este tema también. Muchas gracias de antemano,y esperam	correo (282
6	mendado. Me podrías orientar? Gracias de antemano por tu res|	correo (285
7	ola Sergio. Lo primero darte las gracias por el curso, estoy aprer	correo (286
8	que se ajuste a nuestra realidad. Gracias. Un saludo.	correo (286
9	de base para otros diplomados. Gracias de antemano. Manuel O	correo (287
10	el blog). De antemano... Muchas gracias. Se despide Gonzalo Ada	correo (288
11	lminar y solicitar mi certificado. Gracias por la atención Saludos.	correo (291
12	orrespondiente a CSS? Muchas gracias. abel	correo (292
13	no consigo sacarlo Un saludo y gracias Lía García Vila	correo (297
14	como solucionar aquel error? ... Gracias y disculpa las interrupci	correo (298

Figure 5: Concordance analysis of two words.

Finally, according to the proposed methodology, the tests of the learning model applied to this data set were carried out. The starting point of this phase was the generation of the matrix term document. In Figure 6, we can observe the statistics of the original matrix, as well as the data after the application of the RemoveSparseTerms () function in R. The number of documents is maintained (950) however the number of words is significantly reduced (7852 to 113). Likewise, the percentage of dispersion of terms decreases from 100% to 89%.



```
> Matriz_t_doc
<<DocumentTermMatrix (documents: 950, terms: 7852)>>
Non-/sparse entries: 35731/7423669
Sparsity           : 100%
Maximal term length: 83
Weighting          : term frequency (tf)
> Matrix_sparce
<<DocumentTermMatrix (documents: 950, terms: 113)>>
Non-/sparse entries: 11750/95600
Sparsity           : 89%
Maximal term length: 13
Weighting          : term frequency (tf)
```

Figure 6: Statistics matrix term document

To apply the proposed algorithm, it was necessary to generate a dataframe with the information from the matrix term document. The next task was to randomly define a training set and an evaluation set. For this and according to the revised literature [17], 80% of the total documents were used for training (760), and 20% for the evaluation phase (190). In R this process was done using the library caTools. Finally, the code was written in R for the application of the algorithm SVM and to run the model. To demonstrate the results in R, the function confusionMatrix () was used, the rows show the prediction of the model and the columns shows the actual values of the classification. A key result that shows the function is the degree of accuracy of the executed model, and in our case, it was 75% which means that in 75% of cases the model classified documents correctly.

5: CONCLUSIONS

The analysis of opinion expressed by students in discussion forums, blogs or emails on MOOC, provides useful information for the analysis of such courses. However, very long and dispersed messages complicate the development of methods that can automatically evaluate the sentiment transmitted in the messages. The development of these processes and their application to messages in Spanish in educational environments is the object of growing interest by the researchers.

The present study focused on two areas. Firstly, the analysis of frequency terms, which provided some interesting results as a starting point to improve the application of techniques for the linguistic and semantic analysis of email messages. Secondly, the proposal for a predictive model of the polarity of sentiment for the compiled documents. The degree of precision obtained by the model challenges us to find different algorithms or to make a more efficient preliminary prediction to improve the indicator.

In practice, the analysis of opinion should be used with caution, especially when the texts are very noisy (difficult to categorise) and thematically dispersed. A suggestion for future work is to integrate the texts with additional sources of information external to a course, as more data on students' behaviour and activities could provide greater accuracy in prediction and personalisation. In addition, it is suggested to integrate text analysis on forums and other types of texts produced by the learners in MOOCs, like social networks interactions, messages in the platform, etc.

ACKNOWLEDGMENTS

Special thanks to the UniMOOC group of the University of Alicante, as their support allowed us to obtain the unstructured data which was used in the present study.

REFERENCES

[1] Colin Allison, Alan Miller, Iain Oliver, Rosa Michaelson, Tiropanis Thanassis. 2012. The web in education, Computer Networks. *Computer Networks*. 56(18), 3811–3824. DOI=https://doi.org/10.1016/j.comnet.2012.09.017

[2] Liu Zhi, Zhang Wenjing, Sun Jianwen, Cheng Hercy and Sanya Xian. 2016. Emotion and associated topic detection for course comments in a MOOC platform. In *Proceedings of International Conference on Educational Innovation through Technology (EITT)*. IEEE, Tainan, Taiwan, 15-19. DOI= https://doi.org/10.1109/EITT.2016.11

[3] Tomás Baviera. 2016. Técnicas para el análisis del sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Dígitos: Revista de Comunicación Digital*, 3(2), 33-50

[4] Safwan Shatnawi, Mohamad Medhat G. and Mihaela Cocea. 2014. Text stream mining for Massive Open Online Courses: review and perspectives. Systems Science & Control Engineering, 2(1), 664-676. DOI= http://doi.org/10.1080/21642583.2014.970732

[5] Roberto Hernández and Xiaoou Li. 2015. Sentiment analysis of texts in spanish based on semantic approaches with linguistic rules. In Proceedings of *XXX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. SPLM, Girona, Spain, 1-7.

[6] Ryan Baker, Sydney D'Mello, Mercedes Rodrigo and Arthur Graesser. 2010. Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environment. International Journal of Human-Computer Studies, 68(4), 223-241. DOI= https://doi.org/10.1016/j.ijhcs.2009.12.003

[7] Miaomiao Wen, Diyi Yang, and Carolyn Penstein. 2014. Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. In *Proceedings of the 7th*

International Conference on Educational Data Mining. EDM, London, 130-137.

[8] Panagiotis Adamopoulos. 2013. What makes a great mooc? An interdisciplinary analysis of student retention in online courses. In *Proceedings of the 34th International Conference on Information Systems*. ICIS, Milan,1-21 .

[9] Hogefei Lin, Fengming Pan, Yuxuan Wang, Shaoua Lv, and Shichang Sun. 2010. Affective Computing in E-learning. E-learning InTech Publishing, 117-128. DOI= https://doi.org/10.1016/j.phpro.2012.02.278

[10] Alejandro Maté, Elisa De Gregorio, José Cámara, JuanTrujillo and Sergio Luján-Mora. 2016. The improvement of analytics in massive open online courses by applying data mining techniques. Expert Systems Wiley Publishing, 33(4), 374-382. DOI= http://dx.doi.org/10.1111/exsy.12119

[11] Jana Sinclair and Sara Kalvala. 2016. Student engagement in massive open online courses. Int. J. Learning Technology, 11(3), 218-237. DOI= http://doi.org/10.1504/IJLT.2016.079035

[12] Carlos Arcila-Calderón, Eduar Barbosa-Caro, Francisco Cabezuelo-Lorenzo. 2016. Big data techniques: Large-scale text analysis for scientific and journalistic research. El professional de la información, 25(4), 623-631. DOI= https://doi.org/10.3145/epi.2016.jul.12

[13] Alaa El-Halees. 2011. Mining Opinions in User-Generated Contents to Improve Course Evaluation. In *Proceedings of International Conference on Software Engineering and Computer Systems*. Springer-Verlag, Berlin, 107-115. DOI=https://doi.org/10.1007/978-3-642-22191-0_9

[14] Michael Heilman, Kevyn Collins-Thompson and Maxine Eskenazi. 2008. An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In Proceedings of Third Workshop on Innovative Use of NLP for Building Educational Application. ACL Home Association for Computational Linguistics, Ohio, USA, 71-79.

[15] David Carmel, Avihai Mejer, Yuval Pinter and Idan Szpektor. 2014. Improving Term Weighting for Community Question Answering Search Using Syntactic Analysis. In *Proceedings of Proceeding of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, Shanghai, China, 351-360. DOI=http://doi.org/10.1145/2661829.2661901

[16] Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance. IEEE Transactions on visualization and computer graphics, 14(6), 1221-1228. DOI=http://doi.org/10.1109/TVCG.2008.172

[17] Ivor Tsang, James Kwok, and Pak-Ming Cheung. 2005. Core Vector Machines: Fast SVM Training on Very Large Data Sets. Journal of Machine Learning Research, 363-392.