DE GRUYTER OPEN

**European Journal of Open, Distance and e-Learning**
Vol. 18 / No. 2 – 2015
DOI: 10.1515/eurodl-2015-0014

# A LEARNING ANALYTICS METHODOLOGY FOR DETECTING SENTIMENT IN STUDENT FORA: A CASE STUDY IN DISTANCE EDUCATION

*Vasileios Kagklis [kagklis@gmail.com], Hellenic Open University, Anthi Karatrantou [a.karatrantou@eap.gr], University of Patras, Maria Tantoula [aria_tant@hotmail.com], Hellenic Open University, Chris T. Panagiotakopoulos [cpanag@upatras.gr], University of Patras, Vassilios S. Verykios [verykios@eap.gr], Hellenic Open University, Greece*

## Abstract

Online fora have become not only one of the most popular communication tools in e-learning environments, but also one of the key factors of the learning process, especially in distance learning, as they can provide to the students involved, motivation for collaboration in order to achieve a common goal. The purpose of this study is to analyse data related to the participation of postgraduate students in the online forum of their course at the Hellenic Open University. The content of the messages posted is analysed by using text mining techniques, while the network through which the students interact is processed through social network analysis techniques. Furthermore, sentiment analysis and opinion mining is applied on the same dataset. Our aim is to study students' attitude towards the course and its features, as well as to model their sentiment behaviour over time, and finally to detect if and how this affected their overall performance. The combined knowledge attained from the aforementioned techniques can provide tutors with practical and valuable information for the structure and the content of the students' exchanged messages, the patterns of interaction among them, the trend of sentiment polarity during the course, so as to improve the educational process.

## *Abstract in Greek*

Οι διαδικτυακές ομάδες συζήτησης (Forum) είναι ένα από τα δημοφιλέστερα εργαλεία επικοινωνίας σε μαθησιακά περιβάλλοντα. Αποτελούν έναν από τους βασικούς παράγοντες της μαθησιακής διαδικασίας, ειδικά στην εκπαίδευση από απόσταση, καθώς παρέχουν στους συμμετέχοντες εκπαιδευόμενους τα απαραίτητα κίνητρα συνεργασίας για την επίτευξη κοινών μαθησιακών στόχων. Σκοπός αυτής της εργασίας είναι η ανάλυση δεδομένων από τη συμμετοχή μεταπτυχιακών φοιτητών του Ελληνικού Ανοικτού Πανεπιστημίου σε διαδικτυακή ομάδα συζήτησης στα πλαίσια μαθημάτων του Προγράμματος Σπουδών τους. Το περιεχόμενο των δημοσιευμένων μηνυμάτων των συμμετεχόντων αναλύθηκε χρησιμοποιώντας μεθόδους και τεχνικές εξόρυξης κειμένου, ενώ η αλληλεπίδραση των φοιτητών μελετήθηκε μέσω τεχνικών ανάλυσης κοινωνικών δικτύων. Επιπλέον, στο ίδιο σύνολο δεδομένων εφαρμόστηκαν τεχνικές ανάλυσης συναισθήματος και εξόρυξης γνώμης. Στόχος μας είναι η μελέτη της συμπεριφοράς των φοιτητών απέναντι στο μάθημα και τα χαρακτηριστικά του, καθώς και η μοντελοποίηση της συναισθηματικής τους συμπεριφοράς με την πάροδο του χρόνου και, τέλος, η ανίχνευση του αν και κατά πόσο αυτή επηρεάζει την συνολική τους επίδοση στις σπουδές τους. Ο συνδυασμός των γνώσεων που προκύπτουν από τις προαναφερθείσες τεχνικές μπορεί να προσφέρει στους εκπαιδευτές πολύτιμες και πρακτικές πληροφορίες για τη δομή και το περιεχόμενο των μηνυμάτων που αντάλλαξαν οι φοιτητές, για τον τρόπο αλληλεπίδρασης μεταξύ των φοιτητών, για την τάση της συναισθηματικής πολικότητας τους κατά τη διάρκεια των σπουδών τους, έτσι ώστε να βελτιωθεί η εκπαιδευτική διαδικασία.

**Keywords:** Social Network Analysis, Educational Data Mining, Learning Analytics, Sentiment analysis

## Introduction

According to Hülsmann (2009), the *Achilles heel* of distance education is the weak and limited interaction among students or between tutors and students, due to the distance and the lack of face-to-face communication. For over a decade now, all the educators involved in distance education, in order to overcome the above obstacles, they have shown an increased interest in using several information and communication technological tools over the Internet, as well as new strategies and terms about e-learning (Anderson & Garrison, 1998; Anderson, 2004; Anderson, 2007).

In such a framework, the use of social media and microblogging, virtual worlds, chat rooms, online discussion fora, blogs and other web-based tools, synchronous or asynchronous, are emerging as useful and supportive tools to the educational process (Groves & O'Donoghue, 2009; Carsten et al., 2010). These tools can increase students' motivation and participation by making them able to determine the content of their discussions and define their educational needs (Choi et al., 2005; Bradley & McDonald, 2011).

As communication serves a major role in learning, and especially in distance learning, the online discussion fora provide motivation for collaboration. According to (Brindley et al., 2009), collaborative learning appears to increase the sense of community, which is related to the learner's satisfaction and retention. These factors are important for students in distance education not only for their cognitive improvement, but also because these factors deter them from dropping out. Moreover, useful information could be extracted by analysing the content of the students' exchanged text messages (posts), through their participation and reflection, in order to derive the specific interests of a student or to figure out certain details that are useful for the planning of a personalized help (Abel et al., 2010). These data could help tutors to adjust the educational process so as to fit more in their audience and to become more effective.

Learning analytics (Siemens, 2010) is the research field that combines techniques such as educational data mining, social networks analysis, sentiment analysis and educational data conceptual modelling, in order to gather information and gain knowledge about the function and the results of an educational course at different levels. More specifically, through the learning analytics an early detection of those students who need special help or are in risk of failure may occur or some learning tools can be recommended, appropriate to the students' needs, or further suggestions for any decisions needed to be taken may be made.

Sentiment analysis (Nasukawa & Yi, 2003), also known as opinion mining, refers to the use of text mining and natural language processing (Manning & Schütze, 1999) in order to determine the attitude of a person towards a topic, or determine the overall contextual polarity of a document. The attitude may be the person's judgment or evaluation, the emotional state of the author when writing, or the intended emotional communication. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or corpus level, i.e., identifying whether the expressed opinion in a document, a sentence or a corpus is positive, negative, or neutral. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer services.

When working towards improving the process of distance learning and towards the experience of students participating in educational discussion fora, it is important to know the opinion of the students about their course. Sentiment analysis can reveal useful information about students' attitude and can help in achieving a better understanding of their behaviour during the learning process. Written messages, in the form of posts, in the discussion fora is the most widely used way of communication in distance learning programs. As each individual has its own way of expression, part of our study focuses on identifying the emotional status of the participants and on discovering patterns that can help in modelling the participants' behaviour.

In this paper, an analysis of real data related to the participation of postgraduate students in the online forum of their course at the Hellenic Open University (HOU) is presented. Text mining techniques, social network analysis techniques and sentiment analysis are employed on the same dataset aiming to study the structure and the content of the students' exchanged messages, the patterns of interaction among them, the students' attitude towards the course and its features, as well as to model their sentiment behaviour over time and finally to present how this affected their overall performance.

The rest of the paper is organized as follows. "Related work" section gives a brief background insight of some previous related works. "Studies in Hellenic Open University" section provides details about the HOU, the course, and the module the dataset is originated from. The next section presents the methodology followed for the analysis of the data. "Methodology" section demonstrates the results of the analysis. Finally, last section concludes this work and presents some ideas for future work.

## Related work

Learning analytics refers to the interpretation of a wide range of students' data in order to assess academic progress, predict future performance, and detect potential issues during their studies. Data are collected from explicit student actions, such as completing assignments and taking exams, and also from implicit student actions, such as online social interactions, extracurricular activities, posts on discussion forums, and other activities (Johnson et al., 2011).The methodology of learning analytics includes, (a) the gathering of the data, derived from the students and the learning environment in which they participate, and (b) the intelligent analysis of this data that leads to conclusions regarding the degree of the students' participation in the fora and how this participation affects their learning. The main goal of Learning Analytics methodology is to understand and optimize the learning processes and also to improve the environments in which these processes occur (Siemens & Baker, 2012). Analytics have been used for Prediction purposes, Personalization & adaptation, Intervention purposes, Information visualization.

Educational Data Mining (EDM) is an emerging research area that addresses the development of methods that explore data originating from an educational context. EDM uses different methods from statistics, machine learning and data mining, in order to analyse the data collected during the teaching and learning process. Students' learning data are being explored to develop predictive models and to discover new knowledge based on students' usage data. Such procedures help the tutors to evaluate educational systems, to potentially improve aspects of the quality of education and to lay the groundwork for a more effective learning process (Berland, Baker & Blikstein, 2014). In their research, Baker and Yacef (2009) suggest four goals of EDM: predicting students' future learning behaviour, discovering or improving domain models, studying the effects of educational support that can be achieved through learning systems, advancing scientific knowledge about learning and learners by building and incorporating student models.

Romero, Ventura and Garcia (2008) present a specific application of data mining in learning management systems and a case study tutorial with the Moodle system. In their study they describe how different data mining techniques can be used in order to improve the course and the students' learning. They apply the most general and well known data mining techniques, as well as two other specific data mining methods, the outlier analysis for data cleansing, spotting emerging trends and recognizing unusually good or bad performers, and the social network analysis for the analysis of the structure and context of online educational communities.

The analysis of Social Networks (Social Network Analysis -SNA) views social relationships in terms of network theory, consisting of nodes (representing individual actors within the network) and ties (which represents relationships between the individuals such as friendship, kinship, organizations, sexual relationships, etc.) (D'Andrea et al., 2009; Carlos, 2011).These networks are often depicted in a social diagram, where the nodes represent objects and the ties express relations.

An interesting broad overview of recent studies on social network analysis techniques is presented in Takaffoli and Zaïane (2012). In their study, the authors describe existing works and approaches on applying social network techniques for assessing the participation of the students in the online courses. They present their social network analysis toolbox (Meerkat-Ed) for visualizing, monitoring, and evaluating the students' participation in a discussion forum. In particular, the visualization comprises the depiction of community detection among students on forum, keywords that indicate the topics addressed in the discussion and the relations between them, as well as, the centrality of students in the network. In addition, they present the implementation of Meerkat-Ed on their own case study data. Following this line of research, in our study, both text mining and social network analysis techniques are used, so as to assess the learning process of students' participation in the online discussion.

Turney (2002) and Pang, Lee and Vaithyanathan (2002) are among the first to use sentiment analysis, combined with machine learning algorithms. In Turney (2002) a simple unsupervised learning algorithm is used for review-classification in two groups; recommended or not recommended. The classification of a review is predicted by the average semantic orientation of the phrases in the review. In Pang, Lee and Vaithyanathan (2002) the problem of classifying documents by the overall sentiment is considered. The authors use movie reviews as data, and they discover that standard machine learning techniques can outperform human-produced baselines.

The currently existing approaches of sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques (Cambria et al., 2013). Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored (Ortony, Clore & Collins, 1988). Lexical affinity not only detects obvious affect words, but also assigns arbitrary words a probable *affinity* to particular emotions (Stevenson, Mikels, & James, 2007). Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, *bag of words*, etc. More advanced methods try to detect the holder of a sentiment and the target (Kim & Hovy, 2006). Lastly, concept-level approaches leverage on elements from knowledge representation such as ontologies and semantic networks and, thus, are also able to detect semantics that are expressed in a subtle manner (Cambria & Hussain, 2012).

Ortigosa et al. (2014) present a method for sentiment analysis. Additionally, the authors point out the importance of sentiment analysis and its results in distance learning, for developing educational systems and personalizing the learning process. The sentiment polarity is calculated by applying a combination of a dictionary-based analysis and machine learning techniques on the messages of the users, giving high-accuracy results. The authors model the sentiment polarity in order to detect any variance in the sentiment trend of the users.

A recent study on massive open online courses (MOOCs) is that of Wen et al. (2014). The authors apply sentiment analysis on students' posts, in order to identify students' opinion for specific features of the course, and to evaluate if there is a connection between the sentiments and the students' drop-out rate.

## Studies in Hellenic Open University

### *The Hellenic Open University (HOU)*

HOU was officially established in 1997 and is the only University in Greece offering exclusively distance education courses. It consists of four (4) Schools: Humanities, Social Sciences, Science and Technology, and Applied Arts offering undergraduate and postgraduate courses to adult learners (http://www.eap.gr). Each course consists of modules and students have to submit 4-6 written assignments throughout the 10-month academic year and a compulsory sit exam at the end of it. Furthermore, each course module includes five not compulsory face-to-face Counselling Group Sessions that take place in 9 cities all over the country. Tutor-student communication and interaction is mainly held through e-mail and telephone as well as though the portal (http://online.eap.gr). Students at HOU are provided with a variety of learning materials: especially adapted printed course material, audio and video material, CD-ROMs/software, specially prepared for distance learning.

The portal of HOU is based on the Moodle (Modular Object-Oriented Dynamic Learning Environment) platform and it has been offering services to students and tutors since the academic year 2013-14. Its pilot use had been offered during 2011-2012 although individual efforts to use Moodle services had been made since a decade before in a research framework. (Karaiskakis et al., 2008; Patriarcheas & Xenos, 2009; Lotsari et al., 2014). Through the portal, students have the ability to submit their written assignments, work reports and questionnaires within their academic studies, while tutors are able to return the students graded and annotated assignments or work reports. In addition, work spaces for asynchronous discussions at theme level, discussion groups of students and tutors are available which are managed via automated process. The service is configured properly to keep pace with the academic calendar of themes and to provide students and tutors with direct access to the activities of the every current week.

The frequency of the use of the portal is gradually increasing in the School of Science and Technology and especially in modules related to Computer Science, whereas in the modules of the School of Humanities is still limited. Nowadays, 255 course modules are served by the portal and 1682 tutors and 28666 students are using the portal services.

## The "Information Systems" postgraduate

The aim of the postgraduate course is to offer students the opportunity to acquire specialized knowledge in Information and Communication Technologies, and to prepare them to professionally design, develop and manage integrated information systems. The course is targeted at science and technology graduates and covers the design and development of programs and systems, the management and the quality of system development and advanced issues in telecommunications and networking. The Master's Degree is awarded upon the successful completion of at least four (4) modules and the submission of a dissertation (and subsequent successful examination). Students have to choose 4 modules among: Fundamental Specialization in Theory and Software, Fundamental Specialization in Computer Architecture and Computer Networks, Specialization in Software Engineering, Software Design and Management, Specialization in Networks and Communications.

### *The "Specialization in Software Engineering" module*

The completion of six written assignments during the academic year, and the final written exam are required for the successful fulfilment of the module. Shortly, the content of the module consists of databases (ER, MySQL etc.), the Java programming language, Operating Systems (concurrency and paging), and basic Data Mining techniques (Classification, Clustering, Regression, Association Rules, etc.).

### *The use of the discussion fora*

In the frame of the pilot use of the portal educational services and especially the use of the fora and other collaboration spaces several factors, such as access problems (lack of basic skills and/or inadequate infrastructure), lack of time, and lack of apparent activity in the collaboration space by others, were highlighted as important for user engagement and participation. The role of the tutor is also important and his on-line behaviour concerning his/her attitude towards the use of on-line discussion and collaboration spaces and his/her contribution creating new threads or/and answering to students' threads are crucial (Karaiskakis et al., 2008; Patriarcheas & Xenos, 2009; Lotsari et al., 2014). The number of the registered tutors and students who use the fora in practice, the frequency of their participation as well as the aim, the content, the expressing attitudes and behaviours are under investigation.

## Methodology

### *Description of the data*

The study was conducted using real data in Greek language gathered from the discussion fora of the Hellenic Open University. The data collected from the module "Specialization in Software Engineering" that is one of the four modules of the postgraduate course "Information Systems". At the end of the module, the four tutors of the module were asked for their permission to use the forum data for the purposes of the study. All the data were anonymized by replacing the names of the students with a registration number (ID). The data set consists of the forum activity of 64 students. Forum activity includes participation of each student, which means, starting a topic, creating discussion threads and exchanging messages. Eight (12%) of the students were women and 56 (88%) were men. A total of 371 messages were posted. 89 out of 371 were starting posts, while the rest 282 of them were replies. 198 messages were posted by the students and the rest were posted by tutors. A number of 89 threads consists the forum content with an average value of 3 posts/thread and 6 posts/student. The derived data set was small. This affected the strength of the results and was a limitation to our study. However, regardless its size, the data set was enough to raise questions about and to highlight reflection on the field under investigation and discussion.

## *Student participation in the fora*

Applying descriptive statistics on the set of data, the frequency of the students' and the tutors' participation in the discussion forum, as well as the distribution of the messages exchanged, the distribution of the starting posts and the distribution of the threads, were calculated. Based on these statistics the active students, who regularly participate in the discussions, and the less active students who mainly browse the forum searching for assignment solutions, without really contributing, were identified. Further analysis of the data focused on the active students.

## *Text mining of forum data*

In this study, our analysis started by extracting the text of the messages that appeared in the discussion forum. The text of these messages was converted to a corpus, after removing the punctuations, the numbers and the hyperlinks, and then the corpus was transformed in order to build a document-term matrix. In this matrix, each row represents a term, each column represents a document and an entry in this matrix is the number of the occurrences of the term in the document.

The major problem encountered, concerned the word stemming. In order to overcome this problem as well as to reduce the dimension of the matrix, a dictionary was created. This dictionary included those words with the highest frequency in the corpus, which are directly related with specific terms pertaining to the learning materials of the module under study. Thereby, only the words that were included in the dictionary are presented in the matrix.

Parameters, such as the Term Frequency and Term Associations, were calculated based on the document-term matrix. Terms with frequency greater or equal than ten, were finally included in further analysis and discussion.

## *Network of students*

A network of students based on their co-occurrence in the same thread was created. Each node represents a student and each edge represents a correlation between two students. The label size of vertices in the graph is based on the degree of the students' participation. Thicker edges represent higher degree of correlation.

## *Sentiment analysis*

Sentiment analysis is utilized in different levels: (a) sentence level, (b) document level and (c) corpus level (Liu, 2012). Opinion words are considered important elements for all levels. Such words were identified in text documents, by using a dictionary of the English language, with terms of known polarity. However, there is a limited amount of such material in Greek language. In our analysis, the active students and the less active students were identified. Additionally, the overall polarity of the messages of the active students in relevance to their performance outcome was examined. Lastly, for the analysis of the data available in Greek language, NioSto (Agathangelou et al., 2014), a software application that implements an opinion word extraction algorithm along with a dictionary-based sentiment classification, was used. The approach implemented by NioSto detects opinion words and classifies sentiment polarity of a message in one of the three categories: positive, neutral, or negative.

## Results and discussion

### *Student participation in the fora*

The frequency of the students' participation (including both starting and replying posts) for each student was calculated, detecting the more active and the less active students. Students' participation ranges between 1 and 21 messages, with an average of 6 posts per student. Students with at least 1 message count as active students resulting in a number of 34 active students in the forum of the course. Students' participation with starting posts ranges between 0 and 9 messages, with an average of 2 starting posts per student. Furthermore, tutors were decisively participating in the forum discussions with 13 starting posts and 173 replying posts, with an average of 58 messages per tutor and an average of 4 starting posts per tutor.

Figure 1 presents histograms with the number of messages and started posts written by students or tutors. The red vertical line indicates the average number of the messages or the started posts respectively. A descending trend is observed with the majority of both students and tutors who posted a number of messages lower than the average number of messages. Almost all of the starting posts of the students were related with the homework during each period as well as difficult themes of the educational material. However, there was only a small active group of students who discussed substantially over each other's questions. Tutors contributed to almost all discussions offering advices and indicative solutions.

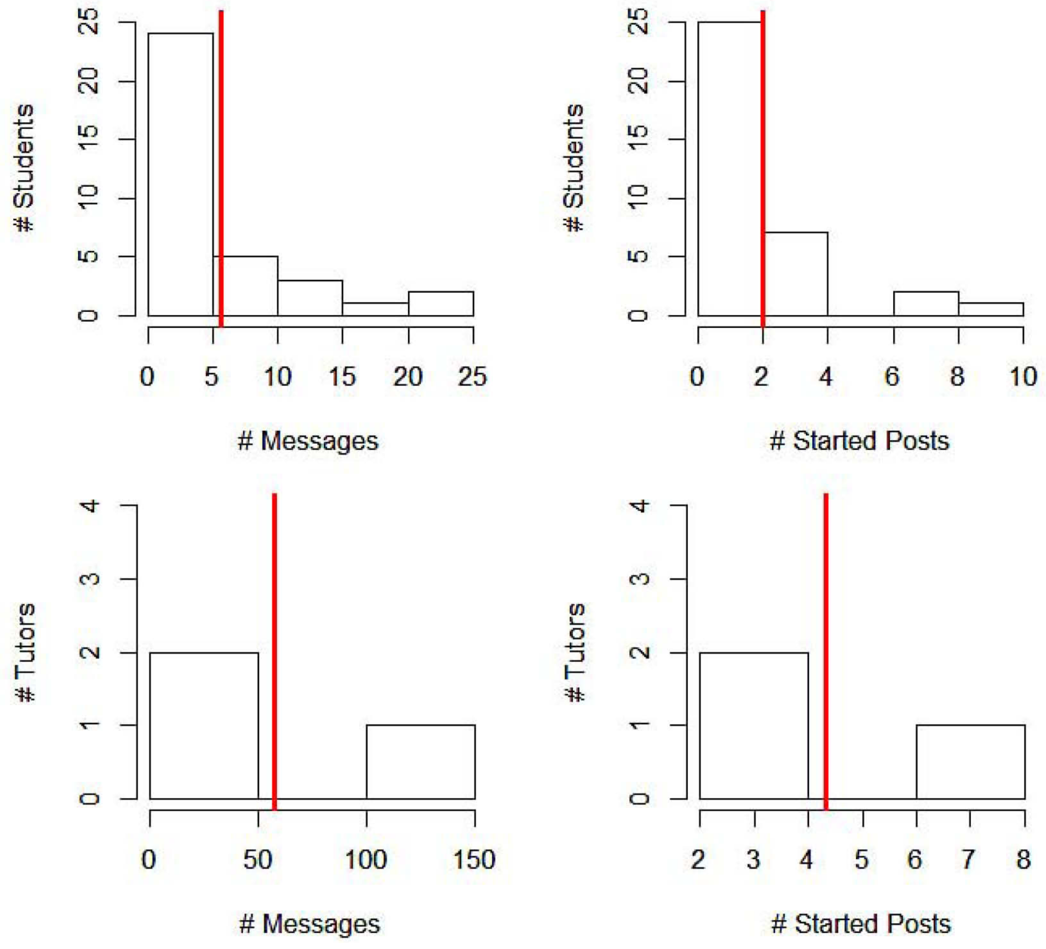

Figure 20. Number of messages and started posts per number of students and tutors

It was considered important to study the relationship between student's participation to the forum and his/her overall performance at the end of the module. As it is mentioned above, 34 students participated in the forum posting at least one message (active participants). The remaining 30 students were non-participants. They either were passive participants, who were possibly viewing the forum posts but were not posting to it or they were not participating at all. According to researchers the number of passive participants may be higher than the number of active participants in distance education courses (Smith & Smith, 2014) and it is worth able to be researched.

Figure 2 displays the number of the non-participants and the participants in the discussion forum and their respective performance. As it can be observed, a higher number of students who participated in the forum achieved very good and excellent results, compared to the number of the non-participant students. However, there were also more students who participated in the forum and failed or quitted the course. By calculating Pearson's correlation coefficient ($r_s = 0.93$, $p = 0.064$), it was concluded that the results are not statistically significant (since $p > 0.05$), and therefore participating in the forum is not directly associated with the students' performance.

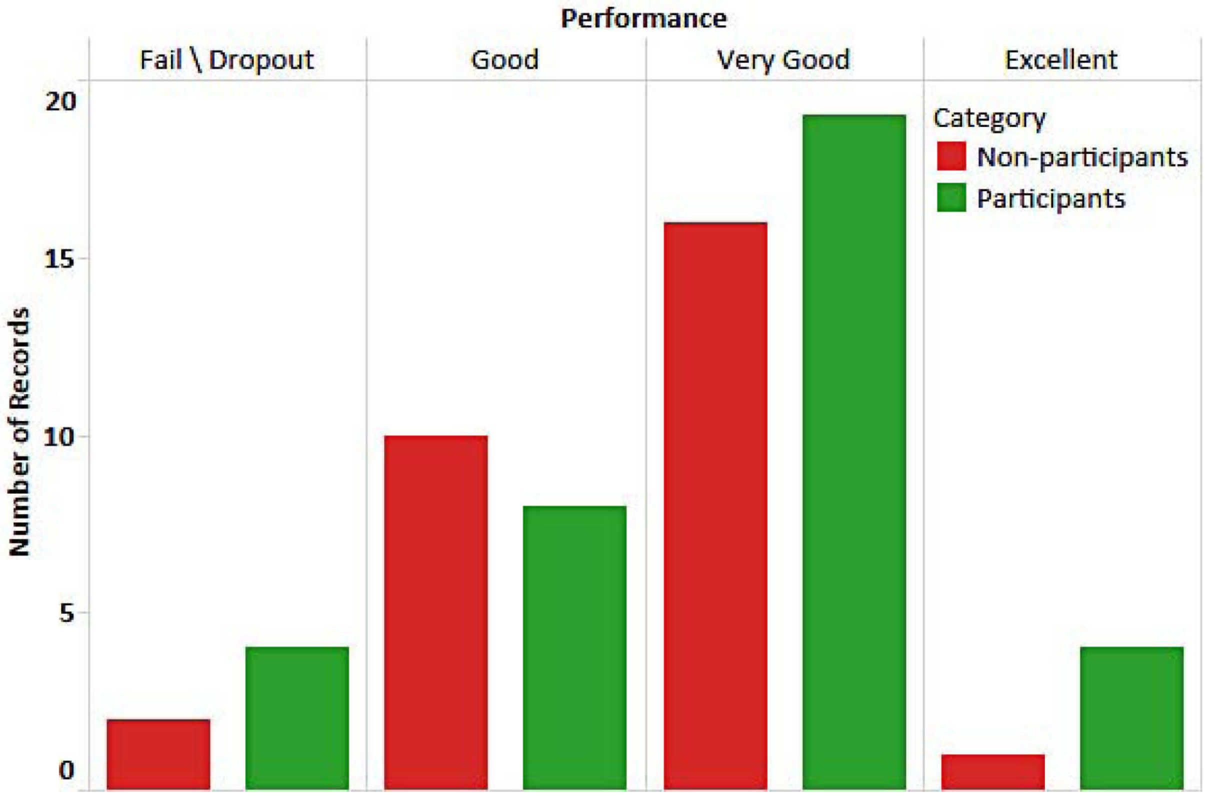

Figure 21. Number of non-participants and participants in the discussion forum and their respective performance

Figure 3 displays the number of messages without the starting posts and the number of starting posts in relevance to the average grade of the participants. It can easily be observed that no specific pattern appears, enhancing the aforementioned results.

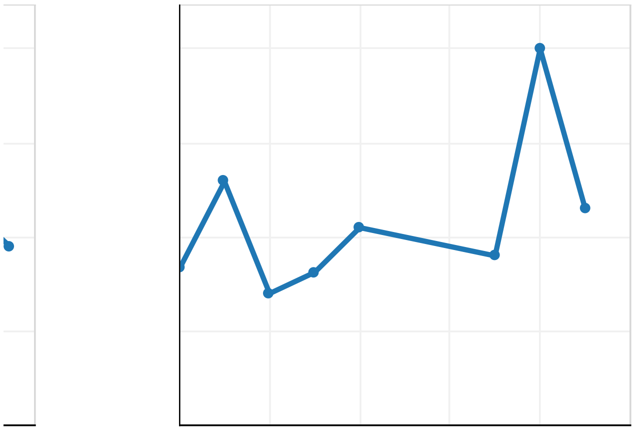

Figure 22. Number of messages (w/o started posts) and started posts in relevance with the average grade

The majority of the messages were posted during the periods of the written assignments with content relative to them. In Figure 4, a pie with the percentage of messages related with each assignment is displayed. Early assignments have a higher percentage of participation compared with the last assignments. Most messages were related with the three first assignments and less messages were related with the three last ones. Students at the beginning of the module and during the periods of the first assignments face many difficulties with the new subjects, the educational material and the study timetable. Therefore, they need to communicate with their fellow students and/or their tutors, in order to get the answers to their questions.

On the other hand, as it has been observed, most of the students put their effort into studying, posing questions and working on the assignments until they reach the threshold they need in order to have access to the final exams. Access to the final exams is given to every student who has obtained a specific cumulative grade from the assignments. As a result, many students slow down their efforts, and this deceleration may affect their participation in the forum, as well.

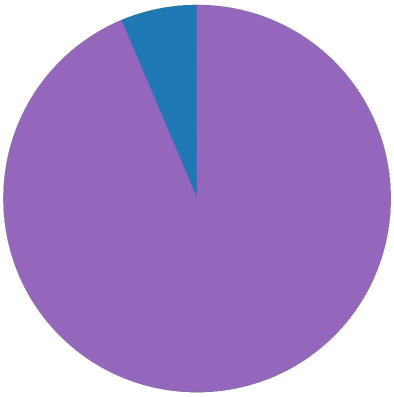

Figure 23. Number of messages per homework period

## *Text mining of forum data*

Table 1 shows the terms in the dictionary including the words with the higher frequency in the corpus, which are directly related with specific terms pertaining to the learning materials of the module under study.

According to the barplot (Figure 5), among the most popular terms are the words “array”, “java”, “schema”, “data”, and “mysql”, revealing issues that are related to difficulties the students might have faced with the educational content. Term frequencies and term associations highlighted topics related to Conceptual Database Modelling (“mysql”, “sql”, “array”, “schema”, “database”, “attribute”, “query”, “key”, “type”, “ternary”, “correlation”, “model”, “diagram”, “entity”), Java Programming Language (“java”, “class”, “superclass”, “object”, blue j”, “constructor”, “ship”, “submarine”) and Operating Systems. Conceptual database modelling, Java programming language and Operating systems are main subjects of the module and strongly related with at least two of the written assignments students ought to submit during the academic year.

Most of the terms mentioned previously are included in the description and the solution of the three first assignments. Most messages in the forum were related with the three first assignments as it has been described in “Description of the data” section.

The purpose of this information is to offer an overall view of the difficulties that students may face and to enable, as a result, each tutor to focus his/her attention on specific concepts of the course by trying to enrich the educational material and to improve the learning process.

Table 1: Words of dictionary

| | | | | | | |
|---|---|---|---|---|---|---|
| array | table | by | ubuntu | management | query | programming |
| bluej | eclipse | database | sequence | complicated | rdbms | code |
| constructor | page | having | server | submarine | constraint | method |
| exception | grep | mysql | objectoriented | data | relation | class |
| fire | bash | java | operating | type | er | battle |
| ide | shell | forum | systems | attribute | erd | slides |
| r | virtual | script | semaphore | ratio | diagram | os |
| resource | machine | sql | process | ternary | relational | pseudocode |
| segment | key | linux | program | weak | schema | tuples |
| ship | awk | from | operator | correlation | model | superclass |
| throws | sed | select | package | entity | algebra | object |

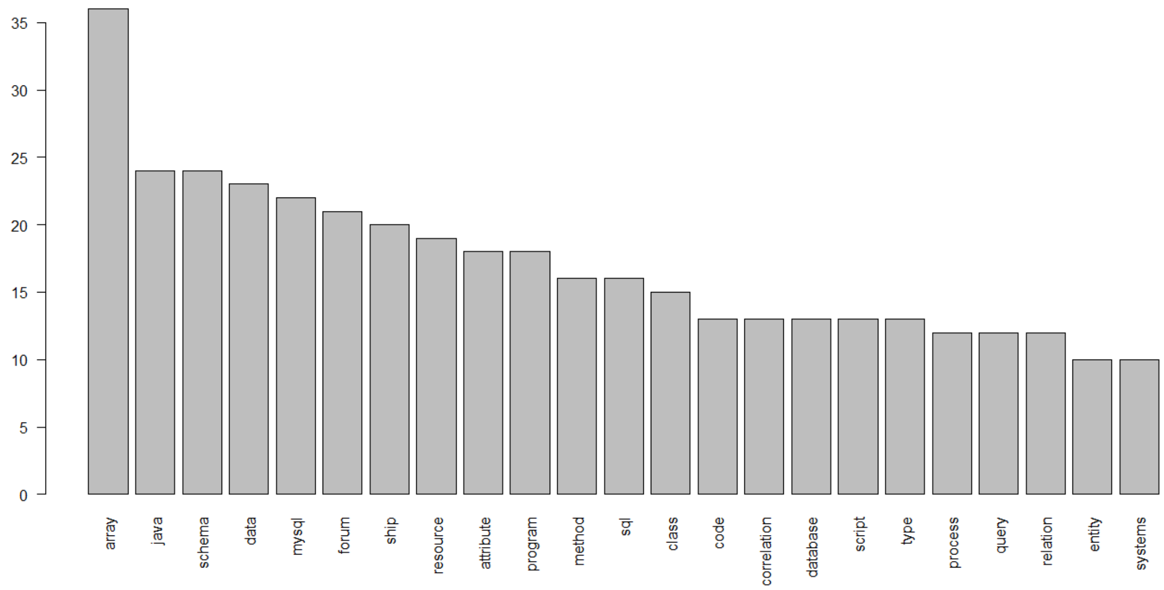

Figure 24. A plot of terms along with their frequency of occurrence (terms appeared at least ten times in the forum discussions are included)

## *Network of students*

Complementary data concerning students' behaviour in the forum are derived from the students' network that is built and presented in Figure 6. The students' network is based on their co-occurrence in the same thread and illustrates the interactions among them. Each node presents a student and each edge presents a correlation between two students. Students with higher levels of participation in the discussion forum are at the centre of the network. For instance, students with IDs 83117, 61122 and 83172 are located close to the centre as they have the highest frequency of participation in the forum being the most active participants. Figure 6 demonstrates the active as well as the peripheral students allowing the tutor to have a visual description of the interactions among his/her students. Furthermore, labels id1, id2 and id3 refer to the course tutors, who seem to be an important factor of the interaction in the forum and they have a central role in the network.

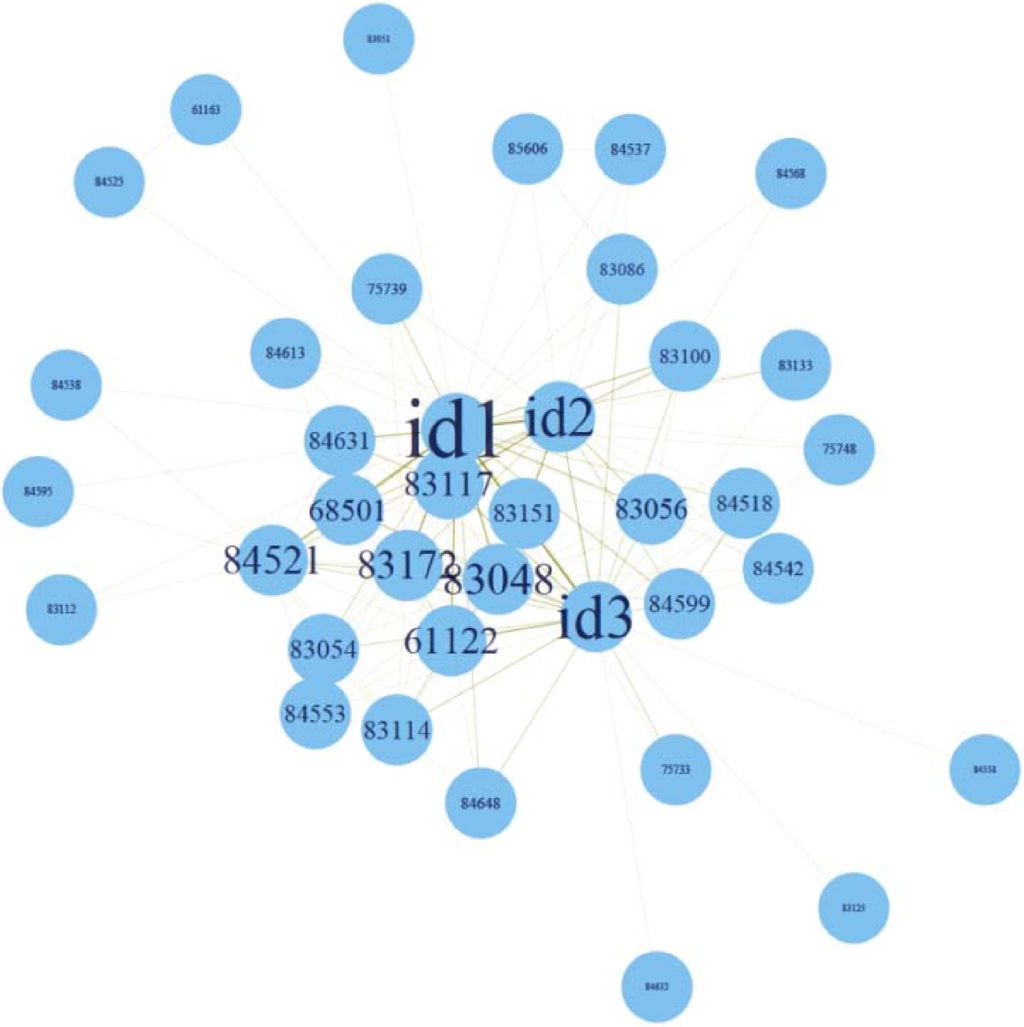

Figure 25. Students' network

## *Sentiment analysis*

Figure 7 visualizes the sentiment classification of the messages posted by students, as it was determined by NioSto. Basically, Figure 7 shows the variance of the proportion of each category in relation to the number of words in students' messages. The total percentages for each category are 27.27% positive, 55.56% neutral and 17.17% negative.

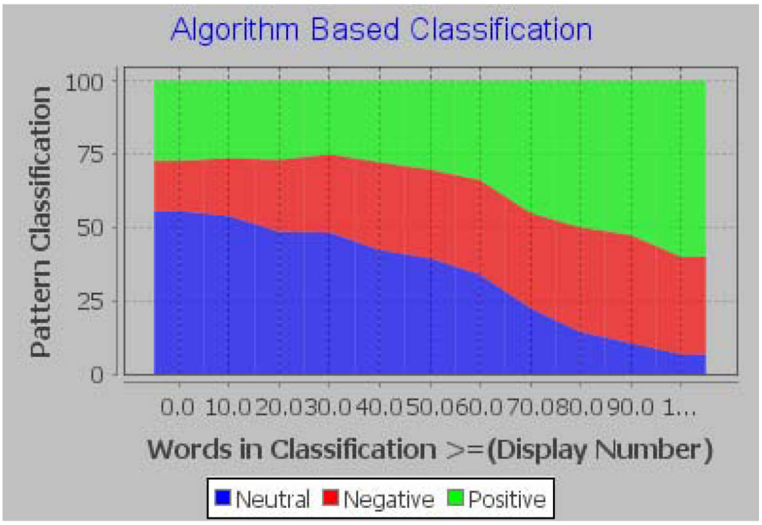

Figure 26. Sentiment classification on students' messages from NioSto

Respectively, Figure 8 shows the variance of the proportion of each category in relation to the number of words in tutors' messages. The total percentages for each category are 42.53% positive, 49.42% neutral and 8.05% negative.

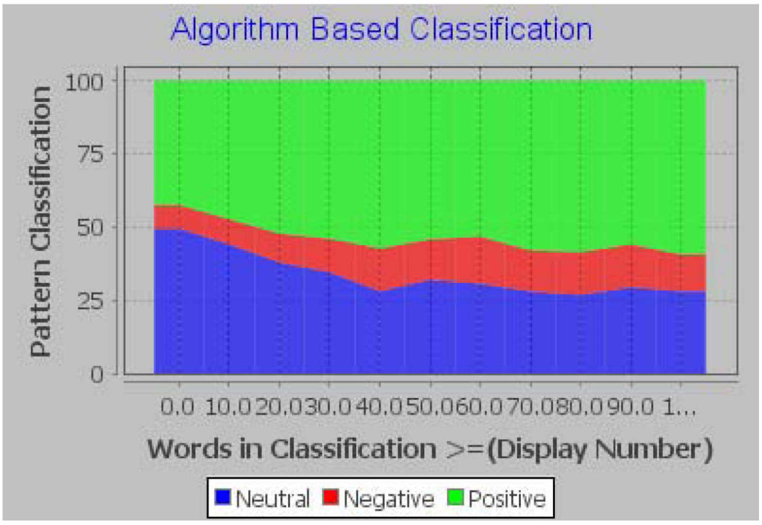

Figure 27. Sentiment classification on tutors' messages from NioSto

The majority of the messages for both groups (students and tutors) are classified as neutral. However, neutral messages cannot play a major role in the analysis. Thus, for understanding better the trend of sentiment polarity, the neutral category was removed completely. Messages containing Christmas wishes were also removed, as this could possibly affect the polarity of the message. Figure 9 displays separately for students and tutors the difference between positive and negative messages, divided by the total number of positive and negative messages per month.

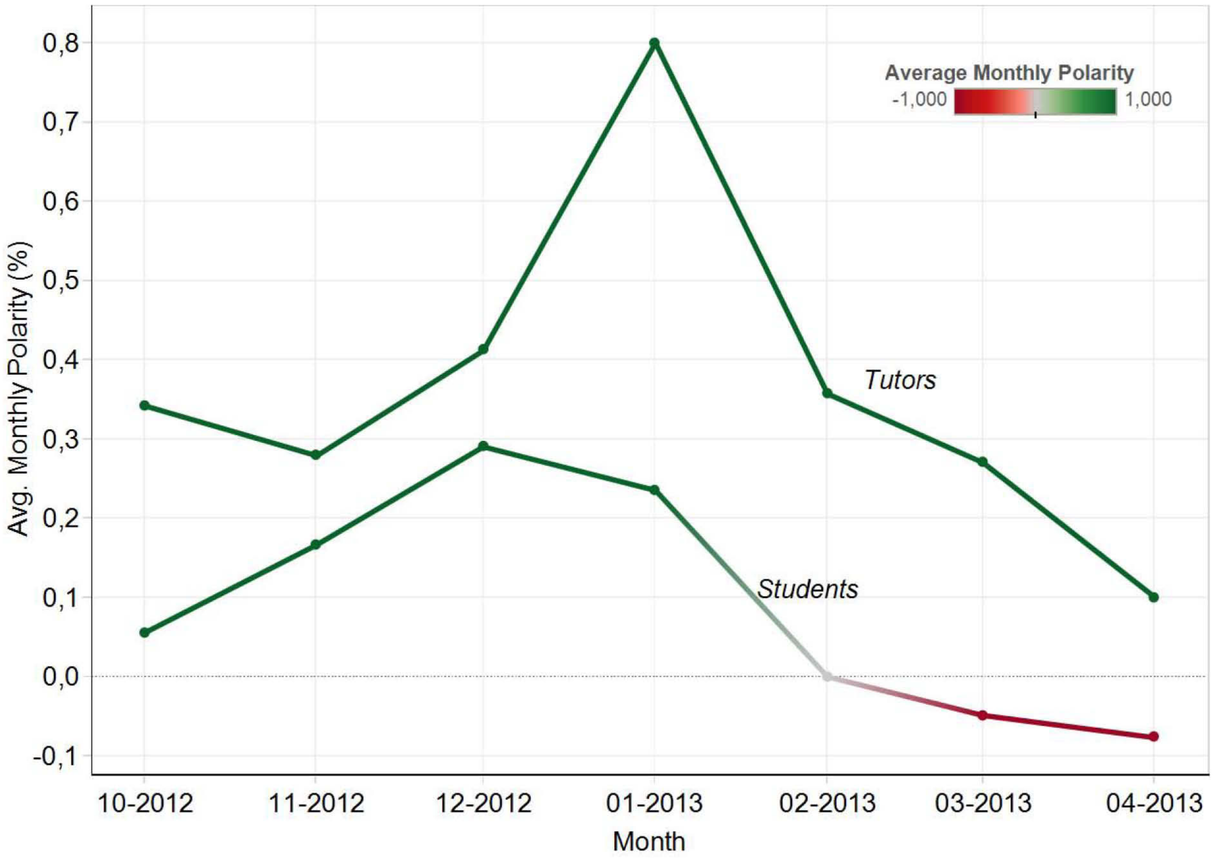

Figure 28. Avg. monthly polarity (%) for students and tutors without posts for Christmas wishes

The overall polarity of the students reaches its maximum, in November 2012 and in January 2013. Then, it follows a descending trend. Negative polarity starts to appear in February 2013 and it peaks at its minimum, in April 2013. A reasonable explanation for this turn is that students were more excited and, therefore, more positive at the beginning of the course. But as the course was in progress, and perhaps due to the workload and the pressure the students started to feel, their perspective changed and, thus, their messages started having a more negative polarity.

As far as the tutors' monthly polarity is concerned, an ascending trend is observed, that reaches its maximum in January 2013. Then, a descending trend follows until the end of the course in April 2013. However, the overall polarity never becomes negative and it is much more positive, compared to the students' polarity. This is not surprising, as the tutors are expected to reply kindly to the students' questions, to give them feedback, offering at the same time psychological support to their students.

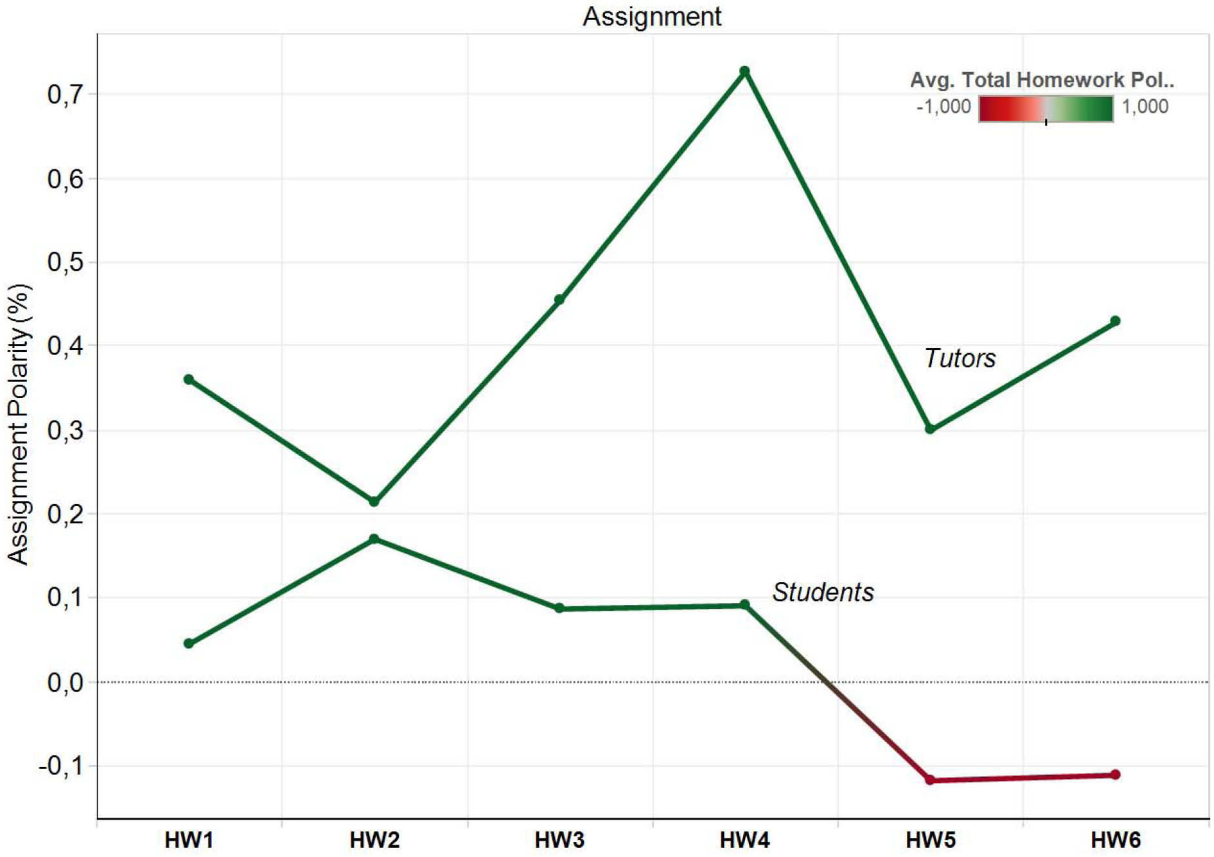

Figure 29. Assignment polarity (%) for students and tutors

Figure 10 displays the difference between positive and negative messages, divided by the total number of positive and negative messages per assignment, separately for students and tutors. The assignments are chronologically ordered depending on their deadline date. More or less, a similar pattern with the one presented in Figure 7 is observed, for both students and tutors. The difficulty of the assignments increases as the assignments are given, meaning that HW1 is much easier than HW6. It looks like there is a relation between the difficulty and the overall polarity.

Figure 11 presents the grade and the polarity separately for each student. The average grade (7.035) of the participants in the discussion forum is indicated by the black bold line. Dotted lines (at 6.5 and 8) make it more clear, if and how polarity is connected with performance. Although for lower grades it is not clear if a specific pattern exists, it can easily be noticed that higher grades are achieved by students with positive polarity. For the data presented in Figure 11, Pearson's correlation coefficient was applied. The results indicate that there is a weak correlation between the polarity and the grade, with a marginally statistical importance ($r_s = 0.334$, $p = 0.049$).

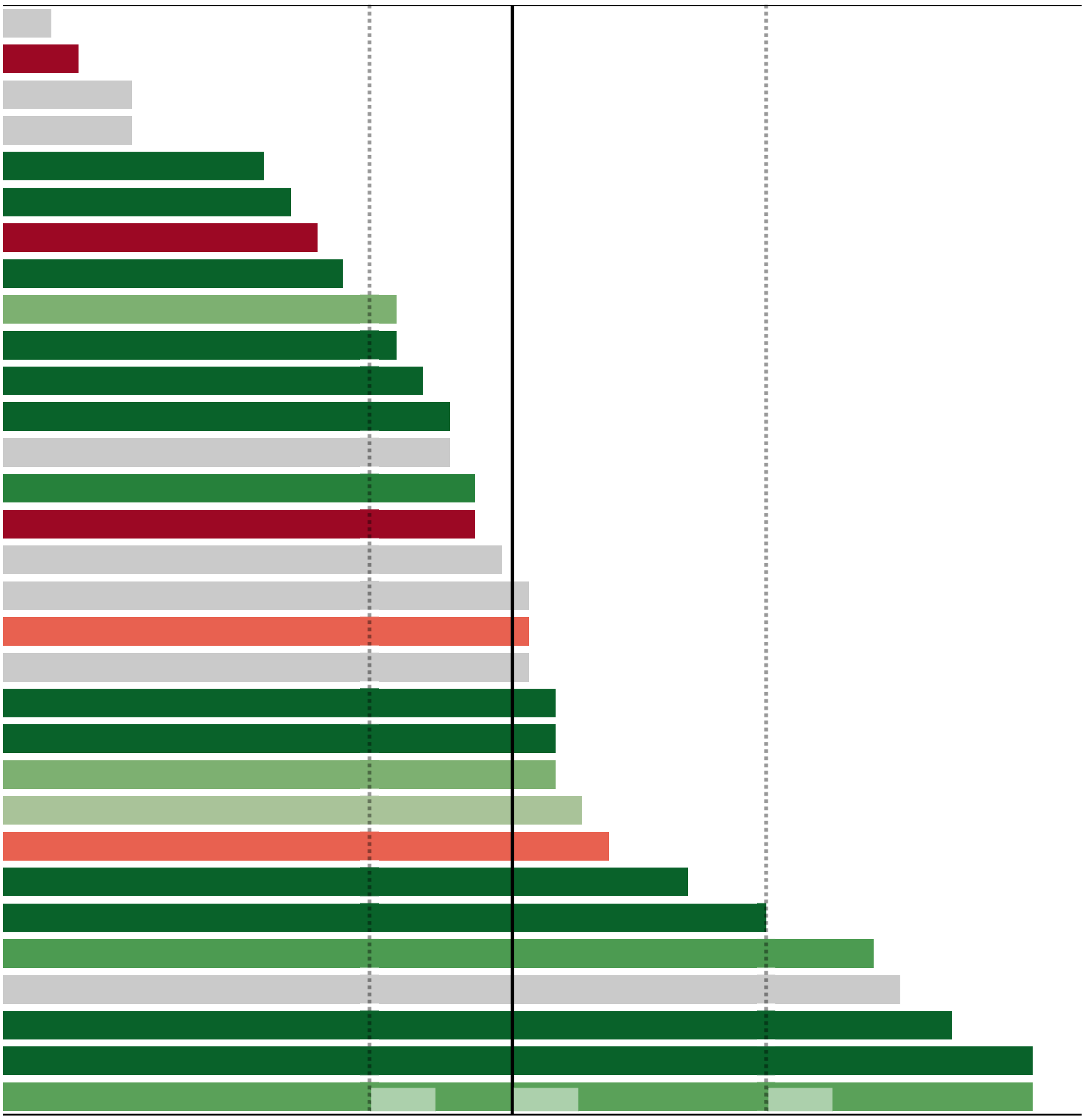

Figure 30. Grade and polarity (%) per student

## Conclusions and future work

In this study, data related to the participation of postgraduate students in the online forum of their course at the Hellenic Open University was analysed. Text mining techniques, social network analysis techniques and sentiment analysis were applied on the same dataset. The combined knowledge attained from the aforementioned techniques can provide tutors with practical and valuable information for the structure and the content of the students' exchanged messages, the patterns of interaction among students, the trend of the sentiment polarity during the course and its affection on the students' performance, aiming at improving the educational process.

The frequency of the participation of the students in the forum and the interaction among them were illustrated, providing tutors with useful information in order to study their participation in the forum and thereby to distinguish the active, the peripheral, and the passive or non-participant students in the discussions. By applying statistical techniques to analyse the content of the exchanged messages in the fora, important terms that reveal the discussion topics emerged, enabling tutors to focus their attention on specific concepts of the educational material. Students' participation in the forum did not prove to be an important factor that affects their final performance. The polarity of the students' messages proved marginally to be related with their performance.

Most of the effort was put on analysing data from a single academic year. Therefore, we used a rather small data set, which affected the reliability of the derived statistical results and correlations. Because of this, our results give some intuition, but there is still space for providing more evidence about whether there is a specific pattern between students' participation, sentiment and performance. Deeper analysis based on a bigger set of data can provide information for students' profiles based on their participation in the fora of the course. It can also provide information about students' interaction among them and their interaction with the educational material. This can be achieved by watching and studying the students' sentiment during the course, and by examining the effect of all these factors on students' performance and learning. Our work will be extended by utilizing data for more than one consecutive years, for the same module of the distance learning course under discussion and for different courses of the H.O.U. as well.

## References

1. Abel, F., Bittencourt, I.I., Costa, E., Henze, N., Krause, D., & Vassilev, J. (2010). Recommendations in Online Discussion Forums for E-Learning Systems. *IEEE Transactions on Learning Technologies, 3*(2), 165-176.

2. Agathangelou, P., Katakis, I., Kokkoras, F., & Ntonas K. (2014). Mining Domain-Specific Dictionaries of Opinion Words. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali & Y. Zhang (Eds.), *Proceedings of the 15th International Conference on Web Information System Engineering (WISE 2014)* (pp. 47-62). Thessaloniki, Greece, 12-14 October, 2014. LNCS 8786, Springer. doi: http://dx.doi.org/10.1007/978-3-319-11749-2_4

3. Anderson, T. (2004). Towards a theory of online learning. In T. Anderson & F. Elloumi (Eds.), *Theory and practice of online learning* (pp. 33-60). CA: Athabasca University Press.

4. Anderson, P. (2007). *What is Web 2.0? Ideas, technologies and implication for education. JISC Technology and Standards Watch.* Retrieved 10-1-2014 from http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf

5. Anderson, T., & Garrison, D.R. (1998). Learning in a networked world: New roles and responsibilities. In C. Gibson (Ed.), *Distance learners in higher education* (pp. 97-112). Madison: Atwood Publishing.

6. Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

7. Berland, M., Baker, R.S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning, 19*(1–2), 205–220. doi:10.1007/s10758-014-9223-7

8. Bradley, A.J., & McDonald, M.P. (2011, October 26). Social Media versus Knowledge Management [Blog post]. Harvard Business Review Blog. Retrieved 10-1-2014 from http://blogs.hbr.org/2011/10/social-media-versus-knowledge/

9. Brindley, J.E., Walti, C., & Blaschke, L. M. (2009). Creating Effective Collaborative Learning Groups in an Online Environment. *IRRODL, 10*(3). Retrieved 10-01-2014 from http://www.irrodl.org/index.php/irrodl/article/view/675/1271

10. Cambria, E., & Hussain, A. (2012). *Sentic Computing: Techniques, Tools, and Applications*. Springer. Retrieved from http://sentic.net/sentic-computing.pdf

11. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems, 28*(2), 15–21. doi:10.1109/MIS.2013.30.

12. Carlos, A.R. (2011). *Social Network Analysis in Telecommunications*. John Wiley & Sons. ISBN 978-1-118-01094-5.

13. Carsten, U., Boreau, K., & Stepanyan, K. (2010). Who students interact with? A social network analysis perspective on the use of Twitter in Language Learning. In M. Wolpers, P. Kirschner, M. Scheffel & V. Dimitrova (Eds.), *Proceedings of 5th European Conference on Technology Enhanced Learning* (pp. 432-437). Sankt Augustin: Springer.

14. Choi, I., Land, S. M., & Turgeon, A. J. (2005). Scaffolding peer-questioning strategies to facilitate metacognition during online small group discussion. *Instructional Science, 33*, 483-511.

15. D' Andrea, A., Ferri, F., & P. Grifoni (2009). An Overview of Methods for Virtual Social Network Analysis. In A. Abraham, A.E. Iassanien & V. Snášel (Eds.), *Computational Social Network Analysis: Trends, Tools and Research Advances*. Springer. ISBN 978-1-84882-228-3.

16. Groves, M., & O'Donoghue, J. (2009). Reflections of Students in Their Use of Asynchronous Online Seminars. *Educational Technology & Society, 12*(3), 143-149.

17. Hülsmann, T. (2009). Access and Efficiency in the Development of Distance Education and E-Learning. In U. Bernath, A. Szücs, A. Tait & M. Vidal (Eds.), *Distance and E-Learning in Transition - Learning Innovation, Technology and Social Challenges* (p. 121). London/Hoboken: ISTE Ltd and John Wiley & Sons, Inc.

18. Johnson, L., Smith, R., Willis, H., Levine, A., & Haywood, K., (2011). *The 2011 Horizon Report*. Austin, Texas: The New Media Consortium.

19. Karaiskakis, D., Kalles, D., & Hadzilacos, Th. (2008). Profiling Group Activity of Online Academic Workspaces: the Hellenic Open University case study. *International Journal of Web-based Learning and Teaching Technology, 3*(3), 1-15.

20. Kim, S.M., & Hovy, E.H. (2006). Identifying and Analyzing Judgment Opinions. *Proceedings of the Human Language Technology / North American Association of Computational Linguistics conference* (HLT-NAACL 2006). New York, NY. http://acl.ldc.upenn.edu/P/P06/P06-2063.pdf

21. Kostourakis, G., Panagiotakopoulos, C., & Vergidis, D. (2008). A contribution to the Hellenic Open University: Evaluation of the Pedagogical Practices and the use of ICT on Distance Education. *International Review of Research in Open and Distance Learning, 9*(2). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/424/1044

22. Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool publishers.

23. Lotsari, E., Verykios, V.S., Panagiotakopoulos, C., & Kalles, D. (2014). A Learning Analytics Methodology for Student Profiling. In A. Likas, K. Blekas & D. Kalles (Eds.), *Artificial Intelligence: Methods and Applications,* 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings (Volume 8445 of the series Lecture Notes in Computer Science pp. 300-312). Retrieved from http://link.springer.com/chapter/10.1007%2F978-3-319-07064-3_24

24. Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

25. Nasukawa, T., & Yi, J. (2003). Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture,* 70-77.

26. Ortigosa, A., Martín, J.M., & Carro, R.M. (2014). *Sentiment analysis in Facebook and its application to e-learning.*

27. Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions.* Cambridge University Press. Retrieved from http://www.cogsci.northwestern.edu/courses/cg207/readings/Cognitive_Structure_of_Emotions_exerpt.pdf

28. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

29. Patriarcheas, K., & Xenos, M. (2009). Modelling of distance education forum: Formal languages as interpretation methodology of messages in asynchronous text-based discussion. *Computers & Education, 52*(2), 438-448. Retrieved from https://scholar.google.gr/citations?view_op=view_citation&hl=el&user=I3UJPkQAAAAJ&citation_for_view=I3UJPkQAAAAJ:u5HHmVD_uO8C

30. R Core Team (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

31. Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education, 51*(1), 368-384.

32. Sahu, C. (2008). An evaluation of selected pedagogical attributes of online discussion boards. *Hello! Where are you in the landscape of educational technology? Proceedings: 25th annual ASCILITE conference, Melbourne 2008*. Retrieved 09-01-2014 from http://www.ascilite.org.au/conferences/melbourne08/procs/sahu.pdf

33. Siemens, G. (2010, August 25). What Are Learning Analytics? [Blog post]. Elearnspace. Retrieved from http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/

34. Siemens, G., & Baker, R.S.J. (2012). Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. *LAK12, 2012*.

35. Smith, D., & Smith, K. (2014). The Case for 'Passive' Learning – The 'Silent' Community of Online Learners. *European Journal of Open, Distance and E-Learning, 17*(2), 85-98. Retrieved from http://www.eurodl.org/index.php?p=archives&year=2014&halfyear=2&article=649

36. Stevenson, R., Mikels, J., & James, T. (2007). Characterization of the Affective Norms for English Words by Discrete Emotional Categories. *Behavior Research Methods, 39*(4), 1020–1024. Retrieved from http://indiana.edu/~panlab/papers/SraMjaJtw_ANEW.pdf

37. Tableau Desktop Software (2015). Retrieved from http://www.tableau.com/products/desktop

38. Takaffoli, M., & Zaïane, O.R. (2012). Social network analysis and mining to support the assessment of on-line student participation. *ACM SIGKDD Explorations Newsletter, 13*(2), 20-29.

39. Turney, P. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the Association for Computational Linguistics,* 417-424.

40. Thomas, M.J.W. (2002). Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning, 18*(3), 351-366.

41. Wen, M., Yang, D., & Penstein Rosé, C. (2014). *Sentiment Analysis in MOOC Discussion Forums: What does it tell us? Educational Data Mining.* Retrieved from http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf