Accepted Manuscript

How automated feedback through text mining changes plagiaristic behavior in online assignments

Gökhan Akçapınar, Ph.D.

PII: S0360-1315(15)00108-6

DOI: 10.1016/j.compedu.2015.04.007

Reference: CAE 2838

To appear in: *Computers & Education*

Received Date: 8 January 2015

Revised Date: 10 April 2015

Accepted Date: 13 April 2015

Please cite this article as: Akçapınar G., How automated feedback through text mining changes plagiaristic behavior in online assignments, *Computers & Education* (2015), doi: 10.1016/j.compedu.2015.04.007.

How automated feedback through text mining changes plagiaristic behavior in online assignments

Gökhan Akçapınar, Ph.D.

Address: Department of Computer Education and Instructional Technologies, Hacettepe University, Ankara, Turkey.

Telephone: +90 312 2977176

E-mail: gokhana@hacettepe.edu.tr

Abstract: This study is intended to decrease learners' plagiaristic behaviour in online assignments by providing automated feedback based on text mining analysis. Document similarity analysis was done at the middle and end of the semester on 4,268 reflection texts (max. 500 characters) written by the participating university students (N=59) about concepts which they had learned in the computer science course. In the analysis conducted in the middle of the semester, the plagiarism ratios in the posts that students wrote during the first six weeks were calculated, and these ratios were used to give the students feedback. The analysis conducted at the end of the semester was used to test whether there was a change in the plagiaristic behaviours of the students after they had received feedback. To calculate the plagiarism ratio, the students' posts were compared to internet search results and posts written by other students. A paired samples t-test was used to identify whether there was a difference in the plagiarised post ratios before and after feedback was provided. In addition, the plagiaristic behaviours of the students before and after feedback were compared proportionally using the McNemar's Chi-square test. The results of the analyses revealed a statistically significant difference in the plagiarised post ratios and the ratios of students performing plagiaristic behaviour before and after feedback. It was found that after feedback was provided, while the ratio of plagiarised posts went down 21.07% on average, 83% of the students had lower plagiarised post ratios. The number of students who did not commit plagiarism increased by 42.37% after feedback was provided.

Keywords: Plagiarism, Copy detection, Document similarity, Educational data mining, Text mining, Online learning, Automated feedback.

The Effect of Automated Feedback Using Text Mining on Plagiaristic Behaviour in Online Assignments

Abstract: This study is intended to decrease learners' plagiaristic behaviour in online assignments by providing automated feedback based on text mining analysis. Document similarity analysis was done at the middle and end of the semester on 4,268 reflection texts (max. 500 characters) written by the participating university students (N=59) about concepts which they had learned in the computer science course. In the analysis conducted in the middle of the semester, the plagiarism ratios in the posts that students wrote during the first six weeks were calculated, and these ratios were used to give the students feedback. The analysis conducted at the end of the semester was used to test whether there was a change in the plagiaristic behaviours of the students after they had received feedback. To calculate the plagiarism ratio, the students' posts were compared to internet search results and posts written by other students. A paired samples t-test was used to identify whether there was a difference in the plagiarised post ratios before and after feedback was provided. In addition, the plagiaristic behaviours of the students before and after feedback were compared proportionally using the McNemar's Chi-square test. The results of the analyses revealed a statistically significant difference in the plagiarised post ratios and the ratios of students performing plagiaristic behaviour before and after feedback. It was found that after feedback was provided, while the ratio of plagiarised posts went down 21.07% on average, 83% of the students had lower plagiarised post ratios. The number of students who did not commit plagiarism increased by 42.37% after feedback was provided.

Keywords: Plagiarism, Copy detection, Document similarity, Educational data mining, Text mining, Online learning, Automated feedback.

1. Introduction

Compared to traditional classes, online learning environments bring certain problems along with their advantages in the acquisition of materials and being independent from time and space (Koulocheri, Soumplis, & Xenos, 2012). Some of these problems are disorientation, lack of motivation (Bolliger, Supanakorn, & Boggs, 2010), technical difficulties (Ferguson, 2012) and high dropout rates (Hew & Cheung, 2014; Santos, et al., 2014). Another important problem with online learning environments is plagiarism on examinations and assignments (Bedford, Gregg, & Clinton, 2009). Plagiarism is a growing problem (Rushby, 2013). Studies show that the rate of plagiarism on examinations and assignments has substantially increased in recent years and the tendency to plagiarise in online learning environments is higher than it is in traditional environments (Batane, 2010; Scanlon, 2003). Some of the major reasons for this increase are easier access to information on the internet

(Akbulut, et al., 2008; Maurer, Kappe, & Zaka, 2006), retrieving information directly from the source by copying and pasting (Watson & Sottile, 2010) and the lack of supervision by teachers (Rowe, 2004).

Even though it is easier to plagiarise in online environments than it is in traditional ones, they also provide an advantage for the detection of plagiarism since any kind of data is digitally stored in these systems (Baron & Crooks, 2004). Various methods have been suggested so far in order to prevent plagiarism, which constitutes a serious problem for the promotion of online learning. The suggestions for examinations include organising them on campuses or in testing centres using traditional methods, having students monitored in examinations by supervisors who are sent by a trusted agency or using technology-based solutions such as a remote proctor (Bedford, et al., 2009). On the other hand, different types of assignments (e.g., portfolios, forums) or technology-based solutions such as plagiarism detection tools have been suggested for online assignments (Baron & Crooks, 2004). For example, some educational institutions require that assignments sent by students be analysed using plagiarism detection software (Whittle & Murdoch-Eaton, 2008).

In this study, the problem of digital plagiarism was addressed in the context of educational data mining, and it was investigated whether the automatic feedback provided by means of text mining was effective in preventing this serious problem that obstructs online learning. The remainder of this paper is organised as follows. Section 2 describes the background knowledge about digital plagiarism and plagiarism prevention methods. Section 3 presents the research methods, including this study's participants, its experimental design, its data, and its data analysis process. Section 4 presents its main results. Finally, Section 5 draws conclusions from this research, indicates its limitations and makes suggestions for further research.

2. Background

Plagiarism is not a new phenomenon (Ercegovac & Richardson, 2004); however, it has become a problem that threatens online learning environments. The term digital plagiarism is used to identify plagiarism in online learning environments (Kauffman & Young, 2015). The most commonly used method for digital plagiarism is taking all or some parts of an assignment from another resource on the internet by copying and pasting text (Jocoy & DiBiase, 2006; Sisti, 2007). Another method frequently used by students is committing plagiarism by copying their classmates or assignments from previous years (Butakov & Scherbinin, 2009; Karim, Zamzuri, & Nor, 2009). Studies show that high ratios of college students commit plagiarism in online assignments (Sisti, 2007; Stephens, Young, & Calabrese, 2007). The plagiarism problem is now has become a serious problem concerning not only the educational institutions where it exists, but society as whole (Batane, 2010).

2.1. Plagiarism Prevention Methods

Studies report that the most common method to prevent plagiarism is commercial plagiarism detection software, which checks whether a text partially or wholly matches internet sources or work turned in by other students (Whittle & Murdoch-Eaton, 2008). There are some studies in the literature in which anti-plagiarism software was used to provide feedback to the students. In his study, Rolfe (2011) studied the effects of instant feedback from Turnitin, a plagiarism detection service, on students' academic literacy levels. The students who participated in the study wrote draft essays, analysed them with Turnitin and revised their writing according to the instant feedback which they received. After his study, Rolfe noted that using Turnitin was not only viewed positively by the students, but also encouraged them to improve their writing skills. Another study that compared students' writing behaviours was conducted with two separate classes that had similar characteristics (Stapleton, 2012). The researcher told one of the classes that their essays would be compared in terms of genuineness; however, he withheld this information from the other class. The results showed that the students who were unaware that their essays would be analysed, included significantly more copied texts and intentional plagiarism in their essays than the students in the other class. In addition to plagiarism detection software, search engines (Heberling, 2002), data mining (Cavalcanti, Pires, Cavalcanti, & Pires, 2012) and text mining methods (Oberreuter & Velásquez, 2013) have also been used to detect plagiarised documents.

2.2. Research Questions

In this study, problems with online plagiarism were studied by taking an educational data mining approach. The author investigated whether automatic feedback using text mining has any effect on students' plagiarism. Feedbacks are crucial variables for changing behaviours, and as Bienkowski, Feng, and Means (2012) demonstrated, individualised, comprehensible feedback can effectively change students' behaviours. A study at Purdue University showed the significant effect on learning performance of simple feedback about students' progress in the class (K. E. Arnold, 2010).

It is now possible to provide instant and frequent feedback to students thanks to improvements in educational technologies (Tanes, Arnold, King, & Remnet, 2011). Educational Data Mining (EDM) and Learning Analytics (LA) are emerging fields that focus on using the data collected in educational environments to improving education (Siemens & Baker, 2012). One crucial goal of EDM and LA studies is identifying students who have problems, and modifying adverse behaviours with timely interventions (Kimberly E. Arnold & Pistilli, 2012). The feedback interventions given in EDM and LA studies usually aims to supply information derived from interaction data that will be useful for teachers and students (Chatti, Dyckhoff, Schroeder, & Thüs, 2012).

Studies show that technology-supported solutions yield successful results in reducing plagiarism (Batane, 2010; Stappenbelt & Rowles, 2010). However, offering such systems without instructional planning might not result in the expected behavioural changes. Moreover, as researchers emphasise (Stapleton, 2012; Whittle & Murdoch-Eaton, 2008), learners can discover means and techniques to abuse these systems. Therefore, this study hypothesised that the automatic feedback mechanism used in this study would have a positive effect on changing students' plagiaristic behaviours and posed the following research questions:

1. How effectively do written and verbal warnings provided by the instructor and the elimination of copying and pasting from online environments prevent plagiarism?
2. Does automatic feedback provided by the system affect students' plagiaristic behaviours?
3. Is there a statistically significant difference between the plagiarised post ratios in the posts written before and after the feedback provided?
4. Is there a statistically significant difference between the ratios of students who did not commit plagiarism before and after the feedback provided?

3. Method

This study was conducted in the Computer Hardware course. This course was held for 14 weeks with 2 hours of face to face class time per week and online activities for the remaining times. For the online portion of the mixed environment course, an online learning environment designed by the researcher was used. The similarity analysis was done twice, in the middle and at the end of the semester. The results of the first analysis, which were calculated in the middle of the semester, were used for giving feedback to students. The results of the second analysis were used to analyse whether the plagiarism ratios had changed after feedback was given. A one-group pretest–posttest experimental design was used to test for these changes. A paired samples t-test was used to find whether there were significant differences between the plagiarised post ratios that were obtained before and after the feedback was provided. In addition to t-test, the McNemar‘s Chi-square test was used to determine if there was a statistically significant difference between the ratios of students committing plagiarism before and after the feedback was provided. The text mining analyses were performed by the RapidMiner data mining tool with the Text Processing package. The paired sample t test was performed using SPSS, and the McNemar test was done with MedCalc statistical software.

3.1. Participants

In 2012-2013, 66 undergraduates (32 female, 34 male) ranging in age from 19 to 23 years who were taking the Computer Hardware course in the Computer Education and Instructional Technology Department participated in this study. During the study, the data obtained from two students who did not log on to the online learning environment during the semester and five students who did not post anything after the feedback date were excluded from the analysis (N = 7). The students were able to use information and communication technologies (ICT) effectively.

3.2. Online Learning Environment

As the online learning environment, a course management system (CMS) developed by the researcher was used. The system was developed with Silverlight using C#.Net programming language. An MS SQL server was used as the database. The students were registered with the system at the beginning of the semester, and provided with a username and a password to log on to the system. Students used this environment to obtain course materials, participate in discussions about the class and to write reflections about the concepts that they learned in the class. The environment is able to record the learning activities of students in detail; however, only the reflection texts that were written by students were analysed in this study.

Figure 1: Online Learning Environment

During the first week of the class, the students were taught how to use the online learning environment. The students were informed that they were supposed to write reflections about the concepts they would learn in the class, participate in discussions, read posts written by their classmates and comment about them. The students were told that the reflections they would write in the online learning environment would be considered in their final grades; however, no limits or requirements were set for these activities. Each reflection written by the students was stored in the database of the learning environment. The text input fields where these reflections are written only allow text input, and it is limited to 500 characters. The copy and paste functions are disabled in these fields across the system. Moreover, when students copy and paste a text, the system automatically gives the warning, "Please do not use direct quotations." The students were also verbally warned not to use direct quotations in their texts, but to express their ideas in their own words.

3.3. Corpus

The corpus used in the research includes a total of 4,268 reflection texts. These were written over 14 weeks about 200 hardware concepts by 59 students. Posts with less than 50 characters in total were not included in the analysis (N=210). Of the posts, 2,342 (M=39.69 SD=22.03) were written before the feedback, and 1,795 (M=30.42 SD=27.06) were written after the feedback. While the posts written before the feedback had an average length of 211 characters, post length after the feedback averaged 248 characters.

Since it would be difficult to search for similarities manually on the internet for the number of posts written by students and to save the results to a database, a tool was developed to automate this process by using the Application Programming Interface (API) of the Bing™ search engine. When the tool is run, it retrieves the post from the database and sends the phrases in it to the search engine using API and if any results are found, it merges the top 20 results in a single text and it is recorded in the analysis database. This process is automatically repeated for all posts.

3.4. Similarity Analysis

Text mining analysis was done the middle and at the end of the semester as shown in Figure 2. During the first stage of the analysis, the texts are retrieved from the analysis database, which includes texts written by students and search results. Then the texts were preprocessed as follows:

- All punctuation is removed.
- All letters were converted to lowercase.
- All stop words (e.g., prepositions, conjunctions) were removed.
- Words under 2 or over 50 characters were removed.

Figure 2: The process of text mining analysis

After pre-processing, TF-IDF (Term Frequency–Inverse Document Frequency) scores were calculated. Then, the vectorial distances between the texts compared were calculated, and a similarity score was produced. There are a number of algorithms that can be used for intertextual similarity analysis. The cosine similarity algorithm was used in this study. Since cosine similarity works fast and is easy to interpret, it is widely used in text mining and information retrieval studies (Dhillon & Modha, 2001). To indicate the similarity between texts, cosine similarity generates similarity scores varying between 0 and 1. While a similarity score of zero indicates that two texts are entirely dissimilar, a score of 1 means they are exactly the same. The similarity analysis was done twice. The

first analysis was made at the sixth week. A total of 2,342 posts and their search results were compared. The resulting similarity scores were recorded to the analysis database automatically by the RapidMiner data mining tool. Since text mining analysis requires so much processing power and time, the results of the analysis were filtered, and values greater than or equal to 0.7 were recorded in the database. The same calculation was repeated for the 1,795 posts written by the students after the feedback was given.

3.5. Feedback

The similarity scores correspond to the plagiarism ratio of the text. For example, if the similarity ratio between a post written by a student and a post written by another student or a text obtained from an internet search result is 0.7, it shows that the plagiarism ratio in the post is 70%. By means of an add-on in the learning environment, the similarity analysis results stored in the database were posted as comments to the posts written by the students. For this purpose, the comment module that the students used to write comments on each other's posts in the online learning environment was used. The sample feedback presented in Figure 3 consists of the plagiarism ratio detected in the text along with the sender, the date sent and a brief message indicating that plagiarism had been detected.

Figure 3: An example of automated feedback (underlined text)

When a comment was written for students' posts, they received an automatic notification sent by the system to inform them that feedback had been provided.

3.6. Plagiarised Posts Ratio

To test whether the feedback provided to the students was effective in changing their plagiaristic behaviours, the plagiarised post ratios in the posts written before and after the feedback were compared. For each student, the plagiarised posts ratio (standardised plagiarism ratios) was calculated using the following formula.

$$Plagiarised\ posts\ ratio = \frac{Number\ of\ plagiarised\ posts}{Number\ of\ total\ posts} \times 100$$

Number of plagiarised posts: Number of posts with 70% and over plagiarism ratios.

Number of total posts: Number of total posts written until the designated date.

The plagiarised post ratios in the posts written before and after the feedback provided were calculated using the formula above at the end of the semester, and the results were recorded in the database for the further statistical analysis.

4. Results and Discussion

Descriptive statistics for the plagiarized post ratios before and after the feedback and their differences are shown in Table 1. Table 1 shows that a 21.07% average decrease was recorded in the plagiarised posts ratios after the feedback was given. When the percentiles of differences are examined, it is possible to see how this decrease is distributed among the group.

Table 1: Descriptive statistics for the plagiarised posts ratios before and after the feedback and their differences

As Table 1 shows, after the feedback, there was a 6% decrease of plagiarised post ratios in the posts written by 75% of the students. While there was at least an 18% decrease in 50% of the students, there was a decrease of 40% or more in 25%. In total, there was a decrease in the plagiarised post ratios in the posts written by 49 of 59 students (83%) after the feedback. Changes in students' plagiarised posts ratios before and after the feedback are shown in Figure 4.

Figure 4: Changes in the plagiarised posts ratios before and after the feedback

Table 2 displays the results of the paired samples t-test, which was done to determine whether the difference between the average plagiarised post ratios before and after the feedback was significant.

Table 2: Results of the paired samples t-test analysis

The result of the paired samples t-test shows that there is a statistically significant difference between the plagiarised post ratios for the posts written by the students before and after the feedback ($t=6.49$, $p=0.001$). The mean plagiarised post ratios for the posts written by the students after the feedback are lower than the mean plagiarised post ratios for the posts written before the feedback. In addition, the McNemar's Chi-square test was performed to test whether there was a statistically significant difference between the ratio of students who committed plagiarism before and after the feedback. Using the plagiarised post ratios, the students were divided into non-plagiarisers and plagiarisers. To transform the data into a categorical form, the cut-point was specified as 15%. Accordingly, the students whose plagiarised post ratio was between 0-15% were categorised as non-plagiarisers, and those whose ratio was between 15-100% as plagiarisers. The distribution of the students based on these categories before and after the feedback is presented in Table 3.

Table 3: Number and percentage of students who committed plagiarism before and after the feedback

As Table 3 shows, while there were only 11 (18.6%) students classified as non-plagiarisers before the feedback, this number increased to 36 (61.0%) after feedback was provided. The results of the McNemar's Chi-square test revealed a statistically significant difference between the non-plagiarising student ratios before and after the feedback was provided, (χ^2 (1, N=59) = 18.58, p<0.0001. In other words, the test showed that the 42.37% increase in the number of non-plagiarising students after the feedback was statistically significant.

5. Conclusions

Studies of anti-plagiarism software, which is widely used in educational institutions as a self-service that provides feedback to students, show that it is viewed positively by the staff and students, and that it also effectively reduces plagiarism (Whittle & Murdoch-Eaton, 2008). Stapleton (2012) noted that even knowing that such services would be used positively influenced students' plagiaristic behaviours. Moreover, Stappenbelt and Rowles (2010) reported that these services help students to develop a sense that avoiding plagiarism is important. When using such services, students either perform checks prior to handing in their assignments or receive feedback afterwards from their teachers.

In this study, problems with digital plagiarism were studied using an educational data mining approach. The effect of automated feedback generated by text mining on preventing plagiarism was investigated. This study focused on two types of plagiarism commonly used by students, plagiarism committed online and students copying each other's work. Its results indicate that individual feedback given to students led to a 21.07% average reduction in plagiarised post ratios. Furthermore, 83% of the students had lower plagiarised post ratios after the feedback. The number of non-plagiarising students went from 11 to 36, a 42.37% increase. The reduced numbers of plagiarised posts and the increased number of students who did not commit plagiarism show that the feedback had a positive effect on the student's plagiaristic behaviour. However, due to the limitations of the one-group pretest-posttest experimental design used in the study, it may not be argued that this decrease certainly resulted from the feedback. For this reason, further investigations are needed to confirm the results of this study by adding a control group that receives no feedback.

Although the students received written and verbal warnings not to commit plagiarism, 37.53% of the posts written before the feedback contained plagiarism at a ratio of 70% or higher. Furthermore, the posts written by 81.4% of the students had plagiarized post ratios ranging from 15% to 81%. This finding can be interpreted to mean that, as Ercegovac and Richardson (2004) claim, verbal and written warnings do so little to prevent plagiarism that they can be regarded as negligible.

The studies in the literature argue that one of the most important factors in the rise of digital plagiarism is the fact that information can be obtained directly from the internet or other sources by copying and pasting (Sisti, 2007). In this study, all the copy and paste functions were eliminated from the online learning environment. Moreover, when the students attempted to paste a text that was copied from somewhere else, it was automatically replaced by a warning that they should not use direct quotations. However, the fact that high plagiarism ratios were detected in the students' posts and a considerable proportion of the students committed plagiarism can be interpreted to mean that merely eliminating copying and pasting is not effective. Weiler (2005) claimed that the learning habits of the new generation of students, who he calls generation Y, are focused on obtaining information rather than reading and internalizing it. The students' tendency to retype information directly without interpreting it, even when the copy and paste functions were unavailable, supports this finding. Similarly, in a recent study, students were observed to commit plagiarism without copying and pasting (Kauffman & Young, 2015).

The companies that sell plagiarism detection services record the analysed texts to their databases for future use without the authors' permissions (Foster, 2002). Institutions that regard this as a copyright violation refuse to use such software (Brown, Robin, & Jordan, 2008). The method used in the study eliminates this problem. The database used for the similarity analysis includes the students' own writings and the search results generated by searching for similar texts on a search engine. RapidMiner software, which was used for the similarity calculations, is open-source and available at no cost. Integrating such systems into feedback mechanisms in online learning environments and making them easy for teachers to use can be an effective way to fight the plagiarism problem in online courses.

The most important limitation here is that text mining analyses require too much processing power and time. This can be a disadvantage in the large-scale learning environments where the number of students and texts may be too large. However, further studies can improve performance of algorithm for detecting text similarities to get faster results.

References

Akbulut, Y., Şendağ, S., Birinci, G., Kılıçer, K., Şahin, M. C., & Odabaşı, H. F. (2008). Exploring the types and reasons of Internet-triggered academic dishonesty among Turkish undergraduate students: Development of Internet-Triggered Academic Dishonesty Scale (ITADS). *Computers & Education, 51*(1), 463-473.

Arnold, K. E. (2010). Signals: Applying academic analytics. In *EDUCAUSE Quarterly* (Vol. 33): EDUCAUSE.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267-270). Vancouver, British Columbia, Canada: ACM.

Baron, J., & Crooks, S. (2004). Academic integrity in web based distance education. *TechTrends, 49*(2), 40-45.

Batane, T. (2010). Turning to Turnitin to Fight Plagiarism among University Students. *Educational Technology & Society, 13*(2), 1-12.

Bedford, W., Gregg, J., & Clinton, S. (2009). Implementing Technology to Prevent Online Cheating: A Case Study at a Small Southern Regional University (SSRU). *MERLOT Journal of Online Learning and Teaching, 5*(2), 230-238.

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. In. Washington, D.C.

Bolliger, D. U., Supanakorn, S., & Boggs, C. (2010). Impact of podcasting on student motivation in the online learning environment. *Computers & Education, 55*(2), 714-722.

Brown, V., Robin, N., & Jordan, R. (2008). A Faculty's Perspective and Use of Plagiarism Detection Software. In *Society for Information Technology & Teacher Education International Conference* (Vol. 2008, pp. 1946-1948).

Butakov, S., & Scherbinin, V. (2009). The toolbox for local and global plagiarism detection. *Computers & Education, 52*(4), 781-788.

Cavalcanti, E. R., Pires, C. E. S., Cavalcanti, E. P., & Pires, V. F. (2012). Detection and Evaluation of Cheating on College Exams using Supervised Classification. *Informatics in Education, 11*(2), 169-190.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning, 4*(5), 318-331.

Dhillon, I. S., & Modha, D. S. (2001). Concept Decompositions for Large Sparse Text Data Using Clustering. *Mach. Learn., 42*(1-2), 143-175.

Ercegovac, Z., & Richardson, J. V. (2004). Academic Dishonesty, Plagiarism Included, in the Digital Age: A Literature Review. *College & Research Libraries, 65*(4), 301-318.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5), 304-317.

Foster, A. L. (2002). Plagiarism-detection tool creates legal quandary. *The Chronicle of Higher Education, 48*(36), A37-A38.

Heberling, M. (2002). Maintaining academic integrity in online education. *Online Journal of Distance Learning Administration, 5*(1).

Hew, K. F., & Cheung, W. S. (2014). Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges. *Educational Research Review, 12*(0), 45-58.

Jocoy, C. L., & DiBiase, D. (2006). Plagiarism by Adult Learners Online: A case study in detection and remediation. *The International Review of Research in Open and Distributed Learning, 7*(1).

Karim, N. S. A., Zamzuri, N. H. A., & Nor, Y. M. (2009). Exploring the relationship between Internet ethics in university students and the big five model of personality. *Computers & Education, 53*(1), 86-93.

Kauffman, Y., & Young, M. F. (2015). Digital plagiarism: An experimental study of the effect of instructional goals and copy-and-paste affordance. *Computers & Education, 83*(0), 44-56.

Koulocheri, E., Soumplis, A., & Xenos, M. (2012). Applying Learning Analytics in an Open Personal Learning Environment: A Quantitative Approach. In *Informatics (PCI), 2012 16th Panhellenic Conference on* (pp. 290-295).

Maurer, H. A., Kappe, F., & Zaka, B. (2006). Plagiarism - A Survey. *J. UCS, 12*(8), 1050-1084.

Oberreuter, G., & Velásquez, J. D. (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications, 40*(9), 3756-3763.

Rolfe, V. (2011). Can Turnitin be used to provide instant formative feedback? *British Journal of Educational Technology, 42*(4), 701-710.

Rowe, N. C. (2004). Cheating in Online Student Assessment: Beyond Plagiarism. *On-line Journal of Distance Learning Administration, 7*(2).

Rushby, N. (2013). Editorial: Plagiarism. *British Journal of Educational Technology, 44*(1), 3-4.

Santos, J. L., Klerkx, J., Duval, E., Gago, D., Rodr, L., #237, & guez. (2014). Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 98-102). Indianapolis, Indiana: ACM.

Scanlon, P. M. (2003). Student Online Plagiarism: How Do We Respond? *College Teaching, 51*(4), 161-165.

Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254). Vancouver, British Columbia, Canada: ACM.

Sisti, D. A. (2007). How Do High School Students Justify Internet Plagiarism? *Ethics & Behavior*, *17*(3), 215-231.

Stapleton, P. (2012). Gauging the effectiveness of anti-plagiarism software: An empirical study of second language graduate writers. *Journal of English for Academic Purposes, 11*(2), 125-133.

Stappenbelt, B., & Rowles, C. (2010). The effectiveness of plagiarism detection software as a learning tool in academic writing education. In *4th Asia Pacific Conference on Educational Integrity (4APCEI)* (pp. 29).

Stephens, J. M., Young, M. F., & Calabrese, T. (2007). Does Moral Judgment Go Offline When Students Are Online? A Comparative Analysis of Undergraduates' Beliefs and Behaviors Related to Conventional and Digital Cheating. *Ethics & Behavior*, *17*(3), 233-254.

Tanes, Z., Arnold, K. E., King, A. S., & Remnet, M. A. (2011). Using Signals for appropriate feedback: Perceptions and practices. *Computers & Education, 57*(4), 2414-2422.

Watson, G., & Sottile, J. (2010). Cheating in the Digital Age: Do Students Cheat More in Online Courses? *Journal Of Computer Information System*, *13*(1).

Weiler, A. (2005). Information-Seeking Behavior in Generation Y Students: Motivation, Critical Thinking, and Learning Theory. *The Journal of Academic Librarianship*, *31*(1), 46-53.

Whittle, S. R., & Murdoch-Eaton, D. G. (2008). Learning about plagiarism using Turnitin detection software. *Medical Education*, *42*(5), 528-528.

Table 1: Descriptive statistics for the plagiarised posts ratios before and after the feedback and their differences

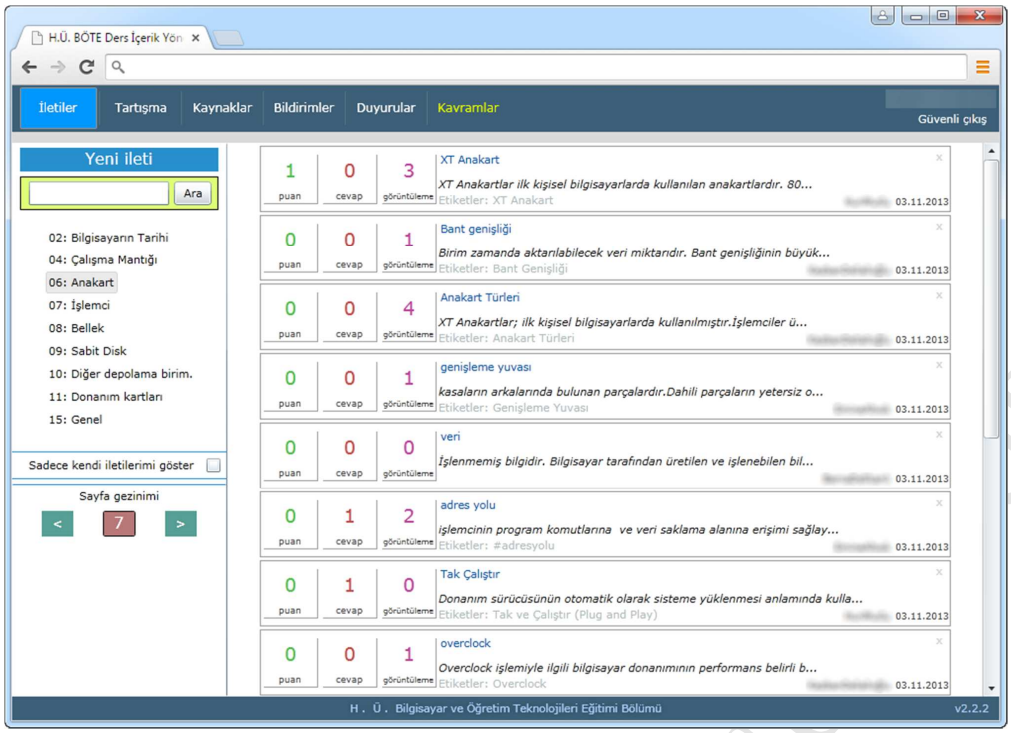
	N	Mean	Std.	Min	Max	Percentiles		
						25	50	75
Before feedback	59	37.53	21.73	.00	81.0	20.00	35.00	56.00
After feedback	59	16.46	15.94	.00	57.0	4.00	13.00	25.00
Differences	59	-21.07	24.92	-74.0	31.0	-40.00	-18.00	-6.00

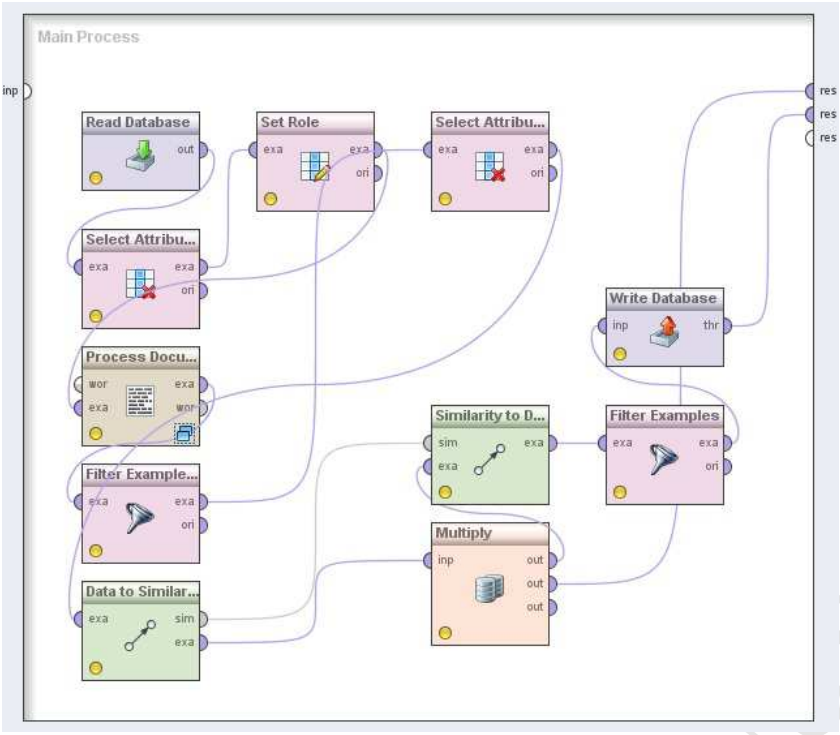
Table 2: Results of the paired samples t-test analysis

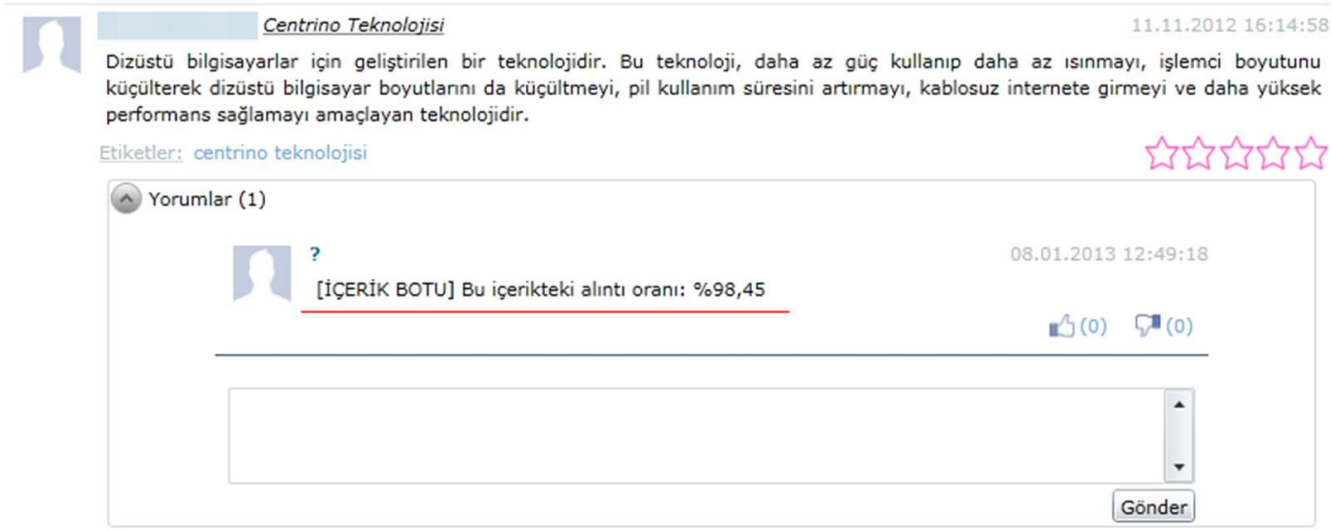
	Before feedback		After feedback		df	t	p
	$\bar{x}$	sd	$\bar{x}$	sd			
Plagiarised post ratio	37.53	21.73	16.46	15.94	58	6.49	0.001

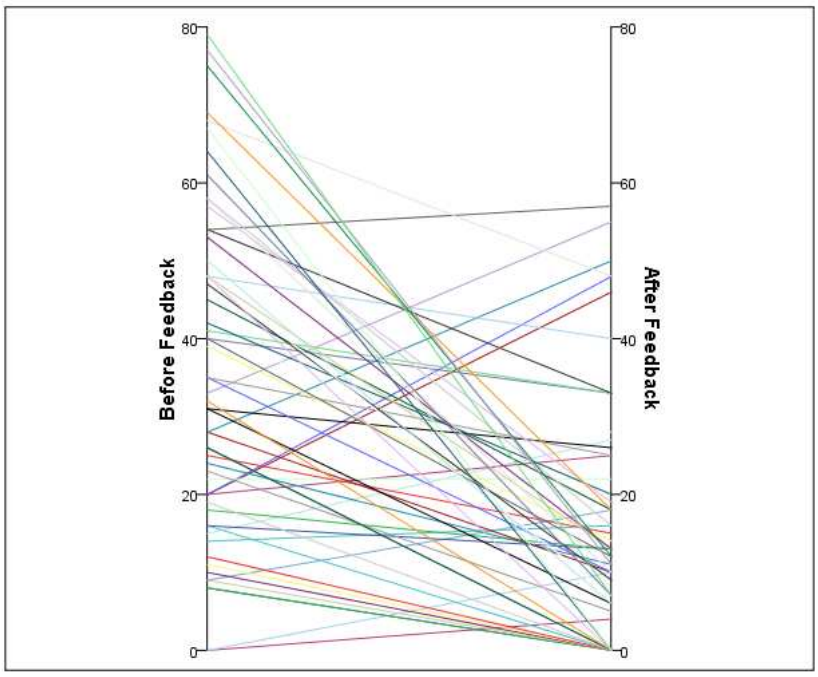
Table 3: Number and percentage of students who committed plagiarism before and after the feedback

	After feedback		
	Non – Plagiarisers	Plagiarisers	**Total**
Before feedback			
Non – Plagiarisers	8 (13.6%)	3 (5.1%)	**11 (18.6%)**
Plagiarisers	28 (47.5%)	20 (33.9)	**48 (81.4%)**
Total	**36 (61.0%)**	**23 (39.0%)**	**59 (100.0%)**









- The ability of feedback generated by text mining to prevent plagiarism was investigated.
- Before the feedback, 81.4% of the students committed plagiarism.
- Disabling “copy-paste” is not an efficient method by itself to prevent plagiarism.
- After the feedback, 83% of the students had reduced plagiarised post ratios.
- The number of plagiarised posts ratio went down 21.07% on average after the feedback.
- The number of non-plagiarisers increased by 42.37% after feedback was provided.