Predicting Academic Performance Based on Students' Blog and Microblog Posts

Mihai Dascalu[1(✉)], Elvira Popescu[2], Alexandru Becheru[2], Scott Crossley[3], and Stefan Trausan-Matu[1]

[1] Faculty of Automatic Control and Computers, University "Politehnica" of Bucharest, 313 Splaiul Independenței, 60042 Bucharest, Romania
{mihai.dascalu,stefan.trausan}@cs.pub.ro
[2] Faculty of Automation, Computers and Electronics, University of Craiova, 107 Bvd. Decebal, Craiova, Romania
popescu_elvira@software.ucv.ro, becheru@gmail.com
[3] Department of Applied Linguistics/ESL, Georgia State University, 34 Peachtree St. Suite 1200, Atlanta, GA 30303, USA
scrossley@gsu.edu

Abstract. This study investigates the degree to which textual complexity indices applied on students' online contributions, corroborated with a longitudinal analysis performed on their weekly posts, predict academic performance. The source of student writing consists of blog and microblog posts, created in the context of a project-based learning scenario run on our eMUSE platform. Data is collected from six student cohorts, from six consecutive installments of the Web Applications Design course, comprising of 343 students. A significant model was obtained by relying on the textual complexity and longitudinal analysis indices, applied on the English contributions of 148 students that were actively involved in the undertaken projects.

Keywords: Social media · Textual complexity assessment · Longitudinal analysis · Academic performance

1 Introduction

Automated prediction of student performance in technology enhanced learning settings is a popular, yet complex research issue [1, 2]. The popularity comes from the value of the predictive information which can be used for advising the instructor about students at-risk, who are in need of more assistance [3]. More generally, automated methods offer instructors the ability to monitor learning progress and provide personalized feedback and interventions to students in any performance state [4]. In addition, individualized strategies for improving participation may also be suggested [3]. Furthermore, a formative assessment tool could be envisaged based on the automatic prediction mechanism [3], which has the potential to decrease instructors' assessment loads [4]. Finally, students' awareness can be increased by providing them prediction results and personalized feedback [4].

K. Verbert et al. (Eds.): EC-TEL 2016, LNCS 9891, pp. 370–376, 2016.
DOI: 10.1007/978-3-319-45153-4_29

Performance prediction has been extensively studied in web-based educational systems and, in particular, in Learning Management Systems (LMS). This is due to the availability of large amounts of student behavioral data, automatically logged by these systems, such as: visits and session times, accessed resources, assessment results, online activity and involvement in chats and forums, etc. [2]. Thus, student performance prediction models based on Moodle log data have been proposed in multiple previous studies [5–7]. Additionally, log data from intelligent tutoring systems (ITS) have also been used for performance prediction [8]. In contrast, students' engagement with social media tools in emerging social learning environments has been less investigated as a potential performance predictor [9].

The current paper aims at analyzing students' contributions on social media tools (i.e., posts on blogs and Twitter) as potential predictors of academic performance. The context of the study is a collaborative project-based learning (PBL) scenario, in which students' communication and collaboration activities are supported by social media tools. Instead of relying only on quantitative usage data, similar to most previous studies, we explore the actual content of students' contributions by applying textual complexity analysis techniques. More specifically, we investigate how students' writing style in social media environments can be used to predict their academic performance. Multiple textual complexity indices (ranging from lexical, syntactical to semantic analyses [10, 11]) are used to create an in-depth perspective of students' writing style. We corroborate these findings with a longitudinal analysis performed on learners' weekly blog and microblog posts in order to obtain a more comprehensive view of academic performance prediction. The scale of our study is quite large, unfolding over the course of six years, as data is collected from six consecutive installments of the Web Applications Design (WAD) course comprising of 343 students. A preliminary study based on only one student cohort yielded encouraging results [12]; this paper is an extension of the pilot study, enriched also with longitudinal analysis of students' contributions.

Details about the study settings are presented in the following section, together with the data collection and preprocessing steps, as well as employed automated methods (textual complexity and longitudinal analysis indices). The results of our in-depth analysis are reported in Sect. 3, while conclusions are outlined in Sect. 4.

2 Methods

2.1 Data Collection and Preprocessing

Data was collected over 6 consecutive winter semesters (2010/2011 – 2015/2016), with 4th year undergraduate students in Computer Science from the University of Craiova, Romania. A total of 343 students, enrolled in the WAD course, participated in this study. A PBL scenario was implemented, in which students collaborated in teams of around 4 peers in order to build a complex web application of their choice. Several social media tools (wiki, blog, microblogging tool) were integrated as support for students' communication and

collaboration activities; all student actions on these social media tools were monitored and recorded by our eMUSE platform [13].

For the current study, the collected writing actions used to assess students' writing styles consisted of their tweets, together with blog posts and comments. The yearly distribution of students and of their social media contributions is presented in Table 1. We focused only on the content written in English. This content was cleaned of non-ASCII characters and spell-corrected. Finally, only students who had at least five English contributions after preprocessing and who used at least 50 content words were considered in order to meet the minimum content threshold needed for our textual complexity analysis. A content word is a dictionary word, not considered a stopword (common words with little meaning - e.g., "and", "the", "an"), which has as corresponding part-of-speech a noun, verb, adjective or adverb. Thus, a total of 148 students were included in our analysis, having cumulatively 3013 textual contributions.

Table 1. Distribution of students and contributions per academic year.

	Year 1 (2010-2011)	Year 2 (2011-2012)	Year 3 (2012-2013)	Year 4 (2013-2014)	Year 5 (2014-2015)	Year 6 (2015-2016)
Number of students	45	48	56	66	53	75
Number of blog posts & comments	166	121	318	1074	451	479
Number of tweets	326	181	1213	1561	956	1233

2.2 Textual Complexity Evaluation

In order to evaluate text complexity, we used the *ReaderBench* framework [10, 11] which integrates a multitude of indices ranging from classic readability formulas, surface indices, morphology and syntax, as well as semantics. In addition, *ReaderBench* focuses on text cohesion and discourse connectivity, and provides a more in-depth perspective of discourse structure based on Cohesion Network Analysis (CNA) [14]. CNA is used to model the semantic links between different text constituents in a multi-layered cohesion graph [15]. We refer readers to [10, 11] for further information about these features.

2.3 Longitudinal Analysis

We used one week as timeframe, due to the schedule of the academic semester in which students had one WAD class per week. The total length of the considered time series is 16 weeks, including 14 weeks of classes and 2 weeks for the winter holidays. For each student, the number of weekly blog and microblog posts was computed in order to obtain his/her time series of social media contributions. The performed longitudinal analysis relies on a wide range of evolution indices including average & standard deviation of contributions, entropy, uniformity, local extreme points, and average & standard deviation of recurrence. We refer readers to [16] for further information about these features that were initially used for keystroke analysis.

3 Results

We split the students into two equitable groups: high performance students with grades ≥ 8, while the rest were catalogued as low performance students. The indices from *ReaderBench* and from the longitudinal analysis that lacked normal distributions were discarded. Correlations were then calculated for the remaining indices to determine whether there was a statistical ($p < .05$) and meaningful relation (at least a small effect size, $r > .1$) between the selected indices and the dependent variable (the students' final score in the course). Indices that were highly collinear ($r \geq .900$) were flagged, and the index with the strongest correlation with course grade was retained, while the other indices were removed. The remaining indices were included as predictor variables in a stepwise multiple regression to explain the variance in the students' final scores in the WAD course, as well as predictors in a Discriminant Function Analysis [17] used to classify students based on their performance.

Medium to weak effects were found for *ReaderBench* indices related to word entropy, number of verbs, prepositions, adverbs, and pronouns, the number of unique words, number of named entities per sentence, and average cohesion between sentences and corresponding contributions measured with Latent Dirichlet Allocation [10] (see Table 2).

Table 2. Correlations between *ReaderBench* and longitudinal analysis indices, and course grade.

Index	*r*	*p*
Word entropy	.416	<.001
Time series entropy	.378	<.001
Average verbs per sentence	.323	<.001
Average cohesion (LDA) between sentences and corresponding contribution	−.274	<.010
Average unique words per sentence	.270	<.001
Average prepositions per sentence	.264	<.010
Time series local extremes	.236	<.010
Average adverbs per sentence	.236	<.010
Average pronouns per sentence	.250	<.010
Average named entities per sentence	.189	<.050

We conducted a stepwise regression analysis using the ten significant indices as the independent variables. This yielded a significant model, $F(3, 143) = 17.893$, $p < .001$, $r = .521$, $R^2 = .272$. Three variables were significant and positive predictors of course grades: word entropy, time series entropy and average verbs in sentence, denoting a higher activity and participation for high performance students. These variables explained 27 % of the variance in the students' final scores for the course.

The stepwise Discriminant Function Analysis (DFA) retained the same three variables as significant predictors of course performance (*Time series entropy* had the highest standardized canonical discriminant function coefficient), and removed the remaining variables as non-significant predictors. These three indices correctly allocated 108 of the 148 students from the filtered dataset, $\chi^2(df = 3, n = 148) = 43.543$ p < .001, for an

accuracy of 73.0 % (the chance level for this analysis is 50 %). For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 105 of the 148 students for an accuracy of 70.9 % (see the confusion matrix reported in Table 3 for results). The measure of agreement between the actual student performance and that assigned by the model produced a weighted Cohen's Kappa of .457, demonstrating moderate agreement.

Table 3. Confusion matrix for DFA classifying students based on performance

		Predicted Performance Membership		Total
		Low	High	
Whole set	Low	48	23	71
	High	17	60	77
Cross-validated	Low	48	23	71
	High	20	57	77

4 Conclusions

This paper investigated how students' writing style on social media tools, corroborated with the time evolution of their posts, can be used to predict their academic performance. Textual complexity and longitudinal analyses were performed on the blog and microblog posts of 148 (out of the total 343) students engaged in a project-based learning activity during 6 consecutive installments of the Web Applications Design course.

The analyses indicated that students who received higher grades in the course had greater word entropy, used more verbs, prepositions, adverbs, and pronouns, produced more unique words, and more named entities. Additionally, students who received higher grades had lower inner cohesion per contribution, indicating more elaborated paragraphs that represented a mixture of different ideas in the context of each contribution. The time series variables denote a more uniform distribution, with weekly fluctuations in terms of participation, which is normal for students that were more actively involved in using the social media tools. Three of these variables (word entropy, time series entropy and average verbs in sentence) were predictive of performance in both a regression analysis and a DFA.

The results are promising as several significant correlations and statistical models were identified in order to predict academic performance (i.e., course grades) based on textual complexity and longitudinal analysis indices. Additional experiments that will consider the learning style of each student, as well as an equivalent textual complexity model for Romanian language, are underway in order to augment the depth of our analyses. This will enable the consideration of a higher sample of students from the total of 343 course participants and will increase the power of the applied mechanisms.

Acknowledgments. This work was supported by the FP7 208-212578 LTfLL project, the 644187 EC H2020 RAGE project, and a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-2604.

References

1. Baker, R.S., Yacef, K.: The state of educational data mining in 2009: A review and future visions. J. Educ. Data Min. **1**(1), 3–17 (2009)
2. Romero, C., López, M.I., Luna, J.M., Ventura, S.: Predicting students' final performance from participation in on-line discussion forums. Comput. Educ. **68**, 458–472 (2013)
3. Yoo, J., Kim, J.: Can online discussion participation predict group project performance? investigating the roles of linguistic features and participation patterns. Int. J. Artif. Intell. Educ. **24**, 8–32 (2014)
4. Xing, W., Guo, R., Petakovic, E., Goggins, S.: Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory. Comput. Hum. Behav. **47**, 168–181 (2015)
5. Calvo-Flores, M.D., Galindo, E.G., Jiménez, M.P., Piñeiro, O.P.: Predicting students' marks from Moodle logs using neural network models. Curr. Dev. Technol. Assist. Educ. **1**, 586–590 (2006)
6. Romero, C., Ventura, S., Espejo, P.G., Hervás, C.: Data mining algorithms to classify students. In: 1st International Conference on Educational Data Mining, pp. 8–17. Quebec, Canada (2008)
7. Zafra, A., Ventura, S.: Predicting student grades in learning management systems with multiple instance genetic programming. In: 2nd International Conference on Educational Data Mining, pp. 309–319. Cordoba, Spain (2009)
8. Pardos, Z.A., Heffernan, N.T., Anderson, B., Heffernan, C.L.: The effect of model granularity on student performance prediction using bayesian networks. In: Conati, C., McCoy, K., Paliouras, G. (eds.) UM 2007. LNCS (LNAI), vol. 4511, pp. 435–439. Springer, Heidelberg (2007)
9. Giovannella, C., Popescu, E., Scaccia, F.: A PCA study of student performance indicators in a Web 2.0-based learning environment. In: 13th IEEE International Conference on Advanced Learning Technologies (ICALT 2013), pp. 33–35. IEEE, Beijing, China (2013)
10. Dascalu, M.: Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence, vol. 534. Springer, Cham (2014)
11. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining texts, learner productions and strategies with Reader Bench. In: Peña-Ayala, A. (ed.) Educational Data Mining: Applications and Trends, pp. 335–377. Springer, Cham, Switzerland (2014)
12. Popescu, E., Dascalu, M., Becheru, A., Crossley, S.A., Trausan-Matu, S.: Predicting student performance and differences in learning styles based on textual complexity indices applied on blog and microblog posts – a preliminary study. In: 16th IEEE International Conference on Advanced Learning Technologies (ICALT 2016). IEEE, Austin, Texas (in press)
13. Popescu, E.: Providing collaborative learning support with social media in an integrated environment. World Wide Web **17**(2), 199–212 (2014)
14. Dascalu, M., Trausan-Matu, S., McNamara, D.S., Dessus, P.: ReaderBench – automated evaluation of collaboration based on cohesion and dialogism. Int. J. Comput. Support. Collaborative Learn. **10**(4), 395–423 (2015)
15. Trausan-Matu, S., Dascalu, M., Dessus, P.: Textual complexity and discourse structure in computer-supported collaborative learning. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 352–357. Springer, Heidelberg (2012)

16. Allen, L.K., Jacovina, M.E., Dascalu, M., Roscoe, R., Kent, K., Likens, A., McNamara, D.S.: {ENTER}ing the time series {SPACE}: uncovering the writing process through keystroke analyses. In: 9th International Conference on Educational Data Mining (EDM 2016). International Educational Data Mining Society, Raleigh, NC (in press)
17. Klecka, W.R.: Discriminant analysis. Quant. Appl. Soc. Sci. Ser, **19**. Sage Publications, Thousand Oaks, CA (1980)