

# A Preliminary Analysis on Student Postings on Facebook and Blogs in an Internship Course in Information Management

Christy W. L. Cheong<sup>1</sup>, Xiao Hu<sup>2</sup>, Samuel K. W. Chu<sup>2</sup>, Nicole Ng<sup>2</sup>

<sup>1</sup>Macao Polytechnic Institute

<sup>2</sup>Faculty of Education, University of Hong Kong

## Abstract

This study analyzes student postings on Facebook and blogs in the same internship course across six years in an Information Management program. A two-level coding scheme on knowledge management and socio-emotional support was adapted to encode the postings. Association rule mining was applied to discover relationships between code categories; automated classification models were built and evaluated in various compositions of the categories. Preliminary findings disclosed that students posted messages of emotional expressions and social support, as well as those showing knowledge capturing and knowledge sharing processes. Preliminary results of classification experiments demonstrate the feasibility of automated analysis. It is also found that postings in the two platforms, Facebook and blogs, are mostly indistinguishable except for those in the social support categories.

**Keywords:** Automated classification; Facebook; blog; knowledge management; socio-emotional support

**Citation:** Cheong, C. W. L., Hu, X., Chu, S. K. W., & Ng, N. (2017). A Preliminary Analysis on Student Postings on Facebook and Blogs in an Internship Course in Information Management. In *iConference 2017 Proceedings* (pp. 517-524). <https://doi.org/10.9776/17225>

**Copyright:** Copyright is held by the authors.

**Contact:** wlcheong@ipm.edu.mo

## 1 Introduction

Empirical evidence has shown that social media, such as blogs and Facebook, were useful in facilitating knowledge sharing and social support building in experiential learning contexts including internships (Chau et al., 2013; Chu, Chan & Tiwari, 2012; Chu, Kwan & Warning, 2012). It is, however, challenging to closely monitor students' activities on these platforms given the amount of information posted. It would be helpful if automated means can be introduced to the monitoring process.

The use of (semi-)automated means has been proposed to analyze student textual inputs of various types (e.g., discussion forum posts, microblogs, essay writings) (Joksimović et al., 2015; McLaren, Scheuer, & Mikšátko, 2010; White & Larusson, 2015) against different learning indicators (e.g., complexity, dialog acts, writing quality) (Dascalu, 2014; Erkens & Janssen, 2008; McNamara, Crossley, & McCarthy, 2010), for the purpose of monitoring student online contributions and discussions. The spines within these methods are Machine Learning algorithms and Natural Language Processing (NLP) techniques (Manning & Schütze, 1999; McLaren, Scheuer, & Mikšátko, 2010; McNamara et al., 2014; Rosé et al., 2008). These techniques have also been applied to automated assessment of student textual responses to assignments, responses to short-answer questions, and essays. Haudek et al. (2012) presented an automated lexical analysis of the textual responses students provided in an assignment and argued that lexical analyses are helpful for teachers to uncover students' knowledge gaps. Pulman and Sukkarieh (2005) experimented with the use of machine learning algorithms to automate the marking of students' short free-text answers. Rosé et al. (2003) proposed an integrated method of text classification drawing on both symbolic and bag-of-words approaches in automating the assessment of qualitative essay answers. Despite the exciting progress in this area, there have been few studies on automated analysis of student input in the context of internship courses where students often work in different environments and gain diverse experiences. In this study, we aim to fill this

gap by applying existing techniques in Machine Learning and NLP to the examination of the feasibility of automated classification of students' textual input in social media in the internship context.

This study took place in a professional experience course in a Bachelor of Science program in Information Management (IM) offered by a university in Hong Kong, where students were required to take up internship posts locally or abroad in related industries during the summer after the third year. As IM skills are needed in virtually all sectors of the society nowadays, students in the IM program often work in different organizations, industries and locations, with different schedules. To facilitate learning, students are required to reflect and share their experience frequently, usually every other day. Besides, socio-emotional support is also important for the students as the internships are likely to be their first full-time working experience and they are mostly dispersed in different places. In this course, blogs and Facebook were used to create a common space for reflective writing and interaction despite geographical distances. The study explored the possibility of using automated means to help analyzing student postings for the purposes of knowledge management and socio-emotional support. It attempts to answer the following three research questions (RQs):

RQ1: Are there any relationships between the categories of postings made by individual students?

RQ2: To what extent can student postings be automatically categorized, and are there any categories more reliable than others in terms of automated classification?

RQ3: Can postings from blogs and Facebook be distinguishable using automated classification?

The study aims to contribute to a deeper understanding of the relationships between posting categories as informed by empirical evidence of a substantial size and over multiple years. The findings, which could help shed light on students' posting patterns, may have the potentials in providing both theoretical and practical implications for further research or enhanced teaching and learning practices. The study can provide empirical results on the feasibility of automated classification of student postings in the described context, which can eventually be of practical use in monitoring student learning. Empirical evidence would also be obtained on the similarities or differences between blogs and Facebook as platforms supporting experiential learning during internships, better informing choice of platforms and instructional design.

## 2 Data Collection and Analyses

The data analyzed in this study consisted of 4,499 postings (including the initial posts and the replies/comments thus triggered) from 109 students enrolling in the professional experience course from 2006 to 2008 and from 2011 to 2013 (see Table 1).

Platform	Year range	No. of students	No. of postings
Blog	2006-2008	45	1,774
Facebook	2011-2013	64	2,725
Total	-	109	4,499

Table 1. Descriptive Statistics of the Dataset

The data, i.e. students' individual postings, were annotated manually according to the two-level coding scheme adapted from Chan et al. (2013) on knowledge management and socio-emotional expression. The coding scheme of Chan et al. (2013) was developed through an iterative process of refinement with reference to the studies of Chu, Chan and Tiwari (2012), Du and Wagner (2007) and Hmelo-Silver (2003) on social and collaborative knowledge construction. Twenty percent of the postings were randomly selected from each year and double-annotated by a second independent coder to ensure coding quality. The inter-rater reliability was measured using Cohen's kappa ( $k$ ), and the results ( $k=0.797$ ) indicated a good level of agreement between coders (Altman, 1991). Figure 1 shows the structure of the coding scheme and the

category definitions while Table 2 shows the distribution of postings across categories. It is noteworthy that the distribution statistics shown in Table 2 include cases with multiple annotations where a posting is concurrently classified into more than one category.

Category	Blog	Facebook	Total
L1KC	413	886	1,299
L2KR	220	348	568
L2EC	196	546	742
L1KSD	364	901	1,265
L2KS	132	267	399
L2PQ	96	205	301
L2PF	170	570	740
L1KAA	35	71	106
L2KC	6	18	24
L2PS	29	53	82
L1EE	733	658	1,391
L2PE	540	476	1,016
L2NE	193	181	374
L1SS	250	231	481
TOTAL	1,795	2,747	4,542

Table 2. Distribution of Postings across Categories

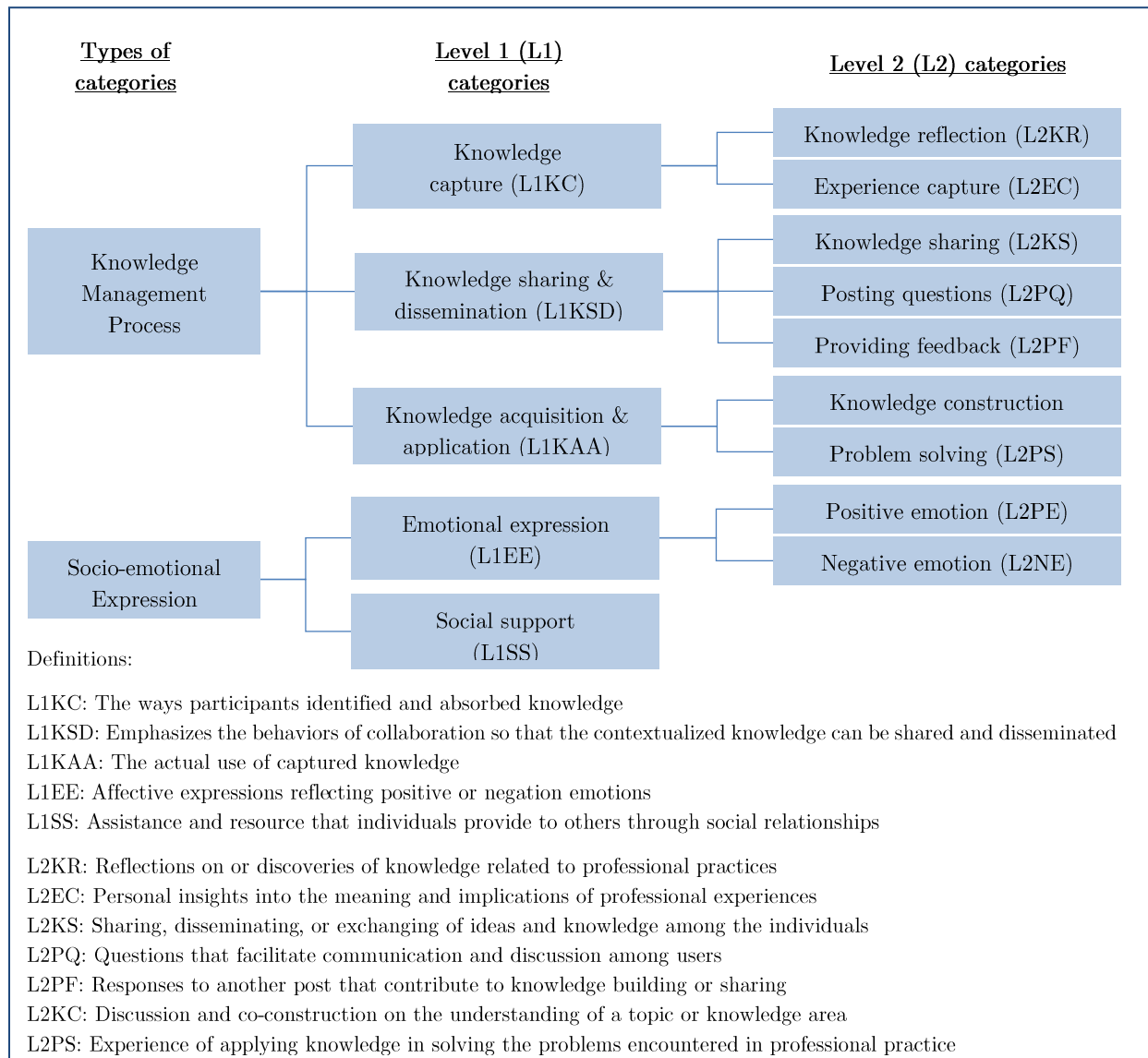


Figure 1. Coding scheme adopted from Chan et al. (2013).

In response to RQ1, association rule mining was conducted. We used individual students as the unit of analysis and the categories of their initial postings and comments as the attributes. We started out mining rules among the Level 1 categories. In light of the findings obtained, subsequent association rule mining processes were carried out to delve into patterns of the Level 2 categories. For RQ2 and RQ3, which are about automated classification of postings, the cases with multiple annotations were first removed from the dataset to simplify the problem at this stage of preliminary exploration. After that, three classification experiments were conducted using LightSide Researchers' Workbench, a text classification tool (Mayfield, Adamson, & Rosé, 2014). The classification performances were measured by accuracy (the ratio of correct predictions to the total number of data samples) and kappa (the level of agreement between manual and automated annotations). To test the generalizability of classification models, each experiment was conducted with a randomized 10-fold cross-validation.

### 3 Preliminary Results

#### 3.1 Relationships between Categories of Postings

Association rule mining was conducted using the FP-Growth algorithm (Han, 2012) to identify interesting and strong rules between the posting categories. The interestingness threshold was set on the level of ‘Lift’ larger than 1 and ‘Cosine’ larger than 0.65 (Merceron & Yacef, 2008).

The results showed that students having posted messages of categories from the socio-emotional domain (L1EE and L1SS) would also have messages reflecting knowledge capture (L1KC) and sharing knowledge (L1KSD). This finding is concordant with previous studies indicating that social interaction promotes cognitive interaction (Chan et al., 2013; Shea et al., 2014). Comparing the platforms, we found that the rules ‘L1KAA  $\rightarrow$  L1KC & L1KSD’ and ‘L1KAA  $\rightarrow$  L1EE & L1SS’ were not present among the blog users. This is probably because there were too few postings in the L1KAA category in the blogs. L1KAA (knowledge acquisition & application) refers to the actual use of captured knowledge. In particular, its subcategory ‘knowledge construction’ (L2KC), by definition, could only appear in discussions, and there were only six L2KC postings in blogs. This finding seems in line with the studies of Hu and Chu (2013) and of Chan et al. (2013) which suggested activities on blogs tended to be more individualistic and involve fewer interactions.

The rules ‘L1EE  $\rightarrow$  L1KSD & L1KC’ and ‘L1SS  $\rightarrow$  L1KSD & L1KC’ were mined from all three datasets: Facebook, blogs and combined. We thus went one step further by mining association rules among their respective Level 2 sub-categories. The results indicated that students posting messages of L1EE, no matter with positive (L2PE) or negative (L2NE) emotional expressions, would also post messages of L2EC (experience capture) and L2KS (knowledge sharing) and, among Facebook users, messages of L2PF (providing feedback) and L2PQ (posting questions) as well. Blog users posting messages of social support (L1SS) would also have messages of knowledge reflection (L2KR) whereas Facebook users would also have messages of L2EC (experience capture) and L2PF (providing feedback). These findings, again, suggested that two-way interactions, as implied by L2PF (providing feedback) and L2PQ (posting questions), were more frequent among Facebook users. This is also supported by the fact that there were more response postings in Facebook (15.9 replies per student) than blogs (8.6 replies per student).

#### 3.2 Automated Categorization

In response to RQ2 and RQ3, to test the feasibility of automated classification of student postings, the following classification experiments were conducted.

Experiment 1 was to determine the best performing combination of classification algorithms and text features. In this experiment, we compared three widely used classification models, namely Naïve Bayes, Logistic Regression, and Support Vector Machines (SVM) with two feature sets: basic text features with or without stopwords removed. The basic text features include: normalized counts of unigrams, bigrams and trigrams of the words contained in each posting, bigrams and trigrams of the Part-of-Speech (POS), word/POS pairs, line length, and punctuations. Stemming was applied to combine words with the same roots (e.g., ‘computer’ and ‘computing’) to reduce the dimensionality of the feature space and thus alleviate the problem of data sparseness.

The dataset used in Experiment 1 consisted of equal number of postings randomly selected from each of the five Level 1 categories. In comparing performances, the basic text features without removing stopwords performed better across algorithms while Naïve Bayes (Avg. Accuracy = 0.604; Avg. Kappa = 0.505) and Logistic Regression (Avg. Accuracy = 0.601; Avg. Kappa = 0.501) seemingly have better performance as compared to SVM (Avg. Accuracy = 0.582; Avg. Kappa = 0.477). The performances show effectiveness in classification as the accuracy of a random prediction baseline would be only 20%. The kappa values also indicate moderate levels of agreement with human annotations. The performances of Naïve

Bayes and Logistic Regression showed no significant difference based on a Student's  $t$ -test ( $p > 0.05$ ). As Logistic Regression had more stable performances across multiple rounds of testing, the combination of Logistic Regression and basic text features without removing stopwords was thus used in subsequent classification experiments.

To test the feasibility of automated classification of student postings on a larger scale and on both category levels, Experiment 2 was conducted on a dataset consisting of the categories with larger sample sizes: L1KC (knowledge capture,  $n=1,299$ ), L1KSD (knowledge sharing and dissemination,  $n=1,265$ ) and L1EE (emotional expressions,  $n=1391$ ) on Level 1; and L2KR (knowledge reflection,  $n=568$ ), L2EC (experience capture,  $n=742$ ), L2KS (knowledge sharing,  $n=399$ ), L2PQ (posting questions,  $n=301$ ) and L2PF (providing feedback,  $n=740$ ) on Level 2. As the datasets were imbalanced, kappa was used to measure the quality of automated classification.

The results suggested that automated classification was more promising with Level 1 categories (with an overall kappa value  $k=0.639$ , indicating a good level of agreement) than with Level 2 categories (with an overall kappa value  $k=0.432$ , indicating a moderate level of agreement) (Altman, 1991). Among Level 1 categories, the performances of L1EE ( $k=0.736$ ) and L1KC ( $k=0.607$ ) were quite good, with kappa values in the 'good' agreement level with human coders. L1KSD ( $k=0.509$ ) had the worst performance which was in 'moderate' agreement with human coders (Altman, 1991). The results once again verify that automated categorization is feasible for Level 1 categories. Among Level 2 categories, the classifier worked well with L2PQ (posting questions;  $k=0.840$ ) whose kappa values indicates 'very good' agreement with human coders (Altman, 1991). This probably benefits from its unique linguistic nature (e.g., question marks). L2KS (knowledge sharing), however, had the worst performance ( $k=0.00$ ). A closer examination on misclassified samples revealed that around 82% of the 214 postings in L2KS were misclassified as L2EC (experience capture, 42.6%) or L2KR (knowledge reflection, 39.8%). As L2KS is about half to one-third of the sizes of L2EC and L2KR, it is possible that its performance is affected by the insufficient amount of examples. Additionally, the high misclassification rates of L2KS as L2EC and L2KR suggested the possibility that L2KS postings might not be linguistically distinguishable from the postings of the other two categories.

Experiment 3 was conducted to test whether postings can be distinguishable between the platforms across Level 1 categories. One binary classification model was built for each of the Level 1 categories to classify whether the postings were from blogs or Facebook. As the datasets of several categories are highly imbalanced (with much more samples from Facebook than blogs), random sampling was applied to the larger set to balance the size of postings from the two platforms, which made the baseline accuracy of random selection as 50%. The results show that L1SS attained an average accuracy of above 70% and the agreement of human and machine annotations ( $k$ ) reached a moderate level in average ( $k=0.465$ ) (Altman, 1991), while performances on other categories stayed close to the baseline. This indicates that postings expressing social support were more likely to be distinguishable between platforms by linguistic features, suggesting the possibility that social support patterns vary between platforms. Similarly, the situation that the knowledge management postings on the two platforms were less distinguishable by automated classifiers may be due to linguistic similarity – the knowledge management processes were more or less expressed in the same words and phrases despite platform differences.

## 4 Conclusion and Future Work

To explore the possibility of automating the analysis of student postings in social media during experiential learning processes, this study collected six years of student postings on blogs and Facebook in an internship course in a Bachelor program in Information Management. These postings were exploited via association rule mining to understand the relationships between categories of practices in knowledge management and socio-emotional support. The preliminary results (1) showed that students posting socio-emotional messages

would also be engaged in reflection and knowledge sharing, and (2) conformed to previous research indicating blog activities tend to be less interactive than those on Facebook. These findings are of referential values in enhancing the instructional design of using social media to support experiential learning.

Given the substantial size and longitudinal nature of the dataset, this study attempted to evaluate the feasibility of using machine learning and NLP techniques to automatically classify student postings. The preliminary results generally supported the viability of automated classification of postings, especially those expressing emotions and posting questions. Impediments were observed in classifying messages indicating knowledge sharing practices, which may not be linguistically distinguishable from messages indicating experience capture and knowledge reflection. This study also found that blog and Facebook postings are more distinguishable via automated means in the social support domain than the knowledge management domain. It evidences that social support patterns varied between postings on the two platforms while knowledge management patterns persisted. In the next stage, we will conduct feature analyses to further the understanding of misclassifications and confusions between categories, which may potentially be helpful for enhancing our understanding on the natures of the postings, as well as improving classification performances.

## 5 References

- Altman, D. G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.
- Chan, R. C. H., Chu, S. K. W., Lee, C. W. Y., Chan, B. K. T., & Leung, C. K. (2013). Knowledge management using social media: A comparative study between blogs and Facebook. In *Proceedings of the ASIS&T 2013 Annual Meeting* (Montreal, Quebec, Canada). Retrieved October 17, 2015 from <http://web.hku.hk/~samchu/docs/Chan-2013-Knowledge-management-using-social-media.doc>.
- Chau, K. L. C., Kan, C. Y. J., Wong, C. M. J., Siu, F. L. C., Chu, S. K. W., & Law, J. (2013). Effective use of Facebook on knowledge transfer in a professional experience internship: A case-based approach to analyze influencing factors. Paper presented at the CITE Research Symposium 2013, The University of Hong Kong, Hong Kong. Retrieved October 17, 2015 from <http://web.hku.hk/~samchu/docs/Chau-2013-Effective-Use-of-Facebook-on-Knowledge-Transfer-in-a-Professional-Experience-Internship.doc>.
- Chu, S. K. W., Chan, C. K. K., & Tiwari, A. F. Y. (2012). Using blogs to support learning during internship. *Computers & Education*, 58(3), 989-1000.
- Chu, S. K. W., Kwan, A. C. M., & Warning, P. (2012). Blogging for information management, learning, and social support during internship. *Journal of Educational Technology & Society*, 15(2), 168-178.
- Dascalu, M. (2014). Analyzing discourse and text complexity for learning and collaborating. In *Studies in Computational Intelligence*. Switzerland: Springer.
- Du, H. S., & Wagner, C. (2007). Learning with weblogs: Enhancing cognitive and social knowledge construction. *IEEE Transactions on Professional Communication*, 50(1), 1-16.
- Erkens, G., & Janssen, J. (2008). Automatic coding of dialogue acts in collaboration protocols. *International Journal of Computer-Supported Collaborative Learning*, 3(4), 447-470.
- Han, J. (2012). *Data mining: Concepts and techniques*. Waltham, Mass.: Morgan Kaufmann Publishers.
- Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid-base chemistry in introductory biology. *CBE-Life Sciences Education*, 11(3), 283-293.
- Hmelo-Silver, C. E. (2003). Analyzing collaborative knowledge construction: multiple methods for integrated understanding. *Computers & Education*, 41(4), 397-420.

- Hu, X., & Chu, S. K. W. (2013). A comparison on using social media in a professional experience course. Paper presented at the International Conference on Social Media and Society 2013, Halifax, Canada. Retrieved October 17, 2015 from <http://ccmir.cite.hku.hk/pubs/SM&S2013.pdf>.
- Joksimović, S., Zouaq, A., Kovanović, V., Gašević, D., Jovanović, J., & Hatala, M. (2015). What do cMOOC participants talk about in Social Media? In *Proceedings of the International Learning Analytics and Knowledge conference*. New York: Poughkeepsie.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA.: MIT Press.
- Mayfield, E., Adamson, D., & Rosé, C. P. (2014). *LightSide Researchers' Workbench: User manual*. Retrieved November 12, 2015 from [http://ankara.lti.cs.cmu.edu/side/LightSide\\_Researchers\\_Manual.pdf](http://ankara.lti.cs.cmu.edu/side/LightSide_Researchers_Manual.pdf).
- McLaren, B. M., Scheuer, O., & Mikšátko, J. (2010). Supporting collaborative learning and e-Discussions using artificial intelligence techniques. *International Journal of Artificial Intelligence in Education*, 20(1), 1-46.
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of writing quality. *Written Communication*, 27, 57-86.
- McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Matrix*. Cambridge: Cambridge University Press.
- Merceron, A., & Yacef, K. (2008). Interestingness measures for associations rules in educational data. *Educational Data Mining*, 8, 57-66.
- Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp.9-16). Association for Computational Linguistics.
- Rosé, C. P., Roque, A., Bhembe, D., & Vanlehn, K. (2003). A hybrid text classification approach for analysis of student essays. In *Proceedings of the HLT-NAACL 03 workshop on Building Educational Applications Using Natural Language Processing* (vol.2; pp.68-75). Association for Computational Linguistics.
- Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-supported Collaborative learning*, 3(3), 237-271.
- Shea, P., Hayes, S., Uzuner-Smith, S., Gozza-Cohen, M., Vickers, J., & Bidjerano, T. (2014). Reconceptualizing the community of inquiry framework: An exploratory analysis. *The Internet and Higher Education*, 23, 9-17.
- White, B., & Larusson, J. A. (2015). Chapter 8: Identifying Points for Pedagogical Intervention Based on Student Writing: Two Case Studies for the "Points of Originality". In J. A. Larusson, & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp.156-190). Springer.