Sentiment analysis on educational datasets: a comparative evaluation of commercial tools

FOTEINI S. DOLIANITI[1], DIMITRIOS IAKOVAKIS[2], SOFIA B. DIAS[3], SOFIA J. HADJILEONTIADOU[4], JOSE A. DINIZ[3], GEORGIA NATSIOU[1], MELPOMENI TSITOURIDOU[1], PANAGIOTIS D. BAMIDIS[5], LEONTIOS J. HADJILEONTIADIS[2,6]

[1]Department of Early Childhood Education
Aristotle University of Thessaloniki
Greece
dolianiti@nured.auth.gr, gnatsiou@nured.auth.gr, tsitouri@nured.auth.gr

[2]Department of Electrical & Computer Engineering
Aristotle University of Thessaloniki
Greece
dimiiako12@gmail.com, leontios@auth.gr

[3]Faculdade de Motricidade Humana
Universidade de Lisboa
Portugal
sbalula@fmh.ulisboa.pt, jadiniz@fmh.ulisboa.pt

[4]Department of Primary Education
Democritus University of Thrace
Greece
shadjileontiadou@gmail.com

[5]Lab of Medical Physics
School of Medicine
Aristotle University of Thessaloniki
Greece
pdbamidis@gmail.com

[6]Department of Electrical & Computer Engineering
Khalifa University of Science and Technology
UAE
leontios.hadjileontiadis@ku.ac.ae

ABSTRACT

Sentiment analysis systems have been gaining increasing popularity for extracting information regarding students' affective state. Developing such systems from scratch is a challenging task, thus, many studies employ commercial, general-purpose tools that are not domain-specific. The aim of the present work is to comparatively evaluate the performance of five well-known *commercial/academic sentiment analysis tools on two educational datasets and contrast it with the performance of educational domain-specific tools, at document and sentence level. Findings suggest that: a) different tools work better for specific datasets and analysis levels, b) depending on the dataset, a general-purpose tool might be a viable solution, and c) any method, domain-specific or general-purpose one, should be evaluated before employed.*

KEYWORDS
Sentiment analysis, opinion mining, education, commercial systems, benchmark

RÉSUMÉ

Les systèmes d'analyse de sentiment, en extrayant des informations relatifs à l'état affectif des élèves, gagnent, de plus en plus, en popularité. Le développement de tels systèmes à partir de zéro est une tâche difficile si bien que de nombreuses études utilisent des outils commerciaux d'usage général, outils non spécifiques à un domaine d'application. Le but de cette étude est a) d'évaluer comparativement la performance de cinq outils bien connus qui analysent des sentiments commerciaux / académiques sur deux ensembles de données éducatives et b) de la comparer aux performances des outils pédagogiques spécifiques, au niveau des documents et des phrases. Les résultats suggèrent que : a) différents outils fonctionnent mieux pour des ensembles de données et des niveaux d'analyse spécifiques, b) selon l'ensemble des données, un outil à usage général pourrait être une solution viable, c) toute méthode, spécifique à un domaine ou non, doit être évaluée avant d'être utilisée.

MOTS-CLÉS
Analyse de sentiment, analyse d'opinion, éducation, systèmes commerciaux, benchmark

INTRODUCTION

In their affective learning manifesto, Picard and her colleagues claimed that machines, even without a fully-fledged theory of affect, can integrate capabilities to recognize emotions, helping us to study and support the affective dimension of learning in ways that were not previously possible (Picard et al., 2004). Recently, sentiment analysis has been gaining increasing popularity as a technique for extracting information regarding students' affective state, becoming an example of how technology can provide alternate ways of emotion measurement other than psychological tools.

Sentiment analysis is an umbrella term covering a large problem space that includes many related -yet, slightly different- tasks (Liu, 2015), such as subjectivity analysis, (i.e., to distinguish between sentences that present factual information and sentences that present opinions and evaluations (Wiebe, Bruce & O'Hara, 1999), emotion classification (i.e., to detect discrete emotions in text, such as happiness, boredom, frustration, and sadness), and polarity classification (i.e., to determine the valence of a text as positive, negative or neutral). The term "*sentiment analysis*" is most commonly used to refer to the latter task of polarity classification (Mohammad, 2016), which is the focus of the present work, as well.

In the educational domain, the potentialities of sentiment analysis mainly arise from two points: a) the capability to automatically analyze even vast amounts of text data, that would otherwise be difficult, labor and time-intense to handle, and b) the opportunity to unobtrusively record and study emotions through the behavioral traces of students without interrupting the learning process. These potentialities have been explored at multiple levels, such as at the classroom-level and the institutional-level, for analyzing text data produced within various teaching-learning contexts -ranging from face-to-face to fully online delivery modes, and from formal to informal settings- as well as through various means of communication, such as synchronous interactions in chats and asynchronous interactions in forums (e.g., Altrabsheh, Gaber, & Cocea, 2013; Colace, de Santo & Greco, 2014; Kagklis et al., 2014; Santos, Lechugo & Silveira-Mackenzie, 2016; Tian et al., 2014; Zarra, Chiheb, Faizi & El Afia, 2016). Instruction evaluation, institutional decision and policy making, enhancement of intelligent information/learning systems, assignment evaluation and feedback improvement, as well as new research insights are among the different task types that sentiment analysis has served within the educational domain so far (Dolianiti et al., 2018).

From the aforementioned it follows that apart from the technical perspective of developing effective sentiment analysis models -which is mainly the focus of Natural Language Processing (NLP) researchers- the practical implementations of these techniques concern diverse groups of stakeholders, including educators and social sciences researchers. On one hand, these groups may not necessarily have the expertise to develop custom sentiment analysis models, one the other hand the very nature of the technical problems that sentiment analysis poses is very challenging, even for the NLP community itself. Complexity and subtlety of language use, creative and non-standard language phenomena (e.g., sarcasm, metaphors, misspellings, abbreviations), and lack of paralinguistic information are some of the challenges to automatically detecting sentiment in text (Mohammad, 2016). Additionally, most successful techniques require ground-truth labels to be assigned to big training datasets, which is a highly costly process (Zhou, 2017). As a result, commercial and/or academic sentiment analysis tools are often utilized as an "*off-the-shelf*" solution for incorporating sentiment scores into various applications tasks.

These popular tools tend to be treated as black boxes and are accepted as valid methods for detecting sentiments without any investigation on their suitability to specific contexts and applications (Ribeiro et al., 2016). However, as sentiment analysis techniques and applications are interconnected, the performance of the former defines the effectiveness of the latter. In the work of Zimbra, Abbasi, Zeng and Chen (2018), for example, top academic and commercial sentiment analysis tools were applied in a Twitter event detection case study and findings suggested that application results (i.e., ability to detect events) correlated with sentiment classification performance. Hence, as commercial, general-purpose tools are increasingly being incorporated into educational research and practice, there is a need to assess their effectiveness in correctly recognizing students' positive, negative and neutral sentiments.

The present work shifts focus from the ways in which sentiment outputs are or can be used for monitoring, studying and making sense of the learning process to whether these sentiment outputs are reliable in the first place. More specifically, the aim of the study is to comparatively evaluate the performance of five well-known commercial/academic sentiment analysis tools on educational datasets and contrast it with the performance of educational domain-specific tools, at document and sentence level.

METHODOLOGY

Datasets

Comparative evaluation of sentiment analysis tools was performed against two datasets, comprised of student forum posts in Moodle Learning Management System. Data were collected across a semester from two courses offered by a public Higher Education Institution (Aristotle

University of Thessaloniki) in Greece. The first dataset was drawn from a postgraduate course related to affective computing and learning while the second was drawn from an undergraduate course focused on teacher education in educational robotics.

In the experiments, each dataset was used in two different versions. In the first version (document-level) each dataset instance was a student post while in the second version (sentence-level) each instance was a student post's sentence. Thus, both versions included the same text content, however, this content was fed into sentiment analysis tools at a different level. Table 1 presents the dataset size and class distribution for each course, both at document-level and sentence-level.

TABLE 1
Dataset size and class distribution, at document and sentence level

Dataset	Analysis level	Size	Class distribution		
			positive	negative	neutral
Affective Computing & Learning	Document	201	50%	15%	35%
	Sentence	881	34%	13%	53%
Educational Robotics	Document	129	30%	26%	44%
	Sentence	383	33%	25%	43%

As student posts were originally in Greek, the datasets were translated into English using the Google Translate machine translation tool. The translated posts were then manually corrected for errors. Previous studies have demonstrated that machine translation systems have reached a level of maturity that allows their integration in multilingual sentiment analysis, and machine translated datasets can yield similar results as their corresponding native-speaker translations (Balahur et al., 2011).

The datasets were independently labeled by two annotators, both at document-level and sentence-level. In order to assess the quality of the labels produced, percent agreement and Krippendorff's alpha (Hayes & Krippendorff, 2007) were calculated and are provided in Table 2. As the inter-rater reliability for the Educational Robotics dataset was not satisfactory at document level, a third reviewer labeled those text instances for which agreement was not originally reached, and a final label was assigned using majority rule. When majority was not reached, neutral sentiment was assigned.

TABLE 2
Inter-rater reliability for each dataset, at document and sentence level

Dataset	Analysis level	Percent agreement	Krippendorff's alpha
Affective Computing & Learning	Document	82.7%	0.753
	Sentence	86.8%	0.773
Educational Robotics	Document	71.3%	0.588
	Sentence	79.6%	0.747

Sentiment analysis tools

This subsection presents the five well-known general-purpose sentiment analysis tools employed in this study as well as the basic architecture and development procedure of the educational domain-specific models.

Table 3 summarizes the main information regarding the five commercial and academic tools. Tool names and references are given in column 1. Column 2 provides a short description while Column 3 reports the output that each tool generates. As column 3 shows, the output form varied across tools. Therefore, these outputs had to be handled in order to get a single class label (i.e., positive, negative, neutral) for each text instance. Output handling for each tool is specified in Appendix I. When tools output a float number, class cut-off values were set. For each tool, multiple cut-offs were tested so as to optimize performance and avoid near-misses, i.e., texts incorrectly classified into a neighboring class (Taboada et al., 2011). In IBM Watson Natural Language Understanding, float scores were considered instead of output labels, as custom cut-off values produced better results. Finally, there existed cases where tools could not analyze some input texts, considering them as undefined. These texts were assigned the neutral class as default.

TABLE 3

Short description and output form of the five commercial/academic tools employed

Tool	Description	Output
Repustate[1]	An API-based commercial tool supporting multiple languages, multiple levels of analysis, and configuration of sentiment rules. It, also, offers the facility to use the API with zero coding via spreadsheet upload.	Float scores ranging from -1 (negative) to 1 (positive) with a score of 0 being neutral.
Microsoft Azure Text Analytics API[2]	An API-based commercial tool supporting multiple languages. Analysis is performed on the input text as a whole, using machine learning techniques. The classification features include n-grams, features generated from part-of-speech tags, and word embeddings.	Float scores between 0 (negative) and 1 (positive)
IBM Watson Natural Language Understanding[3]	An API-based, commercial tool supporting multiple languages. Analysis is performed on the input text as a whole or towards specific targets.	Float scores ranging from -1 (negative) to 1 (positive) as well as three-class labels (positive, negative, neutral)
Sentistrength (Thelwall et al., 2010)	A lexicon-based tool that incorporates sentiment rules to handle linguistic phenomena, such as idioms, repeated punctuation and emoticons. It is available in two versions, a commercial Java version and a free, academic Windows version.	Each text instance is assigned two ordinal scores: i) a positive score from 1 to 5, and ii) a negative score from -1 to -5

Educational domain-specific models were developed from the datasets presented in the previous subsection, using Microsoft Azure's Machine Learning Studio[4]. Specifically, two models were developed for each course, one at document-level and one at sentence-level, resulting in four

[1] https://www.repustate.com/sentiment-analysis/

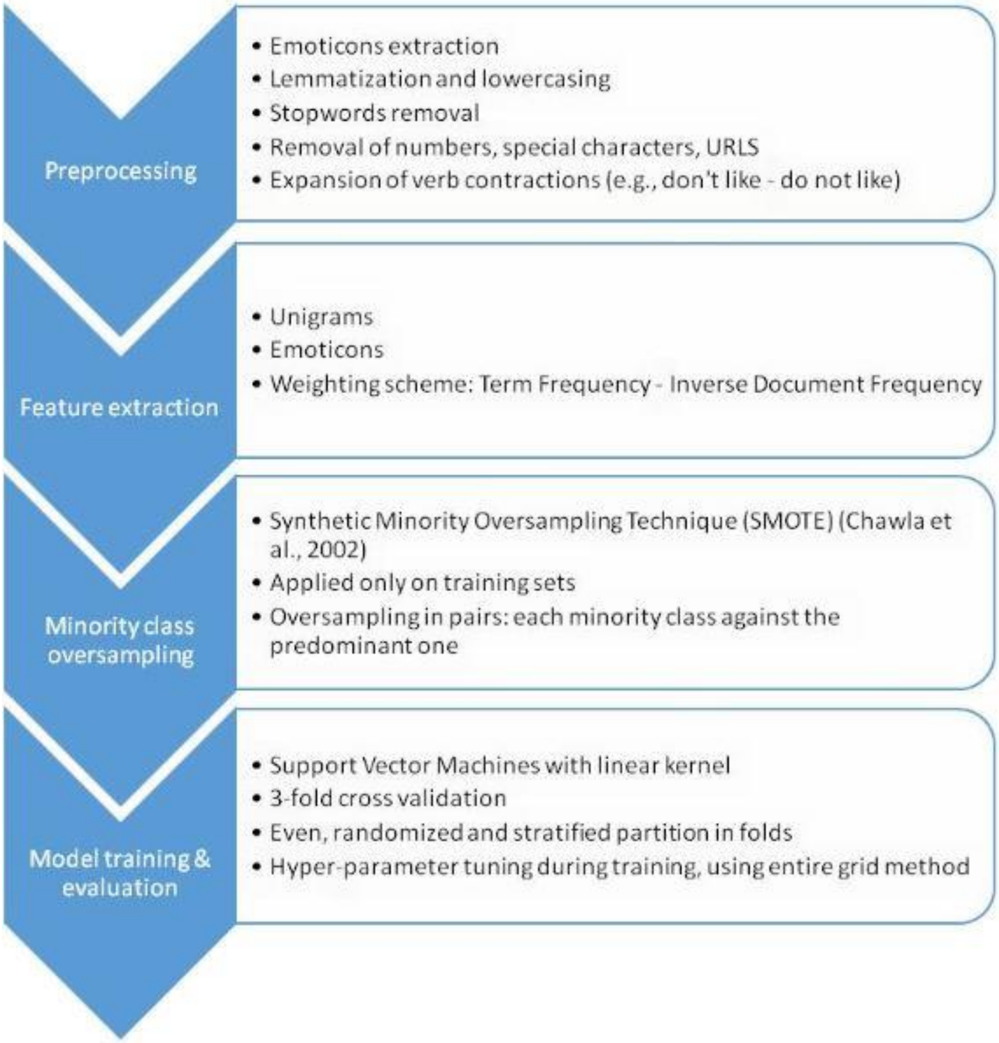
[2] https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

[3] https://www.ibm.com/watson/services/natural-language-understanding/

[4] https://studio.azureml.net/

models in total. All four models shared the same development workflow, comprised of four main steps, presented in Figure 1.

FIGURE 1


Preprocessing
• Emoticons extraction
• Lemmatization and lowercasing
• Stopwords removal
• Removal of numbers, special characters, URLS
• Expansion of verb contractions (e.g., don't like - do not like)
Feature extraction
• Unigrams
• Emoticons
• Weighting scheme: Term Frequency - Inverse Document Frequency
Minority class oversampling
• Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002)
• Applied only on training sets
• Oversampling in pairs: each minority class against the predominant one
Model training & evaluation
• Support Vector Machines with linear kernel
• 3-fold cross validation
• Even, randomized and stratified partition in folds
• Hyper-parameter tuning during training, using entire grid method

Workflow of the educational domain-specific models development

RESULTS

Benchmark evaluation results are reported using the F-measure metric. F-measure values are given both for each individual class as well as on average, as an indicator of the overall performance.

Table 4 presents classification results for the Affective Computing and Learning dataset, both at document and sentence level, with "*+tive*", "*-tive*", and "*avg*" indicating positive sentiment, negative sentiment and average F-score, respectively. On the whole, poor performance was observed, especially for negative sentiment. Comparing general-purpose tools with the educational-domain specific one, the latter outperformed at both levels of analysis as well as across all three classes. Difference between the educational domain-specific model and the second best tool reached 11.3% at document level and 15.4% at sentence level.

Among the commercial and academic tools, Microsoft Azure Text Analytics API exhibited the best average performance at document level, followed by IBM Watson Natural Language Understanding. With the exception of OpinionFinder 2.0, all tools showed a bias towards positive sentiment. Regarding the two other classes, Microsoft Azure Text Analytics API was better at recognizing neutral sentiment while Repustate was more successful with negative sentiment. In fact, Repustate's score against negative class approached the corresponding score of the domain-specific tool.

As far as sentence-level is concerned, Sentistrength showed the highest performance boost of almost 7%, as bias towards positive class was reduced and capability to recognize neutral sentiment was significantly improved. Reduced positive bias and more balanced classification results were, also, observed for IBM Watson Natural Language Understanding and Microsoft Azure Text Analytics API, although latter's overall score was slightly higher at document level. Opinion Finder 2.0 and Repustate did not seem to benefit from the lower analysis level. In fact, bias towards neutral sentiment was strengthened in OpinionFinder 2.0. Similarly to the document level, Microsoft Azure Text Analytics API and IBM Watson Natural Language Understanding achieved the highest average scores after the educational domain-specific model, which makes them the best performing commercial tools for this particular dataset.

TABLE 4

Results for the Affective Computing & Learning dataset, at document and sentence level

Tool	**Document-level**				**Sentence-level**			
	F-measure (%)							
	+tive	**-tive**	**Neutral**	**Avg**	**+tive**	**-tive**	**Neutral**	**Avg**
IBM Watson Natural Language Understanding	71.6	30.5	45.3	49.1	58.1	35.1	58.4	50.5
Microsoft Azure Text Analytics API	72.8	27.9	54.9	51.9	61.7	29.6	61.4	50.9
OpinionFinder 2.0	39.5	15.7	45.7	33.6	22.2	11	65.4	32.9
Repustate	69.6	41.8	34.3	48.5	60.6	32.4	48.4	47.1
Sentistrength	63.4	20	31.5	38.3	59.2	24.9	50.7	45
Educational domain-specific	76.8	42.3	70.5	63.2	69.5	48.8	80.6	66.3

Table 5 presents the comparative results against the Educational Robotics dataset, both at document and sentence level. Beginning with document level, the educational domain-specific model exhibited the best average performance. Among the commercial/academic tools, Microsoft Azure Text Analytics API achieved the highest average F-measure and the more consistent results across all three classes, closely followed by Sentistrength. Difference between the domain-specific model and the second best tool was only 2.3%. Classification results for negative sentiment were considerably low and did not exceed 50% in most cases. The domain-specific model scored higher than all general-purpose tools in terms of negative and neutral class, yet it did not surpassed the positive score of Microsoft Azure Text Analytics API, Sentistrength and Repustate.

TABLE 5

Results for the Educational Robotics dataset, at document and sentence level

Tool	Document-level				Sentence-level			
	F-measure (%)							
	+tive	-tive	Neutral	Avg	+tive	-tive	Neutral	Avg
IBM Watson Natural Language Understanding	52.7	45.5	51.2	49.8	52.8	51.6	57.6	54
Microsoft Azure Text Analytics API	63.2	51.6	61.7	58.8	47.7	55.2	57.2	53.4
OpinionFinder 2.0	27.7	27.3	59.1	38	45.2	59.4	67	57.2
Repustate	62.5	42.9	62.2	55.9	64.4	54.8	61.3	60.1
Sentistrength	63.5	48.4	63.1	58.3	58.8	60.5	69.5	62.9
Educational domain-specific	59.7	60	66.7	62.1	63.7	56.8	65.2	61.9

With regard to sentence level, the domain-specific model was not the best performing tool as Sentistrength achieved the highest average score of almost 63%. With the exception of Microsoft Azure Text Analytics API, whose average performance declined by almost 5.5%, all general-purpose tools exhibited an increased classification performance compared to document level. In all general-purpose tools, capability to recognize negative sentiment was highly improved. On the other hand, the domain-specific model yielded similar results with document level as positive score was improved, yet negative score was declined. OpinionFinder 2.0 had the highest performance boost both overall and at each individual class, and even though it did not outperform the domain-specific model in terms of average performance, it managed to score better at negative and neutral sentiment.

DISCUSSION

The combination between the novel, challenging technical problems that sentiment analysis has provided to the NLP community, and the pervasive real-life applications that sentiment analysis offers to different scientific communities and practitioners is the key factor that makes it one of the most active research areas (Liu, 2015). At the same time, this challenging technical nature has led stakeholders interested in practical implementations of sentiment analysis into adopting commercial/academic tools as an "*off-the-shelf*" solution. Consideration of the following points, arisen from the benchmark evaluation results of this study, may help educators, social sciences researchers and other stakeholders interested in incorporating sentiment scores into educational research and/or practice, to guide decisions and avoid a trade-off between ready availability and performance.

i) *Different tools work better for specific datasets and analysis levels*. Synthesizing the results presented in the previous section, consistent behaviors have been identified in some commercial/academic tools based on two criteria, i.e., the dataset used and the level of analysis. From the dataset perspective, Microsoft Azure Text Analytics API and IBM Watson Natural Language Understanding were the best performing commercial tools for the Affective Computing & Learning dataset as they yielded the best overall scores at both document and sentence level. On the other hand, Sentistrength can be considered the best performing tool for the Educational Robotics dataset as, even though it did not achieve the highest average score at document level, it showed the highest consistency in classification results across the two dataset versions. From the analysis level perspective, Microsoft Azure Text Analytics API was more successful at analyzing datasets at document level while Sentistrength was consistently better at shorter, sentence-level texts. Previous studies have demonstrated that selecting a tool based on its capability to analyze longer or shorter texts is a valuable criterion for effective real-world application scenarios (Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015).

ii) *Depending on the dataset, a commercial/academic tool might be a viable solution*. Previous studies have demonstrated that domain-specific systems outperform general-purpose ones both in the educational (Nasim, Rajput, & Haider, 2017) as well as in other domains, such as technology and pharmacy (Zimbra et al., 2018). Similarly, the domain-specific models outperformed general-purpose tools in almost all cases tested in the present study; however, findings suggested that in some datasets an educational-domain specific tool might be the only way while in others a general-purpose tool could be a viable alternative. More specifically, in the Affective Computing and Learning dataset, the domain-specific tool outperformed all general-purpose ones at both levels of analysis as well as across all three classes, with difference in average performance ranging from about 11% to 15%. On the other hand, the performance gap between the domain-specific model and the general-purpose tools was significantly more narrow in the Educational Robotics dataset, with some general-purpose tools achieving competitive and even superior results. It has been pointed out that expressions of sentiment are being used in different kinds of course contexts to serve different functions; while in a programming course, negative posts reflect students' negative sentiments arisen from problem-solving difficulties, in a literature course negative posts very often contain descriptions of characters in fictions rather than expressions of students' feelings and opinions (Wen, Yang & Rose, 2014). Similarly, Affective Computing and Learning dataset originated from a course in which the presence of emotion-bearing words is inextricably linked to the very nature of the subject matter. Therefore, a domain-specific tool is able to capture this context-related information as opposed to a general-purpose one. On the other hand, in Educational Robotics dataset, the technical vocabulary used by the students did not overlap with students' expressions of sentiments and opinions. Thus, a robust general-purpose tool might work better than a domain-specific tool developed from a small dataset with limited coverage. Although a domain-specific tool developed from a larger and more representative dataset, than the one employed in this study, might outperform the general-purpose tools, many real-world sentiment analysis tasks are faced with the problem of small training sets (Forman & Cohen, 2004). Thus, weighting the nature of the technical vocabulary and the level of available resources for building a specialized tool might be a valuable criterion to inform decisions, applying to the needs of real-world application scenarios.

iii) *Any method, domain-specific or general-purpose one, should be evaluated before employed*. Although tools exhibiting consistent behavior in specific datasets and analysis levels were identified, there existed no single tool performing consistently well across all datasets and analysis levels. Bias towards a particular sentiment class has been pointed out as an issue affecting not only general-purpose tools but also domain-specific ones (Zimbra et al., 2018). Capability to recognize negative sentiment was poor even for the domain-specific tools, especially in the Affective Computing and Learning Dataset, because of the small number of negative examples from which these models were originated. From the aforementioned it follows that before incorporating any tool into educational research and/or practice, it is essential to verify its feasibility and to use the evaluation results as a quality indicator for guiding the interpretation and strength of inferences made using the produced sentiment scores.

CONCLUSIONS

The popularity of commercial and academic tools for analyzing students' sentiments as well as the limited research evidence regarding the validity of these tools gave rise to the present work. Comparative evaluation of well-known academic and commercial tools was conducted against two real-world educational datasets at different levels of analysis, and their performance was contrasted to the performance of educational domain-specific models. Points of consideration, arisen from the results of this benchmark evaluation, may help to guide decisions when application tasks of sentiment analysis in education are considered. Nonetheless, this work was not without limitations. The list of tools included in this study was not exhaustive while drawing data from genuine teaching-learning settings resulted in small and unbalanced datasets. Future studies could extend this effort by using larger educational datasets and by including additional commercial/academic tools.

REFERENCES

Altrabsheh, N., Gaber, M. M., & Cocea, M. (2013). SA-E: Sentiment analysis for education. *Frontiers in Artificial Intelligence and Applications, 255*, 353-362.

Balahur, A., Turchi, M., Steinberger, R., Ortega, J. M. P., Jacquet, G., Küçük, D., Zavarella, V., & El Ghali, A. (2014). Resource creation and evaluation for multilingual sentiment analysis in social media texts. In *Proceedings of Ninth International Conference on Language Resources and Evaluation* (pp. 4265-4269). Reykjavik, Iceland: ELRA.

Colace, F., De Santo, M., & Greco, L. (2014). SAFE: A sentiment analysis framework for e-learning. *International Journal of Emerging Technologies in Learning, 9*(6), 37-41.

Dolianiti, F., Iakovakis, D., Dias, S.B., Hadjileontiadou, S., Diniz, J.A., & Hadjileontiadis, L.J. (2018). *Sentiment analysis techniques and applications in education: A survey*. Paper presented at the 1st International Conference on Technology and Innovation in Learning, Teaching and Education (TECH-EDU 2018), Aristotle University of Thessaloniki, Greece.

Forman, G., & Cohen, I. (2004). Learning from little: Comparison of classifiers given little training. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 161-172). Berlin, Heidelberg: Springer.

Hayes, A. F., & Krippendorff, F. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77-89.

Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C. T., & Verykios, V. S. (2015). A learning analytics methodology for detecting sentiment in student fora: A case study in Distance Education. *European Journal of Open, Distance and E-learning, 18*(2), 74-94.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion Measurement* (pp. 201-237). Sawston, Cambridge: Woodhead Publishing-Elsevier.

Nasim, Z., Rajput, Q., & Haider, S. (2017). Sentiment analysis of student feedback using machine learning and lexicon based approaches. In *2017 International Conference on Research and Innovation in Information Systems* (pp. 1-6). Langkawi, Malaysia: IEEE.

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., & Strohecker, C. (2004). Affective learning – a manifesto. *BT Technology Journal*, *22*(4), 253-269.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science, 5*(1), 1-29.

Santos de Paula, F., Lechugo, C. P., & Silveira-Mackenzie, I. F. (2016). “Speak well” or “complain” about your teacher: A contribution of education data mining in the evaluation of teaching practices. In *2016 International Symposium on Computers in Education (SIIE)* (pp. 1-4). Salamanca, Spain: IEEE.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences, 311,* 18-38.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267-307.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61*(12), 2544-2558.

Tian, F., Gao, P., Li, L., Zhang, W., Liang, H., Qian, Y., & Zhao, R. (2014). Recognizing and regulating e-learners’ emotions based on interactive Chinese texts in e-learning systems. *Knowledge-Based Systems, 55,* 148-164.

Wiebe, J. M., Bruce, R. F., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 246-253). Maryland, USA: Association for Computational Linguistics.

Zarra, T., Chiheb, R., Faizi, R., & El Afia, A. (2016). Using textual similarity and sentiment analysis in discussions forums to enhance learning. *International Journal of Software Engineering and its Applications*, *10*(1), 191-200.

Zhou, Z. H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44-53.

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, *9*(2), 5(1-29).

APPENDIX I

Output handling for each dataset, at document and sentence level

Tool	**Affective Computing & Learning**		**Educational Robotics**	
	Document level	**Sentence level**	**Document level**	**Sentence Level**

Repustate	Cut-off values:
	positive > 0.75;
	0.75 > neutral > 0;
	negative < 0
		Cut-off values:
		positive > 0.65;
		0.65 > neutral > -0.55;
		negative < -0.55
			Cut-off values:
			positive > 0.2;
			0.2 > neutral > 0;
			negative < 0
				Cut-off values:
				positive > 0.45;
				0.45 > neutral > 0;
				negative < 0

Microsoft Azure Text Analytics API	Cut-off values:
	positive > 0.8;
	0.8 > neutral > 0.15;
	negative < 0.15
		Cut-off values:
		positive > 0.8;
		0.8 > neutral > 0.2;
		negative < 0.2
			Cut-off values:
			positive > 0.9;
			0.9 > neutral > 0.3;
			negative < 0.3
				Cut-off values:
				positive > 0.9;
				0.9 > neutral > 0.25;
				negative < 0.25

IBM Watson Natural Language Understanding	Cut-off values:
	positive > 0.45;
	0.45 > neutral > 0;
	negative < 0
		Cut-off values:
		positive > 0.6;
		0.6 > neutral > -0.4;
		negative < -0.4
			Cut-off values:
			positive > 0.45;
			0.45 > neutral > 0;
			negative < 0
				Cut-off values:
				positive > 0.5;
				0.5 > neutral > -0.4;
				negative < -0.4

Sentistrength	First, a single sentiment score is assigned to each text instance, summing its individual positive and negative scores; then, thresholds are set as follows: positive > 0, neutral = 0, negative < 0.			
OpinionFinder 2.0	Class labels are translated into ordinal scores (i.e., -1, 0, 1) and each text instance is assigned the median value of all its constituent clues.			