INFORMATION SYSTEMS EDUCATION JOURNAL

Volume 20, No. 1
February 2022
ISSN: 1545-679X

In this issue:

iscap
Information Systems & Computing Academic Professionals

The **Information Systems Education Journal** (ISEDJ) is a double-blind peer-reviewed academic journal published by **ISCAP** (Information Systems and Computing Academic Professionals). Publishing frequency is six times per year. The first year of publication was 2003.

ISEDJ is published online (https://isedj.org). Our sister publication, the Proceedings of EDSIGCON (https://proc.iscap.info) features all papers, panels, workshops, and presentations from the conference.

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the EDSIGCON conference. At that point papers are divided into award papers (top 15%), other journal papers (top 25%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the ISEDJ journal. Currently the target acceptance rate for the journal is under 40%.

Information Systems Education Journal is pleased to be listed in the Cabell's Directory of Publishing Opportunities in Educational Technology and Library Science, in both the electronic and printed editions. Questions should be addressed to the editor at editor@isedj.org or the publisher at publisher@isedj.org. Special thanks to members of ISCAP/EDSIG who perform the editorial and review processes for ISEDJ.

2022 ISCAP Board of Directors

INFORMATION SYSTEMS
EDUCATION JOURNAL

Using Machine Learning Sentiment Analysis to Evaluate Learning Impact

Ibrahim Lazrig
ilazrig@wtamu.edu

Sean L. Humpherys
shumpherys@wtamu.edu

Computer Information and Decision Management
West Texas A&M University
Canyon, TX 79016, USA

Abstract

Can sentiment analysis be used in an educational context to help teachers and researchers evaluate students' learning experiences? Are sentiment analyzing algorithms accurate enough to replace multiple human raters in educational research? A dataset of 333 students evaluating a learning experience was acquired with positive, negative, and neutral sentiments. Nine machine learning algorithms were used in five experimental configurations. Two non-learning algorithms were used in two experimental configurations. Each experiment compared the results of the algorithm's classification of sentiment (positive, neutral, or negative) with the judgment of sentiment by three human raters. When excluding neutral sentiment, 98% accuracy was achieved using naive Bayes. We demonstrate that current algorithms do not yet accurately classify neutral sentiments in an educational context. An algorithm using a word-sentiment association strategy was able to achieve 87% accuracy and did not require pretraining the model, which increases generalizability and applicability of the model. More educational datasets with sentiment are needed to improve sentiment analysis algorithms.

Keywords: sentiment analysis, educational research, machine learning, learner experience.

1. INTRODUCTION

Sentiment analysis is the identification of attitude, opinions, and emotions in a statement (Tang et al., 2015). Pang and Lee (2004) used sentiment analysis to classify opinions of movies in statements written online by movie viewers. Other uses of sentiment analysis have been to understand the opinions of customers regarding products, sentiments of airline travelers expressing their opinions online, and identifying positive and negative attitudes in tweets. Sentiment analysis has many subfields that solve personality recognition, sarcasm detection, metaphor understanding, aspect extraction, and polarity detection (Cambria et al., 2017). Sentiment analysis has been successfully used in marketing, product development, politics, etc.
Machine learning (ML) is one approach to sentiment analysis that involves a pretraining phase to learn from labeled data. Examples of ML algorithms include naive Bayes, support vector machines, logistic regressions, random forests, etc. Pang and Lee achieved 86% classification of sentiment accuracy in movie reviews with naive Bayes and support vector machine. Neural networks have been applied to sentiment analysis and resolve many of the lower-level NLP tasks, such as tokenization, part of speech recognition, etc. (Zhang et al., 2018).

In contrast to ML, rule-based models are expert systems that use a set of rules to achieve a conclusion or classification (Grosan & Abraham, 2011). Valence Aware Dictionary and Sentiment Reasoner (VADER) is a lexicon- and rule-based sentiment analysis model used to detect sentiments in social media posts from word-emotion associations. VADER is available in the Natural Language Toolkit package (NLTK;

http://nltk.org). NRC Word-Emotion Association Lexicon (EmoLex) uses a list of English emotion lexicon labeled by eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments, negative and positive (Saif, 2021). The labeling was originally performed by crowdsourcing.

Similar to the needs of organizations to understand the opinions of their patrons, educators need to understand the opinions and sentiments of their learners. Sentiment analysis may be able to help in an educational context.

2. LITERATURE REVIEW

Rani and Kumar (2017) propose using natural language processing and ML as tools to help university administrators process student feedback. They used NCR EmoLex to classify emotions and infer students' satisfaction and dissatisfaction with Coursera courses (coursera.org). They observed that class performance (course grade) highly correlates with student course survey results.

Munezero et al. (2013) used sentiment analysis to extract emotions from learning diaries, which are written reflections regarding students' learning experiences. Munezero et al. propose using sentiment analysis to help instructors identify emotions and track changes over time, which can be a labor-intensive task without computational aid.

One avenue of research is the investigation of which sentiment algorithms provide the highest classification accuracies in an educational context: Do the sentiment algorithms achieve the same results as human raters? Can one be a substitute for the other?

One challenge is that sentiment analysis via machine learning requires large quantities of data (Cambria et al., 2017). Existing sentiment analysis algorithms have been trained from data in non-educational domains, often from numerous online product reviews, Twitter feeds, or political forums (Yue et al., 2019). Educational research does not have the large datasets necessary to train machine learning. Different domain data means potentially different patterns and lexicons. Therefore, can existing algorithms trained in non-educational domains perform as well as or better than an ML algorithm trained only on smaller educational datasets?

Transfer learning may help resolve these challenges. Transfer learning takes an algorithm designed in one domain on an unrelated, large dataset and applies it to another domain. The algorithm learns quickly to adapt as the researcher feeds new, smaller but domain-relevant data into the pretrained algorithm for model refinement (Yang et al., 2020; Zhuang et al., 2021). The pretrained algorithm may have been trained on millions of data points and the smaller dataset only on a few hundred. The premise is that the pretrained algorithm may share many of the foundational NLP learning that still apply to the smaller dataset. The smaller dataset offers the algorithm-specific context in which to learn new patterns.

Research Questions— We propose that sentiment analysis be used to investigate the learner's experience of a learning treatment. Instead of using multiple human raters to evaluate the student's opinion about the learning experience, a sentiment analysis algorithm could be used. Specifically, we investigate algorithms to identify the positive/negative sentiments in an experimental treatment on student learning in computer information system (CIS) courses. Our aim is to use algorithms to automate the identification of students' sentiments toward a taught subject from their reviews. We tested a set of machine learning algorithms to answer the following research questions:

- Can sentiment analysis be used in an educational context to possibly help instructors and researchers evaluate students' learning experiences?
- Are sentiment analyzing algorithms currently accurate enough to replace multiple human raters in educational research?
- Can other domain datasets with sentiments be used to train sentiment analysis algorithms to detect sentiments in educational datasets?

3. METHODOLOGY

Participants and Design

Graduate and undergraduate students in three CIS courses (eight sections) were taught and practiced time management as a professional development skill. Quantitative measures of grade performance were analyzed. The main finding in regard to the impact of learning time management skills on grades is reported by Humpherys and Lazrig (2021). In that study, a survey was administered regarding students' perceptions of the learning exercise with the question "Each week you were asked to preplan your study schedule and identify your deliverable. Did this activity help you improve your time management skills? Why or why not? You get points for participation, not for any predefined answer." 180 student reviews were collected, with judgement of sentiment (positive, negative, and neutral) from three human raters. The current study uses machine learning sentiment analysis to compare the performance of algorithms to human raters.

Variables

Sentiment is the construct in question. Sentiment was derived by human raters and algorithms, then compared for *accuracy* as follows:

Human rater-derived sentiment— the sentiment assigned by three human raters regarding the participant's review of the learning experience was encoded as -1 for negative, 0 for neutral, and 1 for positive sentiment. The average of the human rater-derived sentiment is calculated and rounded to the nearest integer. *Positive* indicates a sentiment of improvement in time management, positive results, or valuable learning experience. *Neutral* indicates the participant expressed no improvement in time management or indifference to the learning experience. *Negative* expresses a decrease in time management, negative results, or dissatisfaction with the learning experience.

ML-derived sentiment— encoded as -1 for negative, 0 for neutral, and 1 for positive sentiment derived from an ML sentiment analyzing algorithm. Various algorithms are used and explained later.

Accuracy— how well the ML algorithm predicted the same sentiment score (positive, neutral, negative) as the human raters. The human rater-derived sentiment was considered to be ground truth. Accuracy is a percentage representing the number of sentiments correctly classified by the algorithm divided by the total number of sentiments (Hossin & Sulainman, 2015).

$$\text{accuracy} = \frac{TP + TN + TNu}{TP + TN + TNu + FP + FN + FNu}$$

Accuracy Term	**Matching Results:**	
	Algorithm	**Human**
TP (True Positive)	Positive	Positive
TN (True Negative)	Negative	Negative
TNu (True Neutral)	Neutral	Neutral
FP (False Positive)	Positive	Negative or Neutral
FN (False Negative)	Negative	Positive or Neutral
FNu (False Neutral)	Neutral	Positive or Negative

Table 1. Meanings of accuracy terms

Table 1 shows the definition of terms used when calculating accuracy. Each term is a count (integer). For example, if the algorithm classified a student's comment as negative sentiment but the human rater-derived sentiment was either positive or neutral for the same student's comment, the count of false negatives was incremented. This process was repeated for every data point in the datasets.

Data Collection Procedures

Five datasets were acquired or generated for use in this research (Table 2).

Dataset	Dataset Description	Sample Size
Learning Sentiment	Dataset of students' perceptions of a learning exercise in CIS courses (positive, negative, neutral) augmented with additional negative and neutral ratings of instructors/courses.	333
Learning Sentiment w/o Neutral	Learning Sentiment dataset without neutral sentiments	285
Movies	Pretrain on reviews of movies (positive and negative)	2,000
Airline	Pretrain on tweets about airline service (positive, negative, neutral)	14,640
Airline w/o Neutral	Airline dataset without the neutral sentiments	11,541

Table 2. Datasets.

Learning Sentiment dataset— The dataset has a total of 333 student reviews. 180 students reviewed a time management learning exercise in three CIS courses of which 154 reviews were positive. To increase the number of negative and neutral sentiments, 153 student reviews regarding instructors and courses were collected from rateMyProfessor.com. RateMyProfessor.com lets students write evaluations and comments about courses. In addition to the text-based comments, students select a quality score of 1–5. A quality score of 4 or 5 is labeled "awesome," 3 is considered "average," and 2 or 1 is considered "awful." Furthermore, green, yellow, and red icons are associated with the respective quality scores/labels, which can be equated to positive, neutral, or negative sentiment respectively.

First, the ratings were filtered with the name of the university to match the original data's student population. Next, a random course was selected, but not one of the three CIS courses in the original 180 student review dataset. "Awesome" quality scores (4 and 5) were ignored, given the desire to collect more neutral and negatives comments. If the quality score was a 1, 2, or 3, a human rater read the student's comment. If the human rater agreed that the student's comment was classifiable as a quality score of 1, 2, or 3, the comment and quality score were included in the Learning Sentiment dataset. The quality score was recoded to match the sentiment score in the original dataset. A 1 or 2 quality score was recoded as negative sentiment (i.e., a -1 value in the Learning Sentiment dataset). If the quality score was 3, the sentiment was recoded as neutral (0 value).

These extra reviews were collected to more closely balance the positive and negative reviews and increase the neutral reviews in the dataset. The limitation of the extra review data is that the learning experience reviewed by the students was not just the time management exercise, as originally planned. But since the research questions are about the accuracy of the sentiment algorithms, not about the learning exercise, this limitation should not impact the validity of the sentiment accuracy results. In addition, the threat to validity by an unbalanced dataset where the ML algorithm learns to predict all data as positive sentiments is a greater threat than the limitation of adding extra reviews from different courses. The final sentiment counts in the Learning Sentiment dataset were 154 positive, 48 neutral, and 131 negative. An IRB review process authorized the analysis but did not explicitly permit the dataset to be made public.

Learning Sentiment without Neutral dataset— Neutral sentiments were removed from the Learning Sentiment dataset to compare with publicly available datasets that do not include neutral sentiment and because in past research, neutral sentiments have demonstrated difficulty to evaluate. This resulted in 154 positive and 131 negative data points. The accuracy calculation therefore removed TNu and FNu as terms.

Movie Review dataset— The Movie Review dataset is included in the Natural Language Toolkit (NLTK) package publicly available at https://github.com/nltk/nltk. The dataset was originally collected by Pang and Lee (2004) and has 2,000 reviews with 50% negative sentiment, 50% positive, and no neutral. The movie reviews were written before 2002 on www.rottentomatoes.com by 312 authors with a maximum of 20 reviews per author.

Airline Review dataset— The Airline Review dataset contains 14,640 tweets made about a US Airline in February 2015 with 2,363 classified as positive, 9,178 as negative, and 3,099 as neutral (Crowdflower, 2019). The dataset is publicly available at https://www.kaggle.com /crowdflower/twitter-airline-sentiment.

Airline Review without Neutral dataset— Neutral sentiments were removed from the Airline Review dataset to pretrain some ML models for transfer learning.

Data preprocessing— The preprocessing stage prepares the five datasets for sentiment analysis by cleaning and vectorizing the data. Cleaning the data pertains to removing irrelevant terms, names, and symbols (# and @), and converting all words into lowercase to simplify word matching procedure. In addition, some high frequency words are filtered out, such as stopwords.

A stopword is a commonly used word (such as "the", "a", "an", "in") that adds little value to classification. The NLTK corpus package used has a predefined list of stopwords stored in many different languages, and we used the English stopwords from that list.

Vectorization— We converted the cleaned text into numerical vectors to be used as features in the algorithm. A tokenizer split the text into words, or tokens (known as bag-of-words), then converted them into a feature vector based on word count or term frequency-inverse document frequency (TF-IDF), which is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

Experimental Configurations

To answer the research questions, we ran seven sets of experiments. In experiments 1–5, we used the NLTK for natural language processing and we used the scikit-learn library in Python (https://scikit-learn.org/) for the machine learning algorithms. *Local-training* means we used the Learning Sentiment data for training and testing the algorithms. Local-training gives a baseline to compare against transfer learning using external-training models. *External-training* means the ML models are pretrained (transfer learning) using the Airline Review dataset or Movie Reviews dataset. Then, the model is tested for accuracy with the entire Learning Sentiment dataset. It is anticipated that external-training can overcome the relatively small sample size of the Learning Sentiment dataset and simulate the desired outcome of being able to replace human raters in educational research.

Experiment #1 used the Learning Sentiment dataset for both training and testing. Nine classification algorithms were used (see Appendix A). We employed a 10-fold cross-validation method to calculate the average accuracy: In each fold, the dataset was randomly shuffled and divided into training and testing subsets with the ratio 80:20, then the 10 accuracies were averaged. This process was repeated for each of the nine classification algorithms. Cross-fold validation reduces overfitting and increases generalizability.

Experiment #2 used the Learning Sentiment without Neutral dataset and repeated the procedures of Experiment #1. Since most of the false positives and false negatives in Experiment #1 were due to the misclassification of the neutral sentiments, we decided to investigate the accuracies without neutral reviews. Even human raters can display low inter-rater consistency when classifying neutral sentiments.

Experiment #3 used the Movie Review dataset to pretrain the ML model. All 285 records in the Learning Sentiment without Neutral dataset were used for testing the accuracy of the ML model, since the Movie Review dataset does not have neutral sentiments.

Experiment #4 used the Airline Review dataset for pretraining the ML model. All 333 records in the Learning Sentiment dataset were used for testing accuracy as the Airline Review dataset includes neutral sentiments.

Experiment #5 used the Airline Review without Neutral dataset for pretraining the ML model. All 285 records in the Learning Sentiment without Neutral dataset were used for testing the accuracy of the ML model. This allows for comparison to Experiment #3 regarding transfer learning.
We included two more experiments (Exp#6 and Exp#7) that used rule-based modeling rather than ML, namely VADER and EmoLex. VADER returns a composite real score value ranging between -1 and 1 for the sentiment of a given text with -1 for most negative, +1 for most positive, and around zero for neutral. We set a threshold for the neutral sentiments to be between -0.05 to +0.05. The EmoLex algorithm returned integer scores for positive and negative words in the text. We compared the two scores to determine the overall sentiment of the text. If the positive score is greater than the negative, then the final sentiment will be positive and vice versa. If both are similar or both are zero, the sentiment will be neutral.

Experiment #6 used the rule-based VADER and EmoLex models to test the accuracy of sentiment detection on the Learning Sentiment dataset. Experiment #7 used the rule-based VADER and EmoLex models to test the accuracy of sentiment detection on the Learning Sentiment without Neutral dataset.

The link to the code used in this study is available using the following link:
https://github.com/iLazrig/Sentiment-Analysis-Experiment.git

4. RESULTS

Table 3 summarizes the highest accuracies of sentiment classification achieved in each experiment (#1–7) and the algorithm that performed the best.

Experiment #	Highest Accuracy %	Highest Performing Algorithm
#1 Learning Sentiment	85.1	Naive Bayes, Random Forest, Logistic Regression
#2 Learning Sentiment w/o Neutral	98.3	Naive Bayes
#3 Movies pretraining	77.2	Naive Bayes & AdaBoost
#4 Airline pretraining	55.6	Naive Bayes
#5 Airline pretraining w/o Neutral	61.4	Naive Bayes
#6 Learning Sentiment	72.3	VADER
#7 Learning Sentiment w/o Neutral	86.7	VADER

Table 3. Highest accuracies and algorithms

Algorithm	Accuracy %				
	Local Training		External Training		
	Exp#1 Learning Sentiment	Exp#2 Learning Sentiment w/o Neutral	Exp#3 Movies	Exp#4 Airline	Exp#5 Airline w/o Neutral
Bernoulli-NB	85.1	94.0	54.4	55.6	57.9
Complement NB	85.1	98.3	77.2	52.9	57.9
Multinomial NB	82.1	98.3	77.2	54.7	61.4
K-Neighbors	47.8	57.4	43.9	48.7	61.4
Decision Tree	71.6	89.2	68.4	52.9	50.9
Random Forest	85.1	96.3	61.4	52.0	54.4
Logistic Regression	85.1	94.3	63.2	40.8	52.6
MLP	82.1	96.0	73.7	40.8	54.4
AdaBoost	73.1	93.7	77.2	42.3	56.1

Table 4. Accuracies from experiments #1–5 using sentiment ML algorithms.

The nine ML algorithms and their classification accuracies from experiments #1–5 are shown in Table 4. The highest accuracies in experiments #1–5 are as follows: The naive Bayes, random forest, and logistic regression ML algorithms had accuracies of 85% in experiment #1 and up to 98% when neutral sentiments were removed in experiment #2.

Pretraining the ML model from the Movie Review dataset and validating the accuracy on the Learning Sentiment without Neutral dataset (experiment #3) saw classification accuracies up to 77%. Pretraining the ML model using the Airline Review dataset (with and without neutrals) performed worse. External training did not improve classification algorithms over the local training.

Experiments #6 and #7 used rule-based modeling, specifically VADER and EmoLex (see Table 5). VADER achieved 72.3% accuracy in experiment #6 and 86.7% in experiment #7. EmoLex achieved 55.0% accuracy in experiment #6 and 73.8% in experiment #7. Experiment #6 used the full Learning Sentiment dataset while experiment #7 used the Learning Sentiment without Neutral dataset.

Algorithm	Accuracy % in Exp#6 Learning Sentiment	Accuracy % Exp#7 Learning Sentiment w/o Neutral
VADER	72.3	86.7
EmoLex	55.0	73.8

Table 5. Accuracy of the rule-based models for sentiment

5. DISCUSSION

Sentiment can be positive, negative, or neutral. Sentiment analysis has largely been used in product/service reviews, movie reviews, and politics. This study proposes using sentiment analyzing algorithms to evaluate sentiment in an educational context. Teachers could use sentiment analysis to quickly evaluate sentiment from student reviews after administering a learning exercise or from course evaluations. Researchers could save time and resources when evaluating an educational treatment for sentiment by replacing multiple human raters with a sentiment analyzing algorithm.

Can sentiment analysis perform accurately in an educational context? The experiment with the highest sentiment classification accuracy was Experiment #2, which used the Learning

Sentiment without Neutral dataset for both training and testing. Accuracy of predicting positive and negative sentiment reached 98% using naive Bayes. For predicting positive, negative, and neutral, the highest performing algorithms were in Experiment #1, which used the Learning Sentiment dataset for both training and testing. In Experiment #1, naive Bayes, random forest, and logistic regression produced accuracies of 85%. These results show the potential of using sentiment analysis in education.

From these results we deduce that neutral sentiment is hard to detect. The observed lower accuracies in some experimental configurations was due to misclassification of the neutral sentiments. Our recommendation is that if a teacher or researcher wishes to apply sentiment analysis to an educational context, they are currently limited to only positive and negative sentiment, not neutral, at least until the neutral-detecting algorithms improve.

Another research question is whether or not sentiment analyzing algorithms perform accurately enough to replace human raters. Here, the scenario is a researcher evaluating an educational treatment regarding the sentiment of the learner. The Learner Sentiment dataset originally used three human raters to assess sentiment. Can an algorithm be used to replace the human raters? The requirement for this proposal to succeed is that the researcher should not have to use the target dataset to train the ML model, as in Experiments #1 and #2, since doing so would defeat the purpose of performing a sentiment analysis on unlabeled data and without human involvement. Experiments #3 through #7 tested this scenario. Experiments #3, #4, and #5 used ML models pretrained from movie reviews and airline reviews. Pretraining with those datasets offered tens of thousands of records to refine a sentiment model before applying the model to a target educational dataset. However, accuracy rates were only 77%. The sentiment models trained on the Movie Review dataset and tested on the Learning Sentiment without Neutral dataset (Experiment #3) performed better than the models trained on the Airline Review dataset (Experiment #4 and #5).

We conclude that the Movie Review data is possibly closer in characteristics to the educational dataset than the Airline Review dataset. The pretrained models became domain-dependent. The Airline Review dataset entries are short tweets while the Movie Review dataset entries were longer reviews. The vocabulary distribution across the opinions is different between the two datasets. Neither dataset is sufficient to offer a viable replacement to human raters. Data domain is very important for supervised ML sentiment analysis. Because sentiment domain-specific datasets are sparse in educational research, we opine that if ML algorithms are to be improved, more educational datasets need to be collected and made publicly available, following ethical guidelines for privacy. When a model was trained on educational data, the ML algorithms performed as well as human raters in identifying positive and negative sentiment with the advantage of speed and automation.

Unsupervised algorithms, like VADER, are promising. In Experiments #6 and #7, pretraining the sentiment model was not required. One could take the VADER rule-based algorithm as is and evaluate a target dataset for sentiment. The VADER algorithm performed better than many of the supervised ML algorithms with 72% accuracy for the Learning Sentiment dataset with neutral sentiments included and 87% with neutral sentiments removed. Arguably, 87% approaches an adequate level of accuracy to be useful in an educational context. The VADER algorithm is useful for getting a quick and general (summarized) view or trend of students’ opinions about a topic without the need for human intervention, which could save resources and the instructors’ time. One application of using VADER in the classroom is to have students digitally text opinions about a lecture topic, e.g., business case, scenario, or argument position, and the VADER algorithm can instantly quantify how many students expressed a positive or negative opinion about it. The summary can be presented back to the students as part of the same lecture. Sentiment analysis can also be applied to short essay assignments or to analyze exam responses. Another application can be for administrators to identify struggling teachers and offer assistance after using an automated sentiment analysis for course reviews. With thousands of students' comments, reading all the comments may be fatiguing and ineffective, but an algorithm can identify positive and negative comments to better focus an administrator's care and attention.

In conclusion, instructors desire to evaluate if learning activities (e.g., individual project, group project, service-learning activity, presentation, student research, etc.) have positive impacts on students. Grades are only one measure of learning impact. The sentiment of the student is another measure. After conducting a learning activity, instructors can collect reflective experiences via a short essay or open-ended response from the students. Sentiment analysis can then be used to categorize the students' reflections as positive or negative. Having a count of how many students had a positive or negative experience may guide the instructor in making adjustments to future learning activities and can quantitatively track the impact of adjustments over time. Educational researchers have similar opportunities using sentiment analysis.

Based on the results of this study, we have two recommendations. If the instructor has enough data or prior data from an educational context, they can vectorize the words and train a naive Bayes to detect positive and negative sentiment. Then, they can use that model on the remaining student reflection data. We do not recommend detecting neutral sentiment at this time because the current algorithms and datasets are not sufficiently accurate for neutral sentiment analysis. Future research is needed to accurately identify neutral sentiments. Nor do we recommend augmenting the machine learning training process with data from publicly available sentiment datasets that are outside the educational context (e.g., movie reviews, airline reviews, product reviews, etc.). If the instructor does not have enough student data to pretrain a model, we recommend using the VADER algorithm as is. VADER achieved 87% accuracy in this study and may be sufficient for the instructor's analytical needs. VADER has the advantage of speed and ease as it does not require pretraining a model.

Positive and negative sentiment labels derived from VADER or a naive Bayes algorithm could also be used as input, along with student demographic variables, for clustering algorithms. The clustering algorithm may categorize which subgroups of students had positive or negative experiences from a learning activity. This insight may inform the instructor if certain student populations are disproportionately impacted so that corrective action can be taken. More educational datasets with sentiment are needed to improve future sentiment analysis algorithms.

6. REFERENCES

Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, *32*(6), 74–80. https://doi.org/10.1109/MIS.2017.4531228

Crowdflower. (2019, October 15). *Twitter US Airline Sentiment: Analyze How Travelers in February 2015 Expressed Their Feelings on Twitter*. Twitter US Airline Sentiment. https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Grosan, C., & Abraham, A. (2011). Rule-Based Expert Systems. In C. Grosan & A. Abraham (Eds.), *Intelligent Systems: A Modern Approach* (pp. 149–185). Springer. https://doi.org/10.1007/978-3-642-21004-4_7

Hossin, M., & Sulainman, M. N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2). https://doi.org/: 10.5121/ijdkp.2015.5201

Humpherys, S. L., & Lazrig, I. (2021). Effects of Teaching and Practice of Time Management Skills on Academic Performance in Computer Information Systems Courses. *Information Systems Education Journal*, *19*(2), 45–51.

Munezero, M., Montero, C. S., Mozgovoy, M., & Sutinen, E. (2013). Exploiting sentiment analysis to track emotions in students' learning diaries. *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*, 145–152. https://doi.org/10.1145/2526968.2526984

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 271-es. https://doi.org/10.3115/1218955.1218990

Rani, S., & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. *Computer*, *50*(5), 36–43. https://doi.org/10.1109/MC.2017.133

Saif, M. (2021). *NRC Word-Emotion Association Lexicon*. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: Successful approaches and future challenges. *WIREs Data Mining and Knowledge Discovery*, *5*(6), 292–303. https://doi.org/10.1002/widm.1171

Yang, Q., Zhang, Y., Dai, W., & Pan, S. J. (2020). *Transfer Learning*. Cambridge University Press. 978-1-108-86008-6

Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A Survey of Sentiment Analysis in Social Media. *Knowledge and Information Systems*, *60*(2), 617–663. https://doi.org/10.1007/s10115-018-1236-4

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), e1253. https://doi.org/10.1002/widm.1253

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555