IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

Leveraging Machine Learning Sentiment Analysis to Assess Learning Impact

[1] *Anohu Nuthalapati,* [2]*Phaniraj Kumar,* [3] *DR. B.Naveen Kumar*

Abstract:
In the rapidly evolving landscape of education and professional development, the evaluation of learning impact has become increasingly crucial. Traditional methods of assessment often fall short in capturing the nuanced and dynamic nature of learning outcomes. This paper explores the application of machine learning sentiment analysis as a novel and effective approach to evaluate the impact of learning experiences. By harnessing the power of natural language processing and data analytics, we aim to provide educators and institutions with a robust framework for assessing the emotional and cognitive impact of their educational programs. This paper discusses the theoretical foundations, methodology, and potential benefits of utilizing sentiment analysis in learning impact evaluation.

Keywords: Sentiment analysis(SA), Natural language processing (NLP), Machine learning (ML).

I. Introduction

In today's educational and corporate training environments, the need to evaluate the effectiveness of learning experiences is more pressing than ever. Traditional assessment methods, such as quizzes and exams, while informative, often provide a limited understanding of the holistic impact of learning on an individual. They fail to capture the emotional responses, motivation, and the broader cognitive shifts that occur during the learning process. To address these limitations, we turn to the burgeoning field of machine learning sentiment analysis.

Sentiment analysis, a subfield of natural language processing (NLP), has gained prominence for its ability to extract insights from textual data by determining the sentiment or emotional tone expressed in the text. It has been applied to various domains, such as customer reviews, social media sentiment, and product feedback. In this paper, we propose the use of sentiment analysis to evaluate learning impact, a concept that has gained significant attention but often lacks precise measurement. By analyzing the sentiment expressed in written reflections, survey responses, and other text-based assessments, we can gain deeper insights into the learners' emotional and cognitive states.This paper will present a comprehensive examination of sentiment analysis as an evaluation tool for learning impact, addressing its theoretical underpinnings, the methodology, and its potential benefits for educators, trainers, and institutions. Through this exploration, we aim to highlight the potential of sentiment analysis to enhance the assessment of learning impact and contribute to the ongoing improvement of educational and training programs.

[1] *Assistant Professor Department of CSE, Rise Krishna Sai Gandhi Group of Institutions,* [2] *Assistant Professor Department of CSE, Rise Krishna Sai Prakasam Group of Institutions,* [3] *Professor Department of CSE, Rise Krishna Sai Gandhi Group of Institutions*

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

Sentiment analysis is the identification of attitude, opinions, and emotions in a statement. Pang and Lee 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., 2020). 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.

In contrast to ML, rule-based models are expert systems that use a set of rules to achieve a conclusion or classification (Grosan & Abraham). 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 wordemotion 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 Survey

The literature survey section will provide an in-depth review of existing research and studies related to the application of sentiment analysis in education and learning assessment. It will cover:

Historical Overview: A historical overview of assessment methods and their limitations in measuring learning impact. Sentiment Analysis in Education: A review of studies that have applied sentiment analysis in educational contexts, including those assessing student satisfaction, engagement, and emotional response to learning materials.

Machine Learning and NLP in Education: An exploration of the role of machine learning and natural language processing in educational research and assessment. Theoretical Foundations: An examination of the theoretical foundations underpinning sentiment analysis and its applicability to the assessment of learning impact.

Methodological Approaches: A discussion of the various methodologies and tools employed in sentiment analysis for learning impact evaluation, including data collection, preprocessing, and modeling.

Benefits and Challenges: An assessment of the potential benefits and challenges associated with implementing sentiment analysis in educational and training settings.

One challenge is that sentiment analysis via machine learning requires large quantities of data. Existing sentiment analysis algorithms have been trained from data in non-educational domains, often from numerous online product reviews, Twitter feeds, or political forums. 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

ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

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 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.

The literature survey will provide a comprehensive understanding of the existing research landscape, paving the way for the subsequent sections that delve into the methodology, results, and conclusions of this study.

3. Problem Statement

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? • Is 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?

4. Proposed System

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. 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.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

Accuracy = (TP+TN+TNu) / TP+TN+TNu+FP+FN+Fnu

Accuracy team: TP (True Positive), TN (True Negative), TNu (True Neutral), FP (False Positive), FN (False Negative) and FNu (False Neutral).

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:

Five datasets were acquired or generated for use in this research.

Table:1

Dataset	**Dataset Description**	**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	350
Learning Sentiment w/o Neutral	Learning Sentiment dataset without neutral sentiments	350
Movies	Pretrain on reviews of movies (positive and negative)	3000
Train Line	Pretrain on tweets about train service (positive, negative, neutral)	15500
Airline	Pretrain on tweets about airline service (positive, negative, neutral)	12300

Learning Sentiment dataset— The dataset has a total of 350 student reviews. 190 students reviewed a time management learning exercise in three CIS courses of which 161 reviews were positive. To

ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

increase the number of negative and neutral sentiments, 160 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 190 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 190 positive, 48 neutral, and 131 negatives. An IRB review process authorized the analysis but did not explicitly permit the dataset to be made public.

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 and has 2,000 reviews with 50% negative sentiment, 50% positive, and no neutral. The movie reviews were written before 2014 on www.rottentomatoes.com by 312 authors with a maximum of 20 reviews per author.
Train Review dataset— The Train Review dataset contains 9562 tweets made about a Indian February with 2,363 classified as positive, 5234 as negative, and 4328 as neutral. The dataset is publicly available.
Airline Review dataset— The Airline Review dataset contains 14,640 tweets made about a Airline in February 2015 with 2,363 classifieds as positive, 9,178 as negative, and 3,099 as neutral. The dataset is publicly available.
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 stop words.
A stop word 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 stop words stored in many different languages, and we used the English stop words 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-

ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

IDF), which is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

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 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.

5. Result:

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

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

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 2. Highest accuracies and algorithms

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.

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.

5. 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 Information Systems Education Journal (ISEDJ) 20 (1) ISSN: 1545-679X February 2022 ©2022 ISCAP (Information Systems and Computing Academic Professionals) Page 20 https://isedj.org/; https://iscap.info 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

ISSN 2347–3657
Volume 11,Issue 1,Jan 2023

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.

References

1. Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. IEEE Intelligent Systems, 32(6), 74–80. ttps://doi.org/10.1109/ MIS.2017. 4531228 Crowdflower. (2019, October 15).
2. 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 -3.airline-sentiment Grosan, C., & Abraham, A. (2011).
3. 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).
4. 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).
5. 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.

6. Using Machine Learning Sentiment Analysis to Evaluate Learning Impact, Ibrahim Lazrig, Sean L. Humpherys, Information Systems Education Journal (ISEDJ), ISSN: 1545-679X, February 2022.