computers

Article

Shared Language: Linguistic Similarity in an Algebra Discussion Forum

Michelle P. Banawan [1], Jinnie Shin [2], Tracy Arner [3], Renu Balyan [4], Walter L. Leite [2] and Danielle S. McNamara [3,*]

[1] Asian Institute of Management, Makati City 1229, Metro Manila, Philippines
[2] College of Education, University of Florida, Gainesville, FL 32611, USA
[3] Department of Psychology, Arizona State University, Tempe, AZ 85281, USA
[4] SUNY, Old Westbury, NY 11568, USA
* Correspondence: dsmcnamara@asu.edu

Abstract: Academic discourse communities and learning circles are characterized by collaboration, sharing commonalities in terms of social interactions and language. The discourse of these communities is composed of jargon, common terminologies, and similarities in how they construe and communicate meaning. This study examines the extent to which discourse reveals "shared language" among its participants that can promote inclusion or affinity. Shared language is characterized in terms of linguistic features and lexical, syntactical, and semantic similarities. We leverage a multi-method approach, including (1) feature engineering using state-of-the-art natural language processing techniques to select the most appropriate features, (2) the bag-of-words classification model to predict linguistic similarity, (3) explainable AI using the local interpretable model-agnostic explanations to explain the model, and (4) a two-step cluster analysis to extract innate groupings between linguistic similarity and emotion. We found that linguistic similarity within and between the threaded discussions was significantly varied, revealing the dynamic and unconstrained nature of the discourse. Further, word choice moderately predicted linguistic similarity between posts within threaded discussions (accuracy = 0.73; F1-score = 0.67), revealing that discourse participants' lexical choices effectively discriminate between posts in terms of similarity. Lastly, cluster analysis reveals profiles that are distinctly characterized in terms of linguistic similarity, trust, and affect. Our findings demonstrate the potential role of linguistic similarity in supporting social cohesion and affinity within online discourse communities.

Keywords: math discourse; natural language processing; linguistic similarity; algebra; discussion forums

check for updates

Citation: Banawan, M.P.; Shin, J.; Arner, T.; Balyan, R.; Leite, W.L.; McNamara, D.S. Shared Language: Linguistic Similarity in an Algebra Discussion Forum. *Computers* **2023**, *12*, 53. https://doi.org/10.3390/computers12030053

Academic Editors: Phivos Mylonas, Katia Lida Kermanidis and Manolis Maragoudakis

Received: 13 February 2023
Revised: 24 February 2023
Accepted: 24 February 2023
Published: 27 February 2023

1. Introduction

Interactions between students in collaborative learning activities are often studied within face-to-face learning environments. Considerable evidence indicates that students sharing ideas or questions in pursuit of knowledge demonstrate improved learning outcomes compared to those learning in isolation [1,2]. Thus, teachers are encouraged to incorporate collaborative activities where students engage in the co-development of a shared mental model with one or more peers [3]. Within these collaborative environments, learners engaging in academic discourse have a shared set of rules, including social cues that guide their communication. For example, when one peer is confused about what the other is saying, they may shift their eye gaze or shrug their shoulders [4]. Such exhibition of social cues may result in a dialogic exchange in which participants make connections between utterances to co-construct knowledge. This coordinated discourse is representative of shared language reflecting commonality or connection that unites the members of the learning community [5,6]. Shared language further contributes to the cohesion of a community [7]. Coordinated communication representing community members' shared conceptualizations and perspectives is associated with positive social outcomes not only in face-to-face interactions but also in technology-mediated settings [8]. Given the recent increased adoption of online learning environments, it is increasingly important to better understand the nature of effective communication in asynchronous and collaborative environments such as discussion boards. The quality of discourse in asynchronous, collaborative digital spaces enhances the learning experience and contributes to the attainment of learning goals. Hence, promoting student engagement and participation in academic discourse benefits learning through collaborative problem-solving, peer instruction, and co-construction of knowledge.

Asynchronous discussion boards are often implemented as a mechanism to build learning communities and introduce collaborative learning activities [9]. However, student participation in discussion boards continues to be one of the hurdles to successful discussion board implementation. Students' participation in discussions is linked to course satisfaction, feelings of belonging, and positive learning outcomes [10,11]. While collaboration and interaction among students enhance their overall learning experience, student participation remains a challenge in discussion board deployments [12–14]. Despite the goal of discussion boards to facilitate community development, build social interactions, and scaffold meaning-making, students may not develop a feeling of belonging or a sense of affinity toward their fellow students. However, student perceptions of affinity and belonging are more pronounced and easier to capture within face-to-face classroom settings as compared to an online modality.

This study investigates the factors that are associated with collaboration and social interactions among discourse participants by looking at the discourse through the lens of shared language, i.e., lexical entrainment and semantic similarity. In addition to probing the similarity of students' lexical choices in the discussion threads, we simulate the linguistic markers of affinity and trust within a technology-mediated setting designed for learning algebra. For example, lexical and semantic dimensions of language are indicators of affective expressions (i.e., valence and arousal) from naturalistic dialogues within an intelligent tutoring system [15]. Valence is a dimension of emotion that represents a continuum of negative/unpleasant to positive/pleasant. Arousal is the dimension that depicts the intensity of emotional activation ranging from low/calm to high/excited. Another social factor that we simulate in modeling shared language and social interactions among discourse participants is affinity. Coordinated communication through mimicry, which includes word choice and syntactic structure [16], is linked to affinity [17]. When discourse participants use the same shared language, it serves as evidence of "membership" or inclusion and, at the same time, contributes to the cohesion of a community [7]. Discourse participants must have coordinated notions of how to use and interpret language within the discourse [18]. This coordination is descriptive of a shared language that reflects the commonality of world views or conventions that unite the members of a discourse community [5].

Understanding the nature of language within discourse communities in this study provides valuable insights into students' engagement, trust, and affinity amongst participants, as evidenced by their lexical choices and linguistic similarity. Evidence of engagement, trust, and affinity indicates that students have adequately participated in the discourse, and thus, the pedagogical goals of discussion forums are attained. Hence, the potential benefits of discussion boards, which include increased opportunities for interaction and enhanced learning experiences, will materialize. Multiple methods used for analyzing this discourse leverage natural language processing (NLP) methods, including deep learning for NLP models, and the integration of qualitative analysis to theoretically and contextually validate the findings.

1.1. Lexical Entrainment and Semantic Similarity

In the pursuit of effective communication, parties in a conversation act cooperatively, and this is reflected in their lexical choices to increase understanding and reduce miscommunication. The conversing parties use common terminologies or, at times, agree to a change in terminology to overcome language ambiguities as conversations progress. Linguistic or lexical mimicry describes coordinated communication that involves the repetition of words and phrases among discourse participants [19]. Lexical entrainment is a related phenomenon that encompasses lexical mimicry, the commonality of words used, and the similarity of the linguistic choices pertaining to lexical and syntactic dimensions of language. Lexical entrainment reflects the process of adopting a shared lexicon of terms that parties in a conversation employ. This process represents the shared mental model of discourse as evident in their lexical choices while communicating. Lexical entrainment usually involves coming to an understanding and collaboratively co-constructing knowledge using the agreed-upon lexical symbols or "shared language" throughout the discourse. The presence and extent of lexical entrainment have been observed in discourse communities characterized by trust [20,21] and group success [22].

Semantic entrainment is likewise observed in this linguistic adaptation among discourse participants, leveraging semantic similarity as a measure of entrainment [23]. Semantics, in linguistics, is the study of the meanings of words and sentences. Lexical semantics represents how the meanings of words are understood in the context in which they are used and is derived from the relationships between words in a sentence. Semantic similarity, which represents semantic entrainment, indicates whether two texts have similar or different meanings [24]. It is mathematically derived using word embeddings, sentence embeddings, or both and calculates a similarity metric representing the distance between the texts being compared. Semantic similarity NLP indices are used in various applications, such as text categorization, summarization, question answering, feedback, and information retrieval. Word similarities may be computed based on plain co-occurrence using measures such as raw co-occurrence count, text frequency—inverse document frequency (TF-IDF), or pointwise mutual information (PMI), or using second-order co-occurrences wherein word contexts, also called word embeddings, are used in the distance calculation. Lexical databases, such as Wordnet [25], are also used in calculating the semantic similarity between words. While semantic similarity captures overlap at the word level (i.e., local context), textual similarity reflects the semantic equivalence beyond words (i.e., longer strings of lexemes—sentences and paragraphs). Similarity between texts is also calculated using various distance computation approaches that are also used in word-level similarity. The most noteworthy innovations in similarity computations involve deep learning in NLP, a.k.a. deep NLP. These state-of-the-art models can represent context beyond word occurrence methods through advances in machine learning algorithms and word and sentence embeddings. Textual similarity measures can efficiently approximate semantic, lexical, and syntactic similarity [26]. This work on linguistic similarity represents the discourse participants' shared language based on the similarity of their word choices and understanding of these words (i.e., meaning-making) [27].

1.2. Math Discussion Boards and Math Language Processing (MLP)

Language is an essential resource in mathematics learning and provides the students the ability to participate in math instructional activities. Meaningful discourse within math discussion boards allows the students to ask questions and supplement learning from the lessons and instructional content. Discussion boards facilitate knowledge co-construction and transfer. Further, math knowledge construction is comprised of procedural and conceptual language as students conceptualize abstract ideas and collaborate in real-time problem-solving. In prior work extracting the linguistic profiles of online math discourse, we found that the dominant communicative goals that were prevalent in a math discussion board include elaboration, providing instruction, establishing explicit relations and analogies between mathematical constructs, and presenting information [28].

The application of natural language processing tasks and machine learning in mathematics has been limited. Specifically, in the field of information retrieval and semantic extraction, mathematics emerges among the complex because of the inherent ambiguity and context-dependence of mathematical language. Embeddings in MLP have been effective in capturing term similarity, math analogies, concept modeling, query expansion, and semantic knowledge extraction [29]. However, even with the success of embeddings in MLP, the need for extensive math-specific benchmarks is still prevalent. There is an extensive need for the development of implementable, robust embedding models that accurately represent mathematical artifacts and structures [30].

1.3. Current Study

We measure word and textual semantic similarity in threaded discussions within an online math tutoring platform using NLP and machine learning. The challenge is the scalability of state-of-the-art NLP models to the nuances of mathematical discourse [31]. NLP semantic models need to capture mathematical reasoning on top of extracting relevant information from natural language text [32]. As such, we model the similarity of the students' word choice and lexical processes (i.e., mental modeling and lexical representation of their ideas) within mathematical discourse gathered from Algebra I discussion boards in the Math Nation online tutoring system. In doing so, we determine the presence of shared language by looking for indicators of linguistic entrainment and semantic similarity (i.e., linguistic similarity). We also investigate words that reflect the lexical choices of the students, which could depict lexical entrainment. In addition, we investigate the relationship profiles between the shared language indicators and affect and trust as constructs that represent student engagement and affinity.

The research questions that guide this study are as follows:

1. Does discourse within an online Math tutoring platform exhibit shared language?
    1.1. Do the posts become more linguistically similar as the discussion progresses?
    1.2. Are words indicative of similarity between the discourse participants' posts?
2. How is linguistic similarity associated with known desirable social constructs (affect measures and trust) related to student engagement and feelings of affinity and belonging within discourse communities?

In answering these research questions, we address the limitations of discussion board deployments as a pedagogical tool for collaborative learning. We envision that student participation and engagement can be enhanced with the understanding of the online discourse in math discussion boards.

2. Methods

2.1. Math Nation

Math Nation [33] is an interactive and comprehensive math teaching and learning platform used extensively as either the main curriculum or supplemental material in all school districts in Florida. This learning platform provides video tutorials, practice questions, workbooks, and teacher resources aligned with the Florida Mathematics Standards. Student and teacher use of Math Nation in middle and high school has been shown to increase student achievement on the standardized algebra assessment required for high school graduation by the state of Florida [34,35]. Math Nation features an online discussion forum called Math Wall, where students can collaborate with other students. Participation in this student-facilitated discussion forum is incentivized by the karma points system, where every student who posts is awarded 100 points on their first post. Students subsequently earn more points when they offer meaningful advice or extended help, such as (1) asking guiding questions, (2) critiquing work, and (3) offering feedback. These types of posts often lead to students' queries being fully addressed. Further, Math Wall is designed such that course mentors play a minimal role. Mentors only post when necessary, and their posts are often related to moderating inappropriate online behavior or thread-management tasks. For example, when a student posts about irrelevant topics or replies to an old post, mentors will recall the attention of the students and remind them of the proper protocols for forum participation. Sample mentor posts include: "*The student ambassadors were correct originally*", "*Please don't reply to very old posts, you may start a new post instead*", and "*Where* *is that option? Check your scores*". In addition, the mentors also reply to unanswered and unresolved queries.

2.2. Participants

The participants included 12 Math Nation mentors who monitored the discussions and 1719 students in grade levels 7, 8, and 9 from multiple Florida school districts. This includes only the students who interacted with the Math Wall by either posting a question/query or replying to an original post to resolve a query, or, at times, just posting random replies. Student information was anonymized prior to the analyses; hence, only limited demographic information (e.g., gender, ethnicity, grade level, and school district) was made available.

2.3. Discussion Threads

Math Wall threaded discussions were collected for the 2020–2021 academic year. Discussion threads are chronologically ordered posts with an initial post that represents a query from a student. Other students subsequently post their replies, oftentimes in consideration of or elaborating on the most recent reply within a thread. A thread can be represented as an ordered list of posts (p_1, p_2, . . . p_n) that varies in length. Posts are timestamped to reveal their sequence, but the time interval between each post is varied. There are 4305 threads with a total of 50,975 posts, of which 4244 are the "initial" posts, and 46,731 are the "subsequent replies". The majority of these posts are from the students (50,950), with the mentors having only 25 posts. The average turn-around time from the initial post to the last reply in a thread is 5.72 h. However, some outlying threads took longer than usual, with the greatest time interval between the original post and the last post being 70 days and the shortest interval being 7 min. The number of posts in the threads ranged from 1 to 142, with a mean of 12 posts (SD = 14.5). With the high variability in the number of posts per thread, we used threads with at least 4 posts (1st quartile) and at most 16 posts (3rd quartile). Hence, 2139 threads were included in this analysis. These exclusion criteria eliminated threads that were extremely short or lengthy.

2.4. Natural Language Processing

The NLP methods used in this study were: (1) textual similarity calculation using pretrained deep NLP and neural language models (i.e., spaCy language model and Universal Sentence Encoder); (2) emotion feature extraction (i.e., valence, arousal, polarity) using VADER; (3) derivation of trust affect measures using EmoLex; (4) textual classification modeling using TF-IDF word vectors as features; and (5) explainable AI using the LIME package for model interpretation. In the next section, we describe these methods and how the combination of multiple NLP methods was adopted to address our research questions.

2.4.1. Feature Engineering

Textual similarity features. SpaCy [36] is an open-source NLP library for Python with pipeline components (for tokenization, part-of-speech tagging, parsing, and named entity recognition), modules, and language models that can be used for calculating semantic similarity. SpaCy calculates semantic similarity using cosine similarity as default with the word vectors and context-sensitive tensors in the vector space. Word vectors are multi-dimensional representations of words, which can be implemented using word2vec algorithms. There are many approaches to calculating sentence or document semantic similarity. Straightforward approaches use sentence embeddings or powerful models, such as transformers, to generate their embeddings and subsequently use any distance computation to calculate their similarity metric.

In this study, we purposely use both word embeddings and sentence embeddings to capture our description of semantic similarity involving similar lexical choices (word use) and semantics in terms of the overall meaning of an individual post. Hence, a high similarity score would indicate similar word use and similar meaning between the posts in comparison (i.e., high lexical and semantic similarity). We used spaCy with the en_core_web_lg (release 3.2.0) language model composed of pre-trained statistical models and word vectors. The spaCy pre-trained, multi-task CNN model has been trained on OntoNotes 5 in English with GloVe vectors trained on Common Crawl [37]. One characteristic of spaCy's word similarity calculations is that it averages the token vectors, with words as tokens, that comprise a document or a short sentence. Hence, even if the sentences are semantically different (i.e., have opposite meanings) but use the same words or use common words, spaCy will result in a high similarity score. In the same manner, comparing semantically similar sentences that use dissimilar words will result in a lower similarity score than sentences with different meanings but using the same words. To address this limitation brought about by spaCy's vector averaging, we further employed sentence embedding through the Universal Sentence Encoder (USE) [38,39] to improve spaCy's similarity score computation. Similarity scores of the sentence embeddings range between -1 and 1. Similarity scores closer to 1 reflect greater semantic similarity between items (i.e., posts in our case) and scores closer to -1 reflect extremely dissimilar posts. We used these features to investigate whether the discussion board discourse similarity converges, diverges, or remains the same as the discussion progresses. Further, these features were also used to discover inherent patterns that depict the phenomenon of interest—lexical entrainment or shared language within the Math Wall discourse.

Affect valence, polarity, and arousal features. VADER's Sentiment Intensity Analyzer was used to extract the sentiment polarity (compound score) of the posts. VADER is a rule-based model for general sentiment analysis that has outperformed human raters and generalizes across contexts [40]. VADER provides not just the polarity (compound score) of sentiments that reflect both the valence and arousal levels of a sentiment. VADER is known to accurately classify sentiments in social media contexts; thus, it was an appropriate choice due to the informal discourse in Math Wall, which has some semblance of social media stylistic language. We also used the expanded database for Affective Norms for English Words (ANEW) for valence and arousal scores [41]. To account for varying length, both valence and arousal scores were normalized. Both valence and arousal scores provided finer-grained insights even as the polarity (VADER's compound score) encompassed both measures (i.e., valence and arousal).

Trust features. Trust is a basic emotion and is observed in the social interactions of discourse communities. This study used the NRC Word-Emotion Association Lexicon (EmoLex). EmoLex [42] is a dictionary of words labeled via crowdsourcing that are associated with the 8 basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). This study only focused on trust. EmoLex also provides associations with either negative or positive sentiments but does not provide any sentiment or emotion intensity information. Some example words associated with trust are {advice, congruence, constant, eager, familiar, guide, and helpful}.

2.4.2. Text Classification

We also built a bag-of-words model to represent students' word choices as discourse participants. Bag-of-words is a representation that models the specific words used within documents and can be used to determine the company a specific word keeps across large corpora of texts and language models [27]. We particularly focused on the words they used to understand lexical choice and detect the presence of word mimicry within the threaded discussions. Word choice is a strong indicator of the speakers' lexical processes [43], and better word choice relates to increased precision of their lexical representations at the individual difference level [44]. From the bag-of-words model, we investigated the predictive role of words to aid in the understanding of the role that words play in textual similarity.

Python's Scikit learn library was used for the preprocessing pipeline (i.e., vectorization, feature selection, cross-validation, and logistic regression). Prior to any processing, the posts were anonymized and cleaned such that the data did not reveal the students' identity (i.e., names). The posts were also preprocessed to remove noise and stop words. The TF-IDF vectorizer is calculated from the cleaned posts' unigram and bigram counts. The Similarity outcome variable was coded as a dichotomous class, with **Not Similar** representing threads with a mean similarity score below the overall mean and **Similar** representing threads that are greater than or equal to the overall mean similarity score (i.e., 0 would be threads that are not semantically similar on average, and 1 would be threads that are, on average, semantically similar). We qualitatively evaluated the validity of the dichotomous classes by randomly checking a select number of posts (n = 20) for each class and found that the dichotomous categories appropriately represented the semantic similarity of the threaded discussions. We ensured that the qualitative verification involved threads with outlying extreme aggregated similarity values and values that are nearest to the dichotomization cut-off/threshold mean similarity score. The feature selection/dimensionality reduction method used was SelectKBest, which selects the best features based on the X^2 value. Using the reduced feature matrix, the logistic regression classification model was trained with 10-fold cross-validation at the thread-level repeated three times. Finally, the model performance was evaluated using the F1-score as the metric.

2.4.3. Model Explainability

The Local Interpretable Model-Agnostic Explanations (LIME) package was used to interpret and explain model classifications. While we were able to derive the most relevant words in the model (best features based on X^2), we wanted to supplement the findings by understanding word contributions at the local prediction level. LIME provides a means to explain and visualize individual predictions. Even with acceptable levels of accuracy and other measures of performance, model explainability at prediction-level granularity provides more relevant and useful insights. LIME interprets individual instances of predictions without looking into the model and learns an interpretable model to explain individual predictions [45]. We ran repeated iterations of LIME to find more stable results to facilitate the interpretations.

2.5. Cluster Analysis

One goal was to discover innate patterns that potentially exist between linguistic similarity and valence, polarity, arousal, and trust and extract underlying thread profiles. We implemented a two-step cluster analysis as it performs better than traditional hierarchical clustering methods [46]. This technique determines the optimal number of clusters based on the Bayesian information criterion (BIC) as the measure of fit [47]. Further, we measured cluster quality fitness using the silhouette measure of cohesion and separation to evaluate the overall cluster model quality. In defining the thread profiles based on the resulting clusters, we referenced relevant theories and prior work.

3. Results

3.1. Similarity as Proxy for Shared Language in Discourse

RQ#1 Does discourse within an online Math tutoring platform exhibit shared language?

To answer this question, we calculated linguistic similarity scores to determine if the discourse participants' posts are linguistically similar within and across the threaded discussions. In examining linguistic similarity, we probed into domain-specific semantics used in the Math Wall discourse. Similarity scores were investigated at two levels: (1) similarity between the posts within a threaded discussion and (2) similarity between the threads that comprise the Math Wall discourse. Within both levels, we observed that shared language is present in the Math Wall discourse.

Post Similarity Score. The similarity calculations derived are the similarity of the replies to the original post (OP similarity) and the similarity of the current post to the immediately succeeding post (P2P similarity; post p_n to p_{n+1}). The sequence of pairwise similarity scores reflects the sequence of the posts as revealed in the timestamps. The means of both post similarity measures have minimal differences (OP similarity M = 0.1777; SD = 0.0915;

min = −0.0520; max = 0.7618; P2P similarity M = 0.2108; SD = 0.1134; min = −0.0627; max = 0.7380). The two post similarity scores are positively correlated (r = 0.41, $p < 0.001$). The minimal difference and the positive correlation between OP similarity and P2P similarity reflect that the similarity of individual posts to the original post is almost the same as the similarity of individual posts to subsequent replies (posts). Hence, if a post is similar to the original post, then similarity with the succeeding post is also noted.

Thread Similarity Score. The thread similarity was also calculated in a pairwise fashion such that each thread was compared to all threads in the data (M = 0.3183; SD = 0.1532; min = −0.1834; max = 0.9785). The thread similarity scores exhibit extremely high dispersion with many outlying threads in terms of their pairwise similarity scores. Upon qualitative evaluation of the actual posts, these threads with very high standard deviations (SD > M) are comprised of posts characterized by different communicative intentions and seemingly unrelated. A specific example would be the pairwise comparison between a thread discussing rational numbers versus a thread containing some random or non-algebra-related posts (e.g., "*wassup, wassup*; *how is everyone?*; *goodbye*"). The spread of the thread similarity scores indicates that the Math Wall discourse is comprised of a combination of discussion threads that exhibit shared language and those that do not.

Similarity Scores Variance Analysis

An analysis of the variance of the post similarity scores across the threads was conducted to determine if there are significant differences between the variances. The Shapiro–Wilk test of normality on the similarity scores grouped by thread determined that not all the threads are normally distributed. We do not present the descriptive statistics for the post similarity scores by thread here due to the large number of threads. Subsequently, performing a normality test for the similarity scores (without grouping by thread) also renders a non-normal distribution (statistically significant at $p < 0.001$). Furthermore, we analyzed the variance of both post and thread similarity scores using unequal sample sizes and found statistically significant differences in the variance across the threads (Kruskal–Wallis X^2 (2110) = 3192, $p< 0.001$; Levene's F statistic (2110, 11797) = 2). These results reiterate the significant differences in the variances of similarity scores of posts within the threads. In other words, the linguistic similarity scores of the posts were different across the discussion threads, such that some threads were linguistically similar and some were not.

RQ#1.1 Do the posts become more linguistically similar as the discussion progresses?

Post Similarity across Sequence. We investigated whether the posts (1) become more similar as the discussion unfolds, (2) retain their level of similarity, or (3) become less similar. To answer RQ#1.1, we examined the linear trend of similarity scores over time. From visual inspection of the means of the scores across sequences 1 to 15, a linear trend was observed with notable shifts as the discussion unfolded (see Figure 1). However, this trend cannot be established or validated, even with attempts to calculate linear contrasts, because the homogeneity of variance assumption is violated (Levene's statistic: F (14, 13897) = 2.774, $p < 0.001$; Welch's F (14, 2759) = 1.701, $p = 0.051$). Hence, it cannot be determined if post similarity behaviors exhibit a linear trend, such as becoming more similar, retaining the same level of similarity, or becoming less similar as the discussions unfold.

A post hoc qualitative inspection of the discussion transcripts supports the absence of this linear trend. It is observed that there is usually a notably high similarity at the beginning of the threads, demonstrating that threads usually open with either a common greeting or an elaboration and clarification of the query posted. Only until this common understanding of the question or query is reached will the posts be reflective of diverse answers or solutions wherein individuals contribute to the problem-solving task in the current thread. Hence, shifts in similarity are observed as the discussion progresses. Towards the end, there is also a shift where similarity tends to increase. As in the case of the initiations of discussions, the endings also exhibit common terms and statements by participants, such as *thank you* or *goodbye*. However, this pattern of communication flow is not present in all discussion threads.

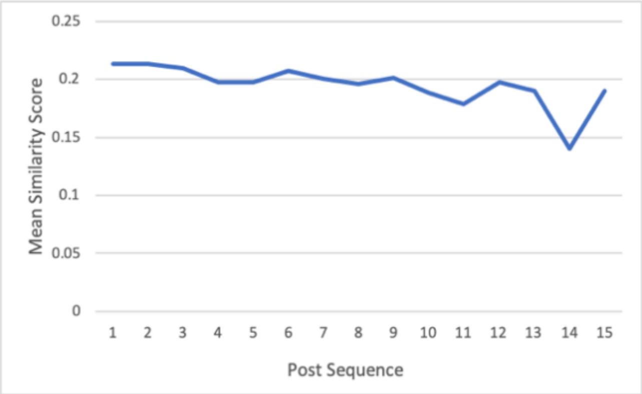

Figure 1. Mean Similarity Score across Sequences (from initial post to the 15th post).

Post Similarity vs. *Thread Similarity.* Both similarity scores depict relatively high variability. However, the variability between threads (thread similarity) is notably higher than the variability within threads (post similarity). This outcome corroborates the lexical entrainment phenomena, which describes how discourse participants adapt their lexical choices to shared conceptualizations within conversations, and, in effect, there is more observed variability across conversations than within conversations [19]. Hence, for the Math Wall discourse, posts within a thread were more linguistically similar. On the other hand, the threads were less similar overall, as compared to posts within threads.

3.2. Similarity Predictive Model

RQ#1.2 Are words indicative of similarity between discourse participants' posts?

We modeled P2P similarity, (i.e., post p_n to p_{n+1}) to determine the turn-by-turn progression of similarity. First, we generated a dichotomous similarity outcome variable by categorizing the similarity scores into two bins (i.e., similar and dissimilar). The category was similar if the score was greater than or equal to the average score; otherwise, the category was dissimilar. This dichotomous categorization distinguishes threads that depict lexical entrainment and shared language versus the threads that were less similar than the average tendency. We then extracted the TF-IDF features from the text and used SelectKBest to use the best features (i.e., higher X^2 values) for classification. The top 10 best word features (i.e., unigrams and bigrams) are presented in Table 1. Logistic regression was used to fit the model with three repeats of 10-fold cross-validation. The final logistic regression model achieved a good fit (accuracy = 0.73, F1-score = 0.67).

Table 1. Top 10 Features (n-grams) based on X^2 values.

Feature	X^2	Feature	X^2
square	3.20	make sure	2.48
square root	3.12	graph	2.47
pemdas	3.07	coordinates	2.35
alright	2.97	equation	2.21
hello	2.55	subtracting sides	1.98

Local Analysis of Predictions

We used the LIME package to explain specific instances of the predictions of the classification model. Example LIME results for three arbitrarily chosen threads are presented in Tables 2–7.

Table 2. Sample Thread 1.

Post	Where can I find help with properties of exponents?
Reply	**Section** 1 **topic** 2
Reply	You can find it by pressing **Section** 1 and then pressing the video of **topic** 2.
Reply	Hi, you can look at the 8th **grade section** and it should be **section** 8.7 of that
Reply	Those are all about **exponents**
Reply	you can refer to **section** 1 **topic** 2
Reply	**Section** 1 **topic** 2,4,5,6 could **help**.

Table 3. Word Probabilities for Prediction on Thread 1.

Most Informative Words	
Class: Not Similar	Class: Similar
'section', −0.0589	'grade', 0.0251
'topic', −0.0397	'help', 0.0211
'property', −0.0219	'exponent', 0.0157

Table 4. Sample Thread 2.

Post	The police department is having a bake sale. Donuts cost \$1.50 each and cinnamon roll costs \$2.50 each. The department uses the algebraic expression 1.50
Reply	The **question** isn't finished... Can you put the rest?
Reply	The police department is **having** a bake sale. Donuts cost \$1.50 each and cinnamon rolls cost \$2.50 each. The department uses the algebraic **expression** 1.50
Reply	a. What does the x **variable represent**? b. What does the y **variable represent**? c. A family buys 3 donuts and 4 cinnamon rolls. What are their total expenses?
Reply	The department uses the algebraic **expression** 1.50

Table 5. Word Probabilities for Prediction on Thread 2.

Most Informative Words	
Class: Not Similar	Class: Similar
'expression', −0.06372	'having', 0.0129
'represent', −0.03780	'question', 0.0060
'variable', −0.0193	

Table 6. Sample Thread 3.

Post	where can i find a section on integers?
Reply	What do you **mean** by **integer**s? Rational and **Irrational**?
Reply	ex. $-3 + -4 = -7$
Reply	I think that deals with rational and **irrational number**s, section 1 topic 3
Reply	ok **thank**s
Reply	No **problem**, **glad** that I could **help**

Table 7. Word Probabilities for Prediction on Thread 3.

Most Informative Words	
Class: Not Similar	Class: Similar
'section', −0.0360	'glad', 0.0365
'thank', −0.0290	'help', 0.0315
'mean', −0.0202	'integer', 0.0263
'problem', −0.0160	'Irrational', 0.0195
	'number', 0.0186

Prediction on Thread 1 (see Table 2*)*. The thread's true class is **Similar** (Pr(**Similar**) = 0.4230), and it has been incorrectly predicted as **Not Similar** (Pr(**NotSimilar**) = 0.5770). However, looking at the probabilities, the difference between the classes is relatively small. The words that contribute to the classification decision are presented in Table 3.

Prediction on Thread 2 (Table 4). The thread's true class is **Not Similar** (Pr(**Not Similar**) = 0.7195; (Pr(**Similar**) = 0.2805), and it is correctly predicted as such. It can be observed that in this correct prediction, the probability of the true class is notably higher than the probability of the incorrect class. The words that contribute to the prediction are presented in Table 5.

Prediction on Thread 3 (Table 6). The thread's true class is **Not Similar,** and it has been correctly predicted as such Pr(**Not Similar**) = 0.5028; Pr(**Similar**) = 0.4972. The difference between the probabilities of both classes is relatively small. The words that contribute to the prediction are presented in Table 7.

The specific predictions illustrate that words pertaining to algebra constructs are the words that are more likely to contribute to classifying the posts in terms of semantic similarity (e.g., "square", "square root", "exponent"). There are also words that occur more often (e.g., "help", "section", "problem") and thus contribute to determining similarity. The top-contributing words to the predictions reveal that the semantic context of the words, relative to the context of the discussion thread, is correctly captured by the classification model. For example, words (or groups of words) that reference an unrelated section, such as the reference to "section 8.7" within thread 1, which is about "section 1", contribute to the dissimilarity of the discussion. In addition, words such as the algebraic terms used in thread 3, e.g., "integer" and "number", contribute to the similarity of the discussion.

3.3. Cluster Analysis

RQ2: How is linguistic similarity associated with known desirable social constructs (affect measures and trust) related to student engagement and feelings of affinity and belonging within discourse communities?

The desirable social constructs included in this analysis are constructs that are associated with the shared language of discourse communities (i.e., valence, arousal, polarity, and trust). The descriptive statistics of these constructs are presented in Table 8. These constructs are not correlated with linguistic similarity scores. Valence and arousal are highly correlated, depicting that the discourse participants expressed words reflecting engagement and arousal when posts were strongly positive or strongly negative. The correlations between these constructs are presented in Figure 2.

Using Schwarz's Bayesian criterion (BIC), the resulting optimal number of clusters is 3 (ratio of BIC change = 0.527) with a fair Silhouette score (Silhouette Score (n = 3): 0.24), revealing some slight cluster overlapping. We performed two-step clustering using the larger collection of threads (total: 3830; number of posts $\geq$ 2). The cluster distribution and profiles (centroids) are presented in Tables 9 and 10, respectively. We report the significant differences between the clusters in terms of all five features and effect sizes in Appendix B. A Bonferroni post hoc test illustrates the specific differences between the mean differences of the three clusters (p < 0.001). Specifically, similarity and trust were significantly different for all three clusters. Valence and arousal were significantly different for clusters 1 and 3. Polarity was significantly different for clusters 1 and 3, as well as clusters 2 and 3. To visualize these differences, we present the cluster mean plots for the five input variables for each cluster (see Figure 3).

Table 8. Descriptive Statistics of the Cluster Model Inputs.

	Mean	**SD**	**Min**	**Max**
Similarity	0.1777	0.0915	−0.0520	0.7618
Trust	0.1617	0.0200	0	0.1622
Valence *	0.1482	0.1470	0	1.1824
Arousal *	0.1231	0.1199	0	0.9200
Polarity	0.1212	0.1259	−0.5303	0.8530

* Normalized measures divided by the number of words.

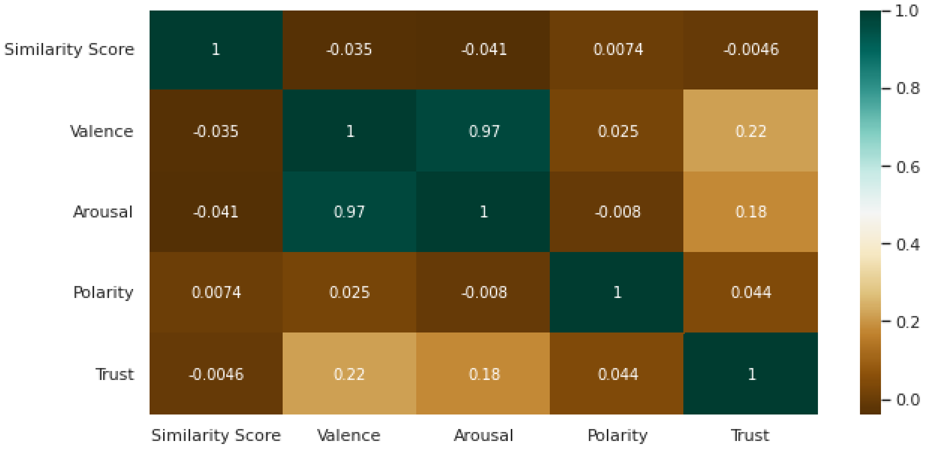

Figure 2. Heatmap of Correlation between Similarity, Valence, Arousal, Polarity, and Trust.

Table 9. Cluster Membership Distributions.

Cluster	**N**	**% of Total**
1	974	24.4
2	461	12.0
3	2395	62.5
Total	3830	

Table 10. Cluster Centroids.

Cluster	**Similarity M (SD)**	**Trust M (SD)**	**Valence M (SD)**	**Arousal M (SD)**	**Polarity M (SD)**
1	0.2944 (0.1496)	0.0187 (0.0267)	0.0430 (0.0725)	0.0363 (0.0620)	0.0143 (0.0206)
2	0.2259 (0.1486)	0.0330 (0.0438)	0.4642 (0.2690)	0.3780 (0.2073)	0.0151 (0.0173)
3	0.1539 (0.0495)	0.0130 (0.0124)	0.1288 (0.0888)	0.1073 (0.0736)	0.0110 (0.0007)

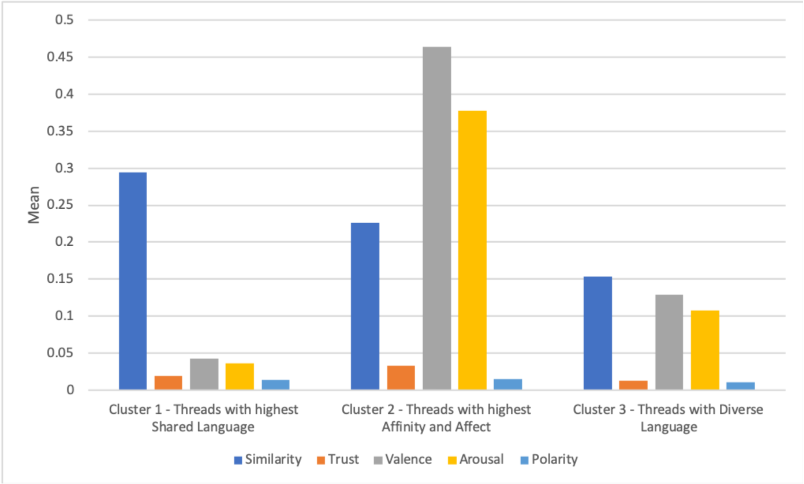

Figure 3. Cluster Plots.

Cluster 1 threads exhibit the most lexical entrainment and have the highest similarity, lowest valence, arousal and polarity, with trust having middle values as compared to the other two clusters. In the sample thread presented in Appendix A—Table A1, the negative affect, confusion, is manifested. Cluster 2 reveals threads that exhibit the most affinity, emotion, and (some level of) lexical entrainment. These threads have the highest trust, valence, arousal, and polarity. Cluster 2 also has relatively high similarity, only slightly lower than the similarity means of Cluster 1. While Cluster 2 had the fewest members (12% of the total number of threads), we find this cluster the most interesting because it corroborates work that indicates trust is associated with feelings of belongingness and affinity to a group or community. High valence and polarity indicate positive emotions. High arousal, when associated with high valence, suggests the intensity of positive emotions, which may reflect engagement. The sample thread in Appendix A—Table A2 manifested trust, appreciation, and collaborative meaning-making between the students. Cluster 3, with the greatest number of members (62.5% of the total number of threads), has the lowest trust and similarity and neutral affect measures, which is indicative of diversity in participants' lexical choices and a lower sense of affinity among the discourse participants (see Appendix A—Table A3 for an example).

4. Discussion

The objective of this study was to investigate the presence of shared language or linguistic similarity in students' posts as they participated in threaded discussions within Math Wall. We analyzed the behavior of linguistic similarity as the discussion progressed. In examining linguistic similarity, we examined math-specific semantics (meaning) and lexical choices of the discourse. Further, through cluster analysis, we investigated the relationship profiles between linguistic similarity as a shared language indicator and constructs that represent student engagement and affinity, namely trust and affect measures of valence, arousal, and polarity.

Shared language in the Math Wall discourse

We found that the Math Wall threaded discussions exhibit shared language. Linguistic similarity exists between the posts in a threaded discussion. Significant differences in the linguistic similarity within the threads are found for the two post similarity scores (i.e., OP similarity and P2P similarity). The observed high similarity at the beginning of the discussion was qualitatively verified to match the nature of the threaded discussions. Specifically, individuals aim to derive an understanding of the query in the initial post and, hence, subsequent replies tend to repeat, rephrase, or elaborate on the initial posts. Subsequent posts are, then, characterized by individual solutions and responses to solve the query. This phenomenon of higher similarity scores at the beginning of turn-by-turn discussions is also observed in similar work by Liebman and Gergle [8]. Further, toward the end of the discussion, there is also an observable shift to increased similarity. This is again qualitatively explained by a common practice of expressing gratitude and appreciation or bidding goodbye as the discussion ends.

Overall, the high variability of thread similarity reveals the dynamic nature of how individuals contribute to these threaded discussions. Even when there is one common goal (i.e., a specific query), some threads (14%) have highly cohesive discussions from beginning to end (i.e., low variability in similarity). Some threads (27%) have varied views and then reach an agreement (i.e., converging similarity), while others (42%) do not come to an understanding (i.e., diverging similarity). Some (17%) start out as confused, then end up in agreement or resolve the confusion (i.e., increasing similarity). These scenarios depict similarity shifts and the high variability of similarity among the threads. Hence, the threaded discussions do not always become more linguistically similar as the discussion progresses.

Lexical choice as a predictor of linguistic similarity

We found that deriving lexical and semantic textual similarity with the state-of-the-art word and embedding models successfully captured the domain-specific semantics within the Math Nation discourse. The non-latent words used in the discourse can predict similarity. Further, the use of the bag-of-words (TF-IDF) model in fitting the predictive model contributed to insights on the nature of the lexicons used in the discourse that aligned with the algebra curriculum and reflected the specific mathematical content and actions. Studying the predictions at the individual level using LIME revealed that words represent algebra constructs (e.g., exponent, integer, irrational), and words that depict frequent action (e.g., help) contribute to a prediction of 1 (i.e., high likelihood of similarity). This can be explained by the students' reference to uniform math constructs and common action words when discussing the topic at hand. Conversely, words that represent specific artifacts in Math Nation (e.g., a section, topic, or problem) contribute to a prediction of 0 (low likelihood of similarity). This could be attributed to the diverse solutions referencing various parts of the curriculum that discourse participants suggest in their attempts to solve a problem.

We also found that specific words can efficiently distinguish the likelihood of similarity of the posts within threads. These words, in particular, are math constructs (rational, square root), Math Nation artifacts (e.g., section, unit), emotion (e.g., glad), and action (e.g., help, mean). These agreed-upon lexical choices echo lexical entrainment as the students did not identify what specific keywords or terms they would adhere to in the conversation. For example, one thread would unanimously use *raise to the 3rd* power, while another thread would use *cube* root. The presence of commonly used and "implicitly agreed-upon" words within a threaded discussion explains the phenomenon of lexical choices that contribute to linguistic convergence [19,48,49].

Linguistic similarity and desirable social constructs

The profiles that emerged, in terms of similarity and the desirable emotional measures of interest (i.e., valence, intensity, polarity, and trust), demonstrate the differences in the threaded discussions. The majority of the threads in the third cluster had the least similar words indicative of trust; however, they were "middle ground" compared to the threads from the other two clusters in terms of emotion measures. After post hoc qualitative analysis, we found that these threads were indicative of posts from students who were confused, did not understand concepts, or had unresolved queries who posted questions hoping to find answers. The second cluster had the highest means for all affect features, including trust, had a relatively high mean similarity, and was comprised of the fewest threads. These threads were cohesive threads (i.e., relatively high similarity) and had the highest presence of words depicting trust. These threads also had the highest means for features of emotion (valence, arousal, intensity) and depicted ideal discussions where collaborative meaning-making was present. The posts reflect a coherent discussion on a single topic and reflect a common goal. This profile shows that Math Wall had some discussion threads (although only 12% of the total number of discussion threads) that depict desirable learning behaviors, i.e., collaboration and coherent discourse. Coherent academic discussions where participants negotiate agreements or disagreements are essential in maintaining the trust and inclusiveness of discourse communities [50]. Cooperative learning helps students with math anxiety and encourages help-seeking behavior [51]. We note from our qualitative evaluation that while some threads resolved the queries in the initial post, some threads remained unresolved, but the progression of the discussion revolved around solutions where the participants collaborated in an attempt to achieve resolution.

The clusters that were significantly different in terms of similarity and trust corroborates prior work [20,21] that describes the trust that members of discourse communities exhibit. Trust, as a social affect, is a factor that motivates students to seek help from their peers within digital and non-digital learning environments [52]. The majority of the discussion threads, i.e., those belonging to Clusters 1 and 2, had low trust values. Math Nation could implement design changes to improve trust within Math Wall. For example, increasing the participation of mentors, study experts, and teachers may increase the level of trust in Math Wall as a student-led discussion board. Emotional valence, polarity, and intensity, along with similarity, uniquely describe and differentiate the three clusters. The clusters representing the profiles of discussion threads show that high similarity evokes both positive and negative emotions that may be intense or subtle. Threads that have high similarity and relatively intense, positive emotions were threads that showed delight in the queries being answered and problems being resolved. At the other end of the spectrum, threads that have high similarity and low-intensity negative emotions were characterized by posts that pertain to difficult concepts where confusion is apparent. It should be noted that the posts were still relevant to the topic at hand, but confusion was present. The low arousal scores of these threads show that even as negative emotions are present, the emotions are at a subtle degree to not bring about frustration and other more intense negative emotions. Low similarity associated with low trust described threads that did not have relevant or related posts. These threads did not resolve any Algebra query nor address any concern, and oftentimes were composed of random posts overall.

5. Limitations, Conclusions, and Future Work

As with all studies, this study has several limitations. First, although the sample is considerably large, the threads are vastly unequal in terms of the number of posts. The large variance and small group sizes, when grouped according to threads and sequence, contributed to marginally significant findings in variance and trend analysis. This is typical of many discussion boards but also begs caution, given that this variance may restrict the likelihood of generalization to other discussion boards. Second, the temporal aspect of the posts was not included in this analysis. The threads have unevenly spaced posts and unequal numbers of posts. For this work, our objective was to characterize the discourse, in general, in terms of targeted language and emotional features independent of their temporal arrangement. We recognize the value of conducting a temporal analysis as future work to derive insights on the implications of timely resolution of queries on discussion boards. Lastly, this study aims to specifically focus on the similarity of content as a valuable construct in discussion board implementations and does not include investigations regarding the impact towards desirable learning outcomes. One reason for this choice is practical; it is currently unavailable in this dataset. Nonetheless, another reason is because our overarching objective is to understand and depict the nature of collaborative language, which may or may not be associated with outcomes depending on multiple complex factors (e.g., student skills and classroom contexts). As we better understand and characterize the nature of this language, we will be better situated to explore the relations between the nature of language and learning outcomes, such as student performance measures.

Implications for Discussion Boards as a Pedagogical Tool

The diverse nature of the discourse, as measured by semantic similarity at the word and semantic textual levels, depicts the dynamic nature of the discussions. While most queries (initial posts) were properly addressed and eventually resolved, the students did not show that they were particularly constrained to only academically related and relevant posts. The students, at times, interjected "unrelated" replies to display humor or just entirely random statements. The threads can be characterized as a rich and diverse collection of posts. Further, while only marginally significant, there was an observed trend of threads starting with high similarity followed by shifts to decreased similarity as the discussion progressed, followed again by shifts to an increase in similarity toward the end of the discussion. We verified this trend with a qualitative validation of the threads depicting the same trend and found that in the early part of the discussion, similarity is "maintained", and a slight increase or decrease can manifest as an elaboration or a summarization transpires. However, towards the end of the discussion, the thread's topic, context, and wordings shift to non-algebra-related content, as most threads end with posts such as "*thank you*", "*goodnight*", "*glad I could help*", and "*bye*". This explains the shift in the word and textual similarity scores as the discussion comes to a close. The unconstrained and free nature of discussion boards may encourage student participation. Even as students occasionally post directly related or irrelevant replies, student engagement is still better than non-participation. Discussion boards, after all, are venues for social relationships within virtual environments.

The presence of similarity and emotional constructs provide evidence of the value that linguistic similarity could bring to the design of discussion boards. Given the importance of trust and positive emotions in contributing to students' feelings of affinity [20,21], future training of instructors and tutors who monitor discussion boards should focus on proper scaffolding of language and leveraging linguistic similarity when responding to students. Specific approaches would include the use of common and widely used words, adding to or referencing a prior idea, elaborating or summarizing a prior idea, and ensuring the resolution of a query within a reasonable timeframe.

Implications to Scalability of NLP Language Models for the Mathematics Domain

State-of-the-art semantic extraction and information retrieval have made grand strides in terms of capturing context across domains. For mathematical domains, however, there is much room for the desired scalability. The complexity of encompassing mathematical context within deep NLP remains challenging because of the inherent structure of mathematical texts [29,30,53]. Our findings corroborate this scalability challenge. While we found that Math Wall discourse similarity was efficiently captured (i.e., similarity scores reflected the presence of similar references to common algebra constructs and intent and actions on math problem-solving activities), we found instances of this scalability problem in our analysis. Specific mathematical nuances that were not captured include relatively lower similarity scores for posts with mathematical jargon referring to the same constructs as "PEMDAS (order of operations)" and "hierarchy of operations", as well as "cubed" and "3rd power". Although, in this study, we addressed this seeming limitation with the use of the bag-of-words (TF-IDF) model, which provided additional insights into the specific words used that existed in our dataset. Our findings, therefore, highlight the need for math-specific deep NLP lexical and semantic models.

In conclusion, we envision and encourage future research in two equally important directions: (1) examining the extent to which the shared language within Math Nation's discourse community is also descriptive of other instances of math discourse communities and (2) building robust deep NLP language models that adapt specifically to math domains and tasks.

Author Contributions: Conceptualization, M.P.B. and D.S.M.; methodology, M.P.B., J.S., W.L.L.; validation, R.B., T.A. and W.L.L.; formal analysis, M.P.B. and J.S.; data curation, M.P.B.; writing—

original draft preparation, M.P.B.; writing—review and editing, J.S., T.A., R.B., W.L.L. and D.S.M.; supervision, D.S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Institute of Education Sciences (IES Grant: R305C160004). Opinions, conclusions, or recommendations do not necessarily reflect the views of the IES.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Arizona State University (STUDY00009419 approved on 1/28/2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to IRB restrictions.

Acknowledgments: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C160004 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Sample Threads Per Cluster

Table A1. Sample Cluster 1 Thread.

Post	I am so confused please help
Reply	Hello <<*Student1*>>, have you watched Section 1, Topic 2?
Reply	Yes, a while ago. I can go watch it again.
Reply	No, that's ok
Reply	Okay, thank you!
Reply	Or you could. Alright, if you still have questions, be sure to come back

Table A2. Sample Cluster 2 Thread.

Post	I need help solving this.
Reply	Hello <<*Student1*>>, your first step is to remove the parenthesis around the exponent 1, so the exponents combine to make 1 times 3/4
Reply	Ok after I do that what do I do?
Reply	Ok, so then I get 3/4 as the answer right?
Reply	Yes I got that
Reply	So, (1/p squared+ 1/p) to the power of p
Reply	I am stuck on this problem too. What would be the next step?
Reply	(28/9) raised to the power 3/4?
Reply	Thank you
Reply	After that step I got (28/9) raised to the power 3/4
Reply	any way to show how you got to 28/9?
Reply	Is the final answer (3.111) raised to the power $\frac{3}{4}$
Reply	Thank you very much
Reply	<<*Student2*>> when you do (1/9/16 + 1/3/4) to the power of 3/4 you will get (28/9) raised to the power $\frac{3}{4}$
Reply	Thank you, <<*Student1*>>

Table A3. Sample Cluster 3 Thread.

Post	How do i solve a to the second power x a to the third power?
Reply	do you mean
Reply	Yes
Reply	when multiplying same base exponents, you add the exponent values
Reply	x¬≤ *x= x¬≥
Reply	were you talking to me? that was an example
Reply	its ok

Appendix B. Two-step Cluster Analysis ANOVA and Effect Sizes

Table A4. ANOVA and Effect Sizes.

Construct		**Sum of Squares**	**df**	**Mean Square**	**F (p < 0.001)**	**Eta-Squared**
Similarity	Between Groups	14.069	2	7.035	761.140	0.285
	Within Groups	35.370	3827	0.009		
	Total	49.439	3829			
Trust	Between Groups	0.158	2	0.079	155.809	0.075
	Within Groups	1.944	3827	0.001		
	Total	2.102	3829			
Valence	Between Groups	57.741	2	28.871	1928.289	0.502
	Within Groups	57.299	3827	0.015		
	Total	115.040	3829			
Arousal	Between Groups	37.885	2	18.943	1987.085	0.509
	Within Groups	36.482	3827	0.010		
	Total	74.368	3829			
Polarity	Between Groups	1.480	2	0.740	41.546	0.021
	Within Groups	68.180	3827	0.018		
	Total	69.661	3829			

References

1. Chi, M.T.; Wylie, R. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **2014**, *49*, 219–243. [CrossRef]
2. Menekse, M.; Chi, M.T. The role of collaborative interactions versus individual construction on students' learning of engineering concepts. *Eur. J. Eng. Educ.* **2018**, *44*, 702–725. [CrossRef]
3. Roscoe, R.D.; Gutierrez, P.J.; Wylie, R.; Chi, M.T. *Evaluating Lesson Design and Implementation within the ICAP Framework*; International Society of the Learning Sciences: Boulder, CO, USA, 2014.
4. D'Angelo, S.; Gergle, D. Gazed and confused: Understanding and designing shared gaze for remote collaboration. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 2492–2496.
5. Bizzell, P. *Academic Discourse and Critical Consciousness*; University of Pittsburgh: Pittsburgh, PA, USA, 1992.
6. Hyland, K. Academic discourse. In *Continuum Companion to Discourse Analysis*; Bloomsbury Publishing: London, UK, 2011; pp. 171–184.
7. Mauranen, A. A rich domain of ELF-the ELFA corpus of academic discourse. *Nord. J. Engl. Stud.* **2006**, *5*, 145–159. [CrossRef]
8. Liebman, N.; Gergle, D. Capturing turn-by-turn lexical similarity in text-based communication. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, San Francisco, CA, USA, 27 February–2 March 2016; pp. 553–559.
9. Palloff, R.M.; Pratt, K. *Building Online Learning Communities: Effective Strategies for the Virtual Classroom*; John Wiley & Sons: Hoboken, NJ, USA, 2007.

10. Chen PS, D.; Lambert, A.D.; Guidry, K.R. Engaging online learners: The impact of Web-based learning technology on college student engagement. *Comput. Educ.* **2010**, *54*, 1222–1232. [CrossRef]
11. Hernández-Lara, A.B.; Serradell-López, E. Student interactions in online discussion forums: Their perception on learning with business simulation games. *Behav. Inf. Technol.* **2018**, *37*, 419–429. [CrossRef]
12. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R. Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Comput. Intell. Neurosci.* **2018**, *2018*, 6347186. [CrossRef]
13. Romero, C.; López, M.I.; Luna, J.M.; Ventura, S. Predicting students' final performance from participation in on-line discussion forums. *Comput. Educ.* **2013**, *68*, 458–472. [CrossRef]
14. Yukselturk, E. An investigation of factors affecting student participation level in an online discussion forum. *Turk. Online J. Educ. Technol.-TOJET* **2010**, *9*, 24–32.
15. D'Mello, S.K.; Graesser, A. Language and discourse are powerful signals of student emotions during tutoring. *IEEE Trans. Learn. Technol.* **2012**, *5*, 304–317. [CrossRef]
16. Garrod, S.; Pickering, M.J. Why is conversation so easy? *Trends Cogn. Sci.* **2004**, *8*, 8–11. [CrossRef]
17. Gonzales, A.L.; Hancock, J.T.; Pennebaker, J.W. Language style matching as a predictor of social dynamics in small groups. *Commun. Res.* **2010**, *37*, 3–19. [CrossRef]
18. Garrod, S.; Anderson, A. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition* **1987**, *27*, 181–218. [CrossRef]
19. Brennan, S.E. Lexical entrainment in spontaneous dialog. *Proc. ISSD* **1996**, *96*, 41–44.
20. Scissors, L.E.; Gill, A.J.; Geraghty, K.; Gergle, D. In CMC we trust: The role of similarity. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Boston, MA, USA, 4–9 April 2009; pp. 527–536.
21. Scissors, L.E.; Gill, A.J.; Gergle, D. Linguistic mimicry and trust in text-based CMC. In Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, San Diego, CA, USA, 8–12 November 2008; pp. 277–280.
22. Friedberg, H.; Litman, D.; Paletz, S.B. Lexical entrainment and success in student engineering groups. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT), Miami, FL, USA, 2–5 December 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 404–409.
23. Liu, Y.; Li, A.; Dang, J.; Zhou, D. Semantic and Acoustic-Prosodic Entrainment of Dialogues in Service Scenarios. In Proceedings of the Companion Publication of the 2021 International Conference on Multimodal Interaction, Montreal, QC, Canada, 18–22 October 2021; pp. 71–74.
24. Lin, D. An information-theoretic definition of similarity. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, WI, USA, 24–27 July 1998; Volume 98, pp. 296–304.
25. Princeton University. *"About WordNet." WordNet*; Princeton University: Princeton, NJ, USA, 2010.
26. Pawar, A.; Mago, V. Challenging the boundaries of unsupervised learning for semantic similarity. *IEEE Access* **2019**, *7*, 16291–16308. [CrossRef]
27. McNamara, D.S. Computational methods to extract meaning from text and advance theories of human cognition. *Top. Cogn. Sci.* **2011**, *3*, 3–17. [CrossRef]
28. Banawan, M.; Shin, J.; Balyan, R.; Leite, W.L.; McNamara, D.S. Math Discourse Linguistic Components (Cohesive Cues within a Math Discussion Board Discourse). In Proceedings of the Ninth ACM Conference on Learning@ Scale, New York, NY, USA, 1–3 June 2022; pp. 389–394.
29. Greiner-Petter, A.; Youssef, A.; Ruas, T.; Miller, B.R.; Schubotz, M.; Aizawa, A.; Gipp, B. Math-word embedding in math search and semantic extraction. *Scientometrics* **2020**, *125*, 3017–3046. [CrossRef]
30. Jo, H.; Kang, D.; Head, A.; Hearst, M.A. Modeling Mathematical Notation Semantics in Academic Papers. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 3102–3115.
31. Ferreira, D.; Freitas, A. Premise selection in natural language mathematical texts. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 6–8 July 2020; pp. 7365–7374.
32. Patel, A.; Bhattamishra, S.; Goyal, N. Are NLP Models really able to Solve Simple Math Word Problems? *arXiv* **2021**, arXiv:2103.07191.
33. Algebra Nation. Available online: https://lastinger.center.ufl.edu/mathematics/algebra-nation/ (accessed on 22 January 2021).
34. Leite, W.L.; Jing, Z.; Kuang, H.; Kim, D.; Huggins-Manley, A.C. Multilevel Mixture Modeling with Propensity Score Weights for Quasi-Experimental Evaluation of Virtual Learning Environments. *Struct. Equ. Model. A Multidiscip. J.* **2021**, *28*, 964–982. [CrossRef]
35. Leite, W.L.; Cetin-Berber, D.D.; Huggins-Manley, A.C.; Collier, Z.K.; Beal, C.R. The relationship between Algebra Nation usage and high-stakes test performance for struggling students. *J. Comput. Assist. Learn.* **2019**, *35*, 569–581. [CrossRef]
36. Honnibal, M.; Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Appear* **2017**, *7*, 411–420.
37. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
38. Available online: https://github.com/MartinoMensio/spacy-universal-sentence-encoder-tfhub (accessed on 28 October 2022).
39. Available online: https://tfhub.dev/google/universal-sentence-encoder/4 (accessed on 28 October 2022).

40. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; Volume 8, pp. 216–225.
41. Warriner, A.B.; Kuperman, V.; Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behav. Res. Methods* **2013**, *45*, 1191–1207. [CrossRef] [PubMed]
42. Mohammad, S.M.; Turney, P.D. Crowdsourcing a word–emotion association lexicon. *Comput. Intell.* **2013**, *29*, 436–465. [CrossRef]
43. Mainz, N.; Shao, Z.; Brysbaert, M.; Meyer, A.S. Vocabulary knowledge predicts lexical processing: Evidence from a group of participants with diverse educational backgrounds. *Front. Psychol.* **2017**, *8*, 1164. [CrossRef] [PubMed]
44. Yap, M.J.; Balota, D.A.; Sibley, D.E.; Ratcliff, R. Individual differences in visual word recognition: Insights from the English Lexicon Project. *J. Exp. Psychol. Hum. Percept. Perform.* **2012**, *38*, 53. [CrossRef] [PubMed]
45. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
46. Gelbard, R.; Goldman, O.; Spiegler, I. Investigating diversity of clustering methods: An empirical comparison. *Data Knowl. Eng.* **2007**, *63*, 155–166. [CrossRef]
47. Benassi, M.; Garofalo, S.; Ambrosini, F.; Sant'Angelo, R.P.; Raggini, R.; De Paoli, G.; Ravani, C.; Giovagnoli, S.; Orsoni, M.; Piraccini, G. Using two-step cluster analysis and latent class cluster analysis to classify the cognitive heterogeneity of cross-diagnostic psychiatric inpatients. *Front. Psychol.* **2020**, *11*, 1085. [CrossRef]
48. Paxton, A.; Roche, J.M.; Ibarra, A.; Tanenhaus, M.K. Failure to (mis) communicate: Linguistic convergence, lexical choice, and communicative success in dyadic problem solving. In Proceedings of the Annual Meeting of the Cognitive Science Society, Quebec City, QC, Canada, 23–26 July 2014; Volume 36.
49. Tosi, A. Adjusting Linguistically to Others: The Role of Social Context in Lexical Choices and Spatial Language. Ph.D. Thesis, The University of Edinburgh, Edinburgh, UK, 2017.
50. Lapadat, J. Discourse devices used to establish community, increase coherence, and negotiate agreement in an online university course. *Int. J. E-Learn. Distance Educ. Rev. Int. E-Learn. Form. À Distance* **2007**, *21*, 59–92.
51. Lavasani, M.G.; Khandan, F. The effect of cooperative learning on mathematics anxiety and help seeking behavior. *Procedia-Soc. Behav. Sci.* **2011**, *15*, 271–276. [CrossRef]
52. Qayyum, A. Student help-seeking attitudes and behaviors in a digital era. *Int. J. Educ. Technol. High. Educ.* **2018**, *15*, 17. [CrossRef]
53. Dadure, P.; Pakray, P.; Bandyopadhyay, S. Mathematical Information Retrieval Trends and Techniques. In *Deep Natural Language Processing and AI Applications for Industry 5.0*; IGI Global: Hershey, PA, USA, 2021; pp. 74–92.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.