Sentiment Analysis on Conversations in Collaborative Active Learning as an Early Predictor of Performance

Nasrin Dehbozorgi  
Dept. of Computer Science  
University of North Carolina at Charlotte  
Charlotte, USA  
ndebozo@uncc.edu

Mary Lou Maher  
Dept. of Software and Information Systems  
University of North Carolina at Charlotte  
Charlotte, USA  
m.maher@uncc.edu

Mohsen Dorodchi  
Dept. of Computer Science  
University of North Carolina at Charlotte  
Charlotte, USA  
mohsen.dorodchi@uncc.edu

ABSTRACT:

This full research paper studies affective states in students’ verbal conversations in an introductory Computer Science class (CS1) as they work in teams and discuss course content. Research on the cognitive process suggests that social constructs are an essential part of the learning process [1]. This highlights the importance of teamwork in engineering education. Besides cognitive and social constructs, performance evaluation methods are key components of successful team experience. However, measuring students’ individual performance in low-stake teams is a challenge since the main goal of these teams is social construction of knowledge rather than final artifact production. On the other hand, in low-stake teams the small contribution of teamwork to students’ grade might cause students not to collaborate as expected. We study affective metrics of sentiment and subjectivity in collaborative conversations in low-stake teams to identify the correlation between students’ affective states and their performance in CS1 course. The novelty of this research is its focus on students’ verbal conversations in class and how to identify and operationalize affect as a metric that is related to individual performance. We record students’ conversation during low-stake teamwork in multiple sessions throughout the semester. By applying Natural Language Processing (NLP) algorithms, sentiment classes and subjectivity scores are extracted from their speech. The result of this study shows a positive correlation between students’ performance and their positive sentiment as well as the level of subjectivity in speech. The outcome of this research has the potential to serve as a performance predictor in earlier stages of the semester to provide timely feedback to students and enables instructors to make interventions that can lead to student success.

Keywords—Active learning, CS1, verbal conversation, Natural Language Processing (NLP), sentiment analysis, affective domain, student performance

I. INTRODUCTION

Research in how people learn reveals that in addition to the cognitive process, social constructs are a part of the learning process [1]. This highlights the importance of teamwork in educational settings such as collaborative active learning. Effective teamwork is vital since it creates knowledge, promotes innovation, enhances productivity and ensures success [2]. A key component of a good team experience is performance measurement. However, the inherently complex nature of teamwork in active learning calls for tools and methods to measure and predict students’ performance at both individual and collaborative levels [3, 2]. In active learning, there are two types of teams: low-stake and high-stake teams. Low-stake teams are typically practiced in introductory-level courses where the goal of teamwork is learning from peers and improving students’ soft skills [4]. Students’ performance in low-stake teams does not contribute much to their final grade. On the other hand, high-stake teams are mainly practiced in upper-level courses where students apply what they have learned into practice in order to make a final product. Evaluating students’ performance in high-stake teams is mainly based on evaluating the final product and assessing the individual’s contribution to the product. However, evaluating individuals’ performance in low-stake teams where no final product is produced is a challenge. To address this issue theories on team performance converge in identifying attitude components that influence team performance such as affective states and behavioral processes [5, 6, and 2]. These essential components of attitude are important to measure since they promote...
team effectiveness and are associated with team performance [2]. Research shows affective attributes have been measured using self-report by having students fill out surveys and by using a Likert scale expressing their feelings [2] or by analyzing offline textual dialogues while students communicate on forums asynchronously [2, 7, 8]. The drawback of surveys is the lack of commitment from students to fill them out in a timely manner, not taking it so seriously to provide precise answers, or even not being aware of their emotional states at the moment. More advanced tools allow researchers to retrieve emotional information signals by capturing facial expressions, gestures, and posture but they have certain drawbacks to be applied in educational settings such as distracting the learning process [9, 10]. Research suggests that speech is a good source to identify affective information, however because of challenges such as environmental noise level it is not often practiced in educational settings [11].

The goal of this research is to operationalize affect (sentiment and subjectivity) from students’ real-time speech in class and investigate if there are correlations between students’ affect in low-stake teams and their individual performance. The correlations can lead us to predict student performance based on the emotions that they express and identify at-risk students earlier in the semester to provide timely feedback to them. To collect the verbal data, we record students’ conversations in multiple active learning class sessions during the semester. By using NLP algorithms, we quantify affective states of sentiment and subjectivity. The results of the data analysis can serve as helpful feedback for both the instructor and students. Research shows that presenting detected emotions to students during their interactions makes them more conscious of their situation and can serve as a prompt to adjust their behavior in the learning process [12]. Beyond providing feedback to the student, the results of our research can guide instructors in applying interventions that improve students’ performance.

In section 2, we review the relevant literature and elaborate on different types of teams that are practiced in active learning. We describe team performance measurement techniques and attitudinal constructs as important metrics in cognition process. We demonstrate the existing gap in the literature and the need for developing measurement tools for low-stake teams in active learning. In section 3, we present our research methodology to operationalize affective metrics from verbal conversations to study the correlations of the affect constructs with students’ performance. Finally, we present the results of data analysis on a CS1 class as a case study.

II. BACKGROUND

Teamwork plays two important roles in active learning: one is peer learning or the social construction of knowledge which helps students to learn from each other and the second is about improving social and soft skills [13, 28]. Teamwork in active learning is being practiced either in the form of high-stake or low-stake teams. The low-stake format is typically applied in introductory-level courses where students interact with each other and improve their social skills as they learn from peers in a socially supported environment. This type of teamwork best suits less-challenging concepts where students get a chance to learn from peers and reduce the gaps between team members’ computing backgrounds. This low-stake teamwork model is also defined as ‘lightweight’ teams in which teamwork has less contribution to students’ final grades [4]. In low-stake teams normally there are no assigned roles and students contribute equally to solve given problems without pressure for grades. In such teams, effective evaluation helps in a fair assessment of individuals besides providing timely feedback to them.

A. Team performance measurement

Researchers have proposed different tools and insight on evaluating team performance over the past 30 years. A review of the literature of teamwork assessment shows there are mainly four ways to collect data on team performance: 1) self-report, 2) peer assessment, 3) observation and 4) objective outcomes. For optimal results, it is suggested to combine different ways of both qualitative and quantitative data collection [2]. In high-stake teams, the most common way to measure success in teams is to evaluate the quality of the artifact generated by those teams. In low-stake teams, there is no significant team-level final product to be evaluated as a team performance measure. In these types of teams, students do not have assigned roles, and given that teamwork outcome has a low contribution to final grades, there is a high chance that team members rely on peers and don’t attend team activities as expected. This scenario has negative consequences in active learning classes since students’ will not learn from peers and do not use class time for learning the course material. In such cases, the emphasis of the team evaluation should be at the individual level, in order to provide timely feedback to students. Here the question arises that what data needs to be collected and what factors should be used for evaluating individuals in low-stake teams to ensure they are having a positive teamwork experience.

There is a large body of recent research around the affective domain in teamwork and how they affect team performance [14]. Research shows the first step to evaluate team performance is identifying the characteristics that individuals in teams pose such as motivation, attitude and personality traits [15].
most common form of measuring attitude is having students fill out surveys by using a Likert scale to express their feelings [2]. However, such tools may not necessarily reveal reliable information about the affective state while having other drawbacks such as the need for direct student involvement and the overhead it may have for students. The challenge calls for methods to quantify the affective metrics at the individual level without individual self-report.

B. Affective domain

Multiple studies have reported that affective states impact the interpersonal relationships in the educational domain [16,17]. The shortage of soft skills among employees in the workplace is another evidence that there is a lack of focus on individual affects in the educational setting which needs to be integrated into curricula [18]. It is reported that students’ attitudes are observable through their behavior in class and how they engage in the class activities [19]. One aspect of attitude is emotion which can influence students’ behavior in collaborative environments and impact performance at both individual and team level [7]. Research shows emotional obstacles can hinder students’ learning process while feelings of joy, happiness, and satisfaction about the given subject positively influence students’ performance [20]. Recent studies claim that students who experience emotional and behavioral difficulties are not identified early and therefore may not receive appropriate feedback and intervention [21]. The affective state is most important in collaborative active learning where learning occurs during teamwork [7].

Capturing attitude information and emotional awareness can benefit both students and instructors [7]. From the students’ perspective, achieving emotional awareness involves the acquisition of skills to manage emotions and establish positive relationships with teammates and to learn how to handle challenging situations [17]. From the instructor’s perspective identifying students’ emotional state can lead to cognitive scaffolding [7]. Researchers propose diverse methods of affect recognition including heart rate, diagnostic tasks, self-reports, facial expressions and knowledge-based methods [11]. Some researchers recognize affective state by doing sentiment analysis on students’ journals and learning diaries or from the chat and discussions in the collaborative conversations in forums or other asynchronous textual data [22]. In these methods, they either identify the polarity of students’ emotions to see whether they are negative, positive or neutral or they identify the expressions most related to the six fundamental emotions of anger, trust, surprise, sadness, joy, fear, disgust, and anticipation [22]. Although sentiment analysis is a more promising way to identify the affective domain, it has not been widely applied in the educational setting compared to the social media domain or review corpora because of limitations in existing educational corpora [20,22]. The selection of a suitable method depends on the type of emotions to be recognized, the required resources to collect data and the context and setting in which the task is performed [11]. For example, in some contexts using self-reports may be more appropriate than using sensors, since sensors can cause interference with the given task [11]. Some cases may need real-time detection which requires computational resources for data analysis. Researchers believe speech is a very good source to identify affective information, however because of challenges such as environmental noise level it is not often practiced [11]. We have identified a way to capture individual student’s speech while in groups in the classroom and use this data to measure their affective states. We investigate correlations between the identified affective metrics in students’ speech and their individual performance.

III. RESEARCH METHOD

The research question we pursue in this study is: Is there a relationship between a student’s positive sentiment or level of subjectivity during in-class collaborative conversation and their individual performance in the course?

To answer this question, we operationalize sentiment and the level of subjectivity from students’ verbal conversations during the class activities and identify their correlation with students’ individual performance. The study protocol involves the processes of identifying modules to measure affect, identifying team characteristics and class design, recording students’ conversations in teams during class and applying voice filtering and transcription modules on audio data for text mining and analysis. “Fig. 1” shows the steps of study protocol.

![Fig. 1. Study protocol](image)

The two hypotheses that we have in this research are:

**H1:** There is a correlation between student’s positive sentiment in teamwork and their individual performance in the course.

**H2:** There is a correlation between the level of subjectivity in student’s speech in teamwork and their individual performance in the course.
The rest of this section describes our methodology to operationalize sentiment and subjectivity, our data collection approach, and analysis of data from the case study.

A. Methodology for operationalizing sentiments

For data collection, the students were recorded while they were working on the class activities with their assigned peers. During the recording process, some environmental, technical or human errors happened which made the data unavailable for analysis. For example, sometimes the TAs who were assigned to set up recorders misplaced the microphone cord. In other cases, students accidentally pressed the stop button on the recorder which led to the loss of data. On top of all these challenges we had high environmental noise level when all team members were talking at the same time while sitting close to each other in the classroom. In order to overcome these challenges we employed some protocols such as assigning the tasks related to recorders only to well-trained TAs and encouraging teams to sit in certain locations to minimize the noise level in recordings. One of the major steps we took was conducting an extensive research on the recording devices. The recorders we required for our study were expected to have bidirectional paired microphones, have built-in noise cancellation feature and lasting battery, be user friendly and cost effective. After trying three different types of recorders we identified the one that best matched our requirements.

In order to transcribe the conversations, first we filtered the audio data and reduced the noise level to improve the quality. Next, we transcribed the audio data by assigning a unique ID to each speaker based on voice recognition. Each team was recorded on one device and so their voices were stored into one audio file. Because the teams were sitting close to each other or sometimes they had questions from the teacher or TAs the transcriptions included the speech utterances from people other than the team members. In the next step, we removed the speech utterances from speakers other than the team members and stored speech utterances related to each ID (i.e. student) into a separate dataset. This resulted in 28 datasets each containing the transcription of speech of that individual in multiple sessions of the class. At this point, the textual datasets (i.e. ‘speech corpuses’) were ready for being imported to the text mining algorithm that we developed for this study.

The text-mining algorithm is shown in “Fig. 2”. The first step of the algorithm is segmentation where we segmented each corpora and separated the part of speech (i.e. vectors) based on speech initiation point such that each vector in the dataset represents the speech of the person until it is finished or interrupted by the other teammate. This means that the number of vectors in each dataset denotes the number of times a person initiated the talk (in both active and reactive mode).

Next, we applied a contraction filtering on the text so that we do not miss any meaningful words while tokenizing the vectors based on the regular expression which is explained in the next step. In the third step of the algorithm, we tokenized the vectors based on a customized regular expression to clean the text and eliminate the extra characters which did not impact the sentiment score. Next, we applied the standard English dictionary to remove the stop words. We also did a word frequency count in each dataset and defined a dictionary based on the most frequently common words that speakers used habitually which did not have any impact in the context of this study and removed those words from the text.

In the last step of the algorithm, we did sentiment analysis on the parsed text by applying TextBlob library, Natural Language Toolkit (NLTK) and Valence Aware Dictionary for sEntiment Reasoning (VADER) tool to measure positive sentiment, and subjectivity score of all vectors in each dataset.

TextBlob which is a rule based sentiment analysis tool has the essential component for the basics of natural-language processing such as calculating polarity and subjectivity [26]. The output subjectivity level is a float number within the range \([0.0, 1.0]\) where 0.0 is very objective and 1.0 is very subjective [27].

We apply TextBlob to measure the subjectivity of the text and use NLTK and VADER for measuring the sentiments (i.e. polarity and valence of the records). NLTK is a tool that allows tokenizing and word

Fig.2. Text-mining algorithm
frequency analysis on the corpus. The Vader sentiment analysis algorithm is applied due to its higher precision and accuracy in particular on short tokens at string-level compared to the other known sentiment analysis tools [23, 24, 25]. Most sentiment analysis tools have either polarity-based or valence-based approaches. Polarity determines if a part of the text is positive or negative, while valence-based approaches determine the intensity of each sentiment class. VADER is both polarity-based and valence-based which outputs sentiment scores into 4 classes of ‘Negative’, ‘Neutral’, ‘Positive’ and ‘Compound’ with values between -1 to 1. The compound value is the normalized value of the sum of valence scores of each word in the lexicon, adjusted according to the rules. Equation (1) shows how compound value is calculated based on the normalized sum of valence scores:

\[
\text{Compound value} = \frac{\text{sum}_\text{val}}{\sqrt{\left(\text{sum}_\text{val}\right)^2 + 15}} \tag{1}
\]

where \(\text{sum}_\text{val}\) is the sum of the sentiment arguments passed to the \(\text{score}_\text{valence}()\) function in Vadar algorithm.

Compound value is the most useful metric for a unidimensional measure of sentiment. Depending on the context, the threshold for neutral compound value normally can be anywhere between -0.05 and 0.05 [32]. In this study, to determine the threshold for neutral compound value, we did a k-means clustering on the compound values using the elbow method. The result indicated the optimum number of clusters would be three. The 3-means clustering of data showed most records fall into the class near to the zero value. Therefore, we consider zero as the threshold for classifying sentences into positive, neutral, or negative meaning any negative compound value is labeled as negative, positive values are labeled as positive and vectors with zero compound value are considered neutral. Finally, we normalized the sum of all compound values with values greater than 1 (i.e. positive compound values) and subjectivity score of all vectors in each dataset.

C. Results

To identify the correlation between positive sentiments and performance, we considered both the intensity and frequency of positive compound values for analysis. The intensity of positive compound value determines the level of positive sentiment, while frequency of positive compound value denotes the occurrence rate of positive sentiment in the corpus. “Fig. 3” plots the distribution of positive sentiment and subjectivity scores of all 28 participants. The horizontal axis shows the participant numbers and the vertical axis indicates the scores (i.e. level of positive sentiment and subjectivity).

As shown in “Fig. 3” the intensity of positive compound values is consistently lower than the

![Fig3. Positive sentiment and subjectivity scatter plot](image-url)
frequency of positive compound values, and the range of subjectivity score is larger than intensity and frequency of positive compound values.

The grade distribution of the participants in this study is shown in Table 1. The normal distribution of the grades shows equal number of high performing and low performing students, while most of the participants are in medium range of grades B and C.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>17.86 %</td>
</tr>
<tr>
<td>B</td>
<td>9</td>
<td>32.14 %</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
<td>32.14 %</td>
</tr>
<tr>
<td>DFW</td>
<td>5</td>
<td>17.86 %</td>
</tr>
</tbody>
</table>

In “Fig. 4”, “Fig. 5”, “Fig. 6”, and “Fig. 7” the kernel density plot of the intensity of compound values for the four grade categories (A, B, C and DFW) are presented. In these plots the horizontal axis shows the intensity level of compound value and the vertical axis denotes the density level. The plots show more density in positive compound values for students with higher grades, and as the students’ grades decrease the density of positive compound values decrease.

To answer the research question, we measure the correlation of the positive compound value (intensity and frequency) and the performance score as well as the correlation of the subjectivity score and the performance score by applying Spearman's rank correlation coefficient. For hypothesis testing, we applied a two-tailed p value statistical method and measure the p value for each correlation.

Spearman’s rank correlation coefficient is a nonparametric (distribution-free) rank statistic for measuring the strength of an association between two variables [31]. It assesses how well the relationship between two variables can be described using a monotonic function, without making any assumptions about the frequency distribution of the variables [31]. The coefficient value is signified by $r_s$, where $r_s$ can be anywhere between -1 and 1. The interpretation is that the closer is $r_s$ to +1 and -1, the stronger the monotonic relationship is between the two variables. The strength of the correlation can be described using the following guide for the absolute value of $r_s$ [30]: $r_s = 0.0-0.19$ “very weak”, $r_s = 0.20-0.39$ “weak”, $r_s = 0.40-0.59$ “moderate”, $r_s = 0.60-0.79$ “strong”, $r_s = 0.80-1.0$ “very strong”. The $r_s$ value is calculated using equation (2): ($n =$ number of cases)
Where \( d_i \) is the difference in ranks for variables.

By using Spearman’s correlation coefficient equation, we measured the correlation between positive compound values (intensity and frequency) as well as subjectivity with performance score. The coefficient values \( (r_s) \) are presented in Table 2.

**TABLE 2. COEFFICIENT VALUES OF POSITIVE SENTIMENT, SUBJECTIVITY AND PERFORMANCE**

<table>
<thead>
<tr>
<th></th>
<th>Coefficient value ( (r_s) )</th>
<th>Strength of Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity of positive compound value</td>
<td>0.61</td>
<td>Strong</td>
</tr>
<tr>
<td>Frequency of positive compound value</td>
<td>0.64</td>
<td>Strong</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>0.40</td>
<td>Moderate</td>
</tr>
</tbody>
</table>

The coefficient values \( (r_s) \) related to intensity of and frequency of positive sentiments indicate both have a strong positive correlation with performance. The calculated coefficient value of subjectivity shows a moderate positive correlation with performance.

“Fig. 8”, “Fig. 9” and “Fig. 10” visualize linear regression analysis of positive sentiment and subjectivity vs performance. In these plots the horizontal axis specifies the positive compound value (intensity and frequency) and subjectivity while the vertical axis shows performance score of each participant.

In “Fig. 8” and “Fig. 9”, we observe a homogeneous pattern between positive sentiment (intensity and frequency) and performance (i.e. higher performance scores have higher positive sentiments). On the other hand, in “Fig. 10”, the regression plot of the subjectivity and performance does not show any consistency between these two metrics.

For hypothesis testing we applied 2-tailed \( p \) value statistical method. The \( p \) values of positive sentiment and subjectivity scores and performance are presented in Table 3.

**TABLE 3. \( p \)-VALUES OF POSITIVE SENTIMENT, SUBJECTIVITY AND PERFORMANCE**

<table>
<thead>
<tr>
<th></th>
<th>( 2 )-tailed ( p ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intensity of positive compound value</td>
<td>0.001</td>
</tr>
<tr>
<td>Frequency of positive compound value</td>
<td>0.002</td>
</tr>
<tr>
<td>Subjectivity</td>
<td>0.03</td>
</tr>
</tbody>
</table>

In testing H1 the null hypothesis states:

H1\(_0\): There is no correlation between student’s positive sentiment in teamwork and that student’s individual performance in the course.

The calculated \( p \) value for intensity of positive compound value is 0.001 and the \( p \) value for frequency of positive compound value is 0.002. Since the result of the \( p \) values for both the intensity and frequency of positive compound values are statistically significant, therefore the null hypothesis is rejected which confirms that there is a correlation between students’ positive sentiment and their performance.

For validating H2 the null hypothesis states:

H2\(_0\): There is no correlation between the level of subjectivity in student’s speech in teamwork and that student’s individual performance. Considering the subjectivity and the performance metrics, the
calculated $p$ value is 0.03. Since the $p$ value is less than the confidence level of 0.05 therefore the null hypothesis is rejected which indicates there is a correlation between students’ level of subjectivity in speech and their individual performance.

In another observation from data analysis, we found that there is a negative yet strong correlation between the frequency of students’ neutral sentiment and their performance. The calculated coefficient value ($r_s$) for the neutral sentiment is -0.61, and the $p$ value is 0.001. This means that students with higher performance score had fewer number of records with neutral sentiment scores. “Fig. 11” shows the regression plot of the frequency of neutral sentiment values vs performances.

However, we did not find any correlation between students’ negative sentiments and their individual performance.

IV. CONCLUSION AND FUTURE WORK

The purpose of this study is to analyze the correlation between students’ positive sentiments as they talk during teamwork and their individual performance in the course. The identified correlations help instructors to make novel interventions in order to encourage students’ engagement in teams and also identifying at risk students to provide timely feedback and support for them.

The novelty of this study is: 1) using verbal conversations as a medium to measure affective states, and 2) focus on low-stake teams (i.e. lightweight teams). There has been much focus on methods to evaluate students’ performance in capstone teams. However, due to low grade contribution and lack of student engagement in low-stake teams, evaluating students in such settings have been a challenge. Supported by the research, we argue that capturing students’ affective states in low-stake teams can help us to evaluate individuals’ performance.

By developing a text-mining algorithm as shown in “Fig. 2”, we operationalized the positive sentiment and subjectivity level for each student as they spoke in teams. Based on the result of this work, we conclude that there is a strong correlation between performance and positive sentiment as well as the level of subjectivity. This method helps in evaluating students’ level of involvement in low-stake teams and by doing frequent formative analysis on their speech in teamwork, we can identify lower performers and provide more learning opportunities to them.

In future work we will conduct multi-class sentiment analysis on the collected data by studying diverse classes of sentiment such as “joy, anger, anxiety, etc.” rather than just the polarity of the sentiments. We will analyze different sentiment classes to identify their correlation with students’ performance.

In the next step we will conduct aspect-based sentiment analysis on the corpus to identify in which areas or themes students show more positive or negative sentiments. The result of this analysis would help instructors to get a more precise feedback from the context of speech in teams and enables them to implement cognitive interventions.

References:


