SENTIMENT ANALYSIS ON VERBAL DATA FROM TEAM DISCUSSIONS AS AN INDICATOR OF INDIVIDUAL PERFORMANCE

by

Nasrin Dehbozorgi

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computer Science

Charlotte

2020

Approved by:

____
Dr. Mary Lou Maher

____
Dr. Mohsen Dorodchi

____
Dr. John Gero

____
Dr. Samira Shaikh

____
Dr. Bojan Cukic

____
Dr. Jessica Schlueter

ABSTRACT

NASRIN DEHBOZORGI. Sentiment Analysis on Verbal Data from Team Discussions As an Indicator of Individual Performance. (Under the direction of DR. MARY LOU MAHER)

The inherently complex nature of evaluating teamwork calls for methods to measure and predict students' performance and provide timely feedback. Analyzing students' individual performance, particularly in low-stake teams, is a challenge since the main goal in such team settings is knowledge acquisition and social skill development rather than final artifact production. Students' attitude plays an important role in their cognition process and is commonly measured by self-report tools or asynchronous communications in textual discussion platforms. These methods have certain drawbacks such as distracting the learning process, demanding time and commitment from students or lack of emotional awareness which reduces the reliability of such tools. Research suggests speech is the best way to measure attitude accurately since it captures behavior directly rather than self-report. This study focuses on operationalizing attitude constructs of affect, self-efficacy, and personality and analyzing their correlation with students' performance in order to identify which attitude constructs serve as performance predictor metrics. Personality and self-efficacy are measured using standard self-report tools, and affective states are captured from students' conversations as they work in low-stake teams in a CS1 active learning class and discuss course content with peers. The novelty of this research is its focus on students' speech in class and how to operationalize affective states from verbal conversations as a metric that is related to individual performance. We record students' conversations during teamwork in multiple sessions throughout the semester. By applying Natural Language Processing (NLP) algorithms we conduct sentiment analysis to detect valence, polarity and multiple classes of affective states from the conversations. The result of data analysis shows that students with higher levels of positive sentiment during the semester had higher performance scores. The result of personality and self-efficacy self-report tools, however, does not indicate a statistically significant correlation with performance. This outcome supports the research argument that self-report tools are not reliable in capturing attitude constructs. The result of data analysis in this study helps in identifying the attitudinal components that are correlated with performance to be applied in developing predictive models.

ACKNOWLEDGEMENTS

I express my sincere gratitude to my advisor Dr. Mary Lou Maher, for her tremendous support throughout my Ph.D. program. Her immense knowledge, wisdom, patience, and motivation played a crucial role in my successful completion of the Ph.D. Under her tutelage, I gained research skills to accurately identify the research issue, apply methodologies, and publish the results at various conferences. Beyond being the best advisor, she has also been a role model embodying the principles I look up to.

I am very grateful to my co-advisor Dr. Mohsen Dorodchi, to mentor and guide me before and during my Ph.D. studies. I had the privilege of researching with him on different projects, which led to professional publications. His timely support significantly contributed to my success in the Ph.D. program.

Besides my advisors, I would like to offer my sincere thanks to the committee members, Dr. John Gero, Dr. Samira Shaikh, and Dr. Bojan Cukic for taking the time to provide well-considered advice and professional feedback during the process. Understanding their different perspectives helped me confidently maintain the research's cohesion while setting up future opportunities for expanding this work. My Special thanks goes to Dr. Jessica Schlueter for serving on my dissertation committee and her optimism and encouragement on this research path.

An essential contribution to my PhD work's success was because of the collaboration with the CCI faculty who provided resources for the data collection. Mainly I would like to thank Dr. Siddharth Krishnan, Dr. Nadia Najjar, Dr. Bojan Cukic, Dr. Mary Lou Maher, and Dr. Mohsen Dorodchi for granting the opportunity to collect data from their class participants. Furthermore, I would like to thank faculty members of the Active Learning Academy, Center of Educational Innovations, my

HCI lab peers and CCI students at UNC Charlotte for the stimulating discussions, insightful feedback, and participating in the user studies.

My sincere appreciation goes to Dr. Srinivas Raghavan, who dedicated his precious time to serve as my external mentor. His wisdom, advice, and timely motivation encouraged me to persist during the process's most challenging times.

Most importantly, I convey my deepest gratitude and love to my family for their critical role in this dissertation. My father, mother, and sisters encouraged me to pursue my goals, and selflessly supported me in all my endeavors. Their unending emotional and financial support made me succeed in this journey.

Finally, I would like to acknowledge that this work is supported by the National Science Foundation Award 1519160: IUSE/PFE: RED: The Connected Learner: Design Patterns for Transforming Computing and Informatics Education.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Research in how people learn reveals that in addition to the cognitive process, social constructs are an essential part of the learning process [1]. This highlights the importance of collaboration in educational settings. Considering the inherently complex nature of teamwork in active learning, there is a need for tools and methods to measure and predict students' performance at both individual and collaborative levels [2, 3].

Theories on team performance converge in identifying attitude components that influence team performance such as affective states, behavioral processes, and cognition [3,4, 5, 6, 7, 8 and 9]. These essential components of attitude are important to measure since they promote team effectiveness and are associated with team performance [3, 10, and 11]. Self-efficacy is another attitude construct that has drawn the attention of many researchers in the field. Self-efficacy is defined as how individuals judge and perceive their ability to perform a specific task, which impacts the performance [12, 13]. There is also robust research on the impact of personality type on individuals' and team's performance [14]. In general, it is claimed that homogeneous team composition leads to more positive team experience which improves team performance [14].

1.1 Motivation

Teamwork in active learning can be categorized as low-stake and high-stake teams. Low-stake teams are mostly practiced in introductory-level courses where the goals of teamwork are learning from peers and improving students' soft skills [15]. 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 to make a final product. Evaluating students' performance in high-stake teams is basically 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. Existing research on evaluating teams and identifying the impact of attitude on team performance mainly focuses on the high-stake teams where team members contribute to producing a final product.

This research takes a holistic approach to measure the impact of attitude on students' individual performance in low-stake teams. We will analyze students' affective states, self-efficacy, and personality types, while they are assigned to work in small-sized teams and identify how these factors impact students' performance. In the following section, my thesis statement and the hypotheses are discussed.

1.2 Thesis Statement

This thesis claims that the following three constructs can characterize individual student's success in the course: 1) student affect in conversations during team activities, 2) self-efficacy and 3) combination of personality traits in a team.

We focus on three constructs of affect, efficacy, and personality to take a deeper look at students' behavior in teamwork by analyzing their verbal conversations to identify the correlations with their performance in the course. For this purpose, we record verbal conversations in teams and utilize standard self-report tools to operationalize attitude components.

1.3 Research Hypotheses

We identify correlations between these constructs and students' performance in low-stake teams. Identifying these relationships can help predict student performance and identify at-risk individuals. We formulate 4 Null hypotheses:

H1$_0$. There is no correlation between students' positive sentiment (frequency and intensity of compound values) in low-stake teams and their individual performance.

H2$_0$. There is no correlation between students' initial self-efficacy and individual performance.

H3$_0$. There is no correlation between students' initial self-efficacy and positive sentiments (frequency and intensity of compound values) in low-stake teams.

H4$_0$. There is no correlation between the combination of students' personality types in low-stake teams and their individual performance.

The research question in this study is how to operationalize these constructs as metrics to be measured. Research shows traditionally attitude has been measured by having students fill out surveys and by using a Likert scale expressing their feelings [3]. 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, posture, and periods of silence [16, 17, 18, and 19].

In this research, we record students' collaborative conversations in multiple active learning class sessions during the semester. Students' affective states are extracted from transcribed data by sentiment analysis tools. For measuring self-efficacy, we employ a standard efficacy tool at the beginning and at the end of the semester. Finally, we identify students' personality traits by a standard personality instrument that is self-reported by students at the beginning of the semester. The emerging patterns from data analysis show if there is a correlation between students' performance with attitude, personality trait, and self-efficacy.

The emergent result from data analysis can serve as helpful feedback for both the instructor and students. Research shows that presenting detected emotions during students' interaction to them makes them more conscious of their situation and serves as a prompt to change their behavior in the learning process [20]. The impact of this research is to enable instructors to apply interventions to influence students' attitudes and behavior in teams in earlier stages of the semester to improve their performance.

In chapter 2, we discuss different aspects of collaborative active learning and how it leads to students' engagement and motivation in the class. We elaborate on dimensions of teamwork and introduce different types of teams that are practiced in educational settings. Finally, we talk about diverse team performance measurement techniques and attitudinal constructs that can be considered when evaluating the team's performance. We demonstrate the existing gap in research and the need for developing measurement tools for low-stake teams in active learning.

Chapter 3 talks about the attitude constructs of affect, self-efficacy, and personality traits that have been identified in a robust body of research as important factors in the cognition process. In Chapter 4, we present our data collection protocol and methodology to operationalize the attitude from conversations in low-stake active learning teams and the correlation of attitude to individual performance. The result of the data analysis from this study is presented in Chapter 5 and in Chapter 6 we conclude by summarizing the study, discussing the takeaways, limitations, and the plan for future work.

CHAPTER 2: COLLABORATIVE ACTIVE LEARNING

Active learning affords students the opportunity to understand and apply course topics during class time in the presence of the instructor and teaching assistants [21]. It requires students to engage in meaningful learning activities and think about what they are doing, either in the form of teams or individually [22]. Active learning has two primary benefits: 1) utilizing class time to address students' misconceptions and help in a deeper understanding of the knowledge during class activities, 2) create a more engaging learning experience for students and improve their social skills [23]. Student engagement and collaboration are features of active learning that are often contrasted by traditional lecture settings where students passively receive information [23]. It is a challenge for students to maintain their attention for the entire class period and they tend to lose their focus after the halfway point of a long lecture [24]. This has been a motivation for integrating more activities into lectures so that students remain engaged in the class. The class activities are done either individually or in teams to solve a given problem. Incorporating activities during scheduled class time is a unique opportunity where students can work together without schedule conflicts under the supervision of an instructor. This indicates that active learning can be considered as a continuum along which varying amounts of activity can be included in a class period. Although there is some variation in terms of how active learning is defined and discussed, there are some generally accepted definitions that help to distinguish it from non-active learning [23]. There are many different types of pedagogy that could be classified as active learning, such as team-based learning (TBL) [25], cooperative learning [26, 27], collaborative learning [23], problem-based learning [23] or studio-based learning [28]. Although there are instances where students may work on activities alone, most forms of active learning have an emphasis on collaboration, learning from peers, and social construction of knowledge. Frequent application of teams in active learning highlights the importance of having structured protocols for team formation, adjusting the team size, and evaluating team performance.

The diversity of collaborative active learning methods makes it a challenge to formalize, adopt, and share goal-oriented practices in a structured format. Pedagogical design patterns are one way to formalize pedagogical practices [21]. Pedagogical Design patterns represent known problems and solutions in a standardized way to enable the sharing of emerging best practices. They allow designers to look up a problem that they are currently facing and use practiced solutions which are often rooted in learning theories or empirical rationale. Mapping the experience and practice to the theories of learning and motivation is not easy, especially for new instructors. Design patterns provide a framework to formalize this connection between problems and existing solutions based on theories or experience [21, 29].

Our review on 235 existing pedagogical design patterns in the literature reveals that only 5% of them addressed the problems related to active learning features like teamwork and collaborative class activities [21, 29]. In order to fill this gap and capture the best practices of active learning and disseminating them, we have developed a set of 18 collaborative active learning design patterns [30]. The patterns fall into four categories of 1) preparation design patterns, 2) in-class activity design patterns, 3) teamwork and collaboration design patterns, and 4) reflection and feedback design patterns. These patterns are formalized into an object-based design pattern model. This model and its attributes will be discussed in detail in section 2.1.

In the following section, teams in education, different dimensions of teamwork in the educational setting, as well as the teamwork attributes of our object-based design pattern model, are discussed.

2.1. Teams in Education

Forming teams that work well together are a hallmark of effective team-based learning. In Michaelsen's seminal work, he posits that cultivating cohesion within the team is essential to the success of those teams [31]. But depending on the context and goal of teamwork, success in teams can mean several things, including good performance as well as the social construction of knowledge and positive teamwork experiences.

From the performance perspective, teams are formed to maximize the performance of students. With this goal, there are a number of guidelines for team size, formation, and roles. Choosing an appropriate size for teams is task-dependent, and recommendations for optimal sizes vary widely in the literature [32]. Dyads are a popular choice for CS courses in the form of pair programming. There are two other common recommendations for team size: 3-5 and 5-7. 3-5 is typically recommended for activities that require less structure [32], and for more structured interaction, larger teams can be considered. LeJeune recommends 5-7 to ensure that the team has enough breadth of skills to complete the task while minimizing social loafing and promoting positive interdependence [33]. We refer to these two functional sizes as small and medium respectively. Large can be considered a catchall for other sizes; however, it is generally associated with class-wide activities, such as discussions. When initially forming teams, students can be grouped together randomly, by each individual's preference, or by the instructor. Randomly formed teams are often preferred because they reduce coalitions [31] and homophily [34]. A compromise between randomly selected and instructor selected teams are teams that are chosen algorithmically. For example, CATME attempts to integrate instructor specified criteria while avoiding scheduling conflicts within teams [35]. Roles are one way to ensure positive interdependence in teams [36]. By giving students roles it is possible to ensure that they work collaboratively and rely on each other. The use of roles has shown to improve cohesion in programming teams [37]. We identify two types of roles: task-specific and team-specific. Examples of task-specific roles are driver and navigator for pair programming or programmer, tester, and documenter in the traditional programming team. Team specific roles are designed to keep the team on track. Examples of team-specific roles are the timekeeper, encourager, and devil's advocate [32]. Roles can be assigned by preference, personality tests, and randomly. Cruz et al. provide a review of personality tests in software engineering education and name Myers Briggs, Kersey Temperament Sorter, and Neo Five-Factor Model as the three most common tests for forming teams [38].

To achieve social and collaborative benefits from the team experience, different factors should be considered when forming teams. For example, there is a large and significant negative correlation between teams in which some members have pre-existing friendships and performance on a group project [39]. This is one of the reasons that self-selected teams should be avoided; however, these teams can also provide opportunities for students to develop their friendships so that they are more connected to other students in their major. Barker et al. [40] previously identified student-to-student interaction as one of three variables that were significantly correlated with the intention to major in computer science. Similarly, persistence in an academic program has also been correlated with a student's sense of social support [41]. Although, some of these factors may not lead to high performing teams they may not necessarily lead to unsuccessful team-based learning experiences.

Teams may have multiple goals and attributes, which can be emphasized depending on the context and the goal of pedagogical practices. As discussed in the previous section design patterns can help in formalizing and structuring goal-oriented pedagogical practices. For this purpose, we have developed an object-based design pattern model with two-level attributes which captures the teamwork features in active learning [21, 29]. The team attributes of this model are formation, size, composition, duration, individual grade, contribution to the final grade, activity progression, and roles. Each of these attributes has a set of defined values based on the research and empirical evidence. Figure 1. illustrates this model, its components, attributes, and related values [21, 29].

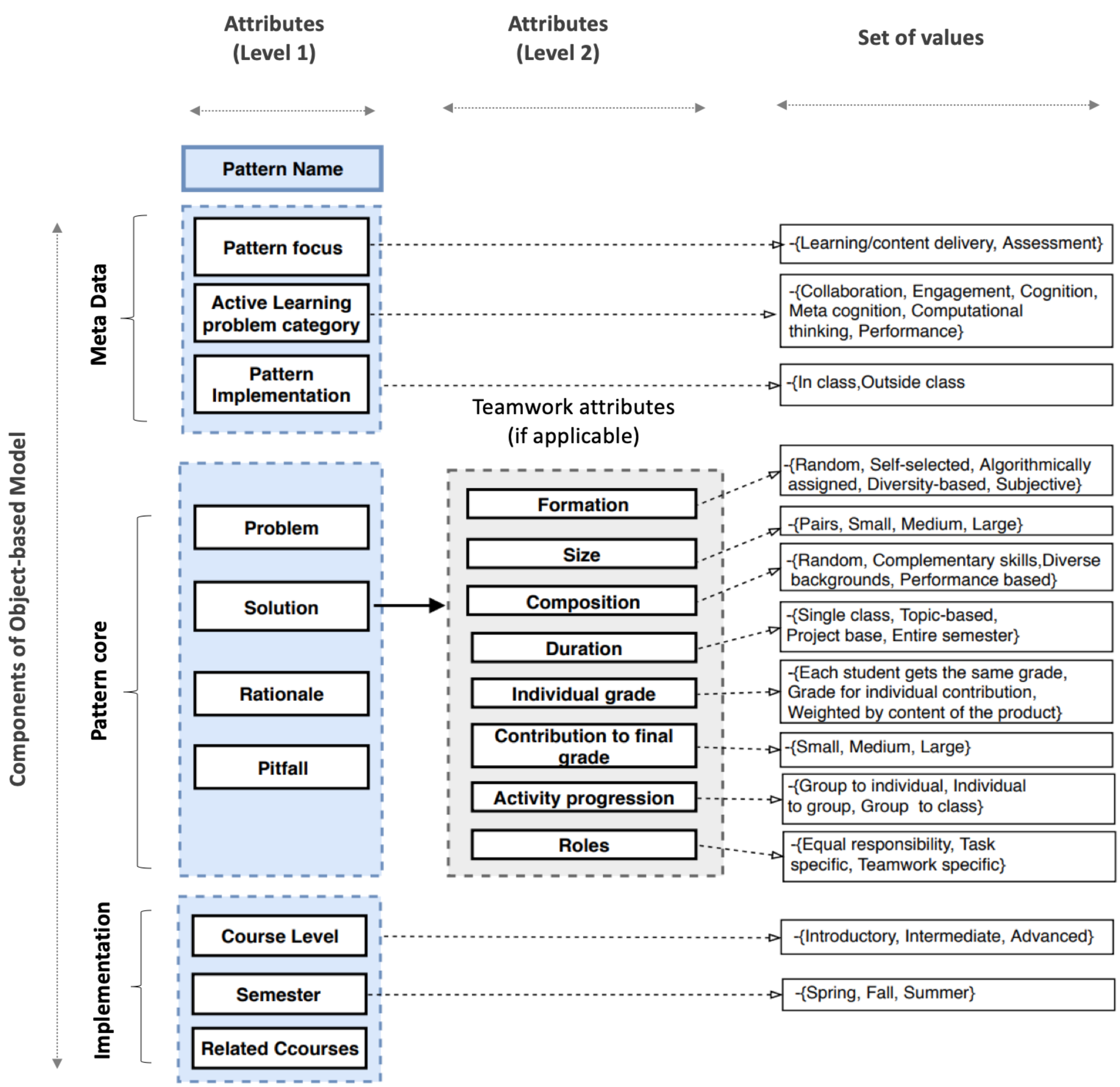


Figure 1. Object-based design pattern model with teamwork attributes [21]

In summary, teams have different dimensions of size, role, and diverse formation algorithms which vary depending on the context of teamwork. The goals for teams in education could be either collaboration that leads to a high-quality report or product, the social construction of knowledge or positive collaborative experience. Another important aspect is the goal of teamwork and how teamwork impacts students' evaluation in the course. These differences in team goal and outcome lead to the next session where we introduce two main types of teams that are being practiced in educational settings.

2.2. Team types

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 one is improving social and soft skills. Teamwork in active learning has been introduced by different terms such as cooperative learning or collaborative learning, depending on the emphasis of teamwork.

Collaborative learning refers to any pedagogical technique which requires students to work together in small groups to accomplish a common goal [23]. Therefore, collaborative learning can include all group-based instructional methods, including cooperative learning [26, 42]. However, some researchers believe there are distinctions between collaborative and cooperative learning and that they have different philosophical roots [43, 44].

Collaborative learning has more emphasis on students' interaction rather than merely performance and assessment. On the other hand, cooperative learning is defined as structured teamwork where students pursue common goals while being assessed individually [26]. This type of teamwork has five specific aspects, which are individual accountability, mutual interdependence, face to face interaction, the appropriate practice of interpersonal skills, and regular self-assessment of team functioning. While different cooperative learning models exist [45], the core element held in common across them is a focus on cooperative incentives rather than competition. Therefore, roles, applying individual skills in team product, and assessment are more emphasized on cooperative learning.

Based on the given definitions, we believe collaborative learning is a form of teamwork mainly applied in introductory-level courses in which students need to interact with each other and improve their social skills while learning from peers in a socially supported environment. This social aspect is very important since it improves students' communication skills and makes them prepared for upper-level classes and future professions. It also makes learning a fun experience while motivating students to be actively involved due to social pressure in teams. This type of teamwork best suits less challenging and sophisticated concepts where students get a chance to learn from peers and fill up the gaps between team members' backgrounds. This low-stake teamwork model is also defined as 'lightweight' teams in which teamwork has less contribution to students' final grades [15]. In lightweight teams normally there are no assigned rules and students contribute equally to solve given problems without pressure for grades.

On the other hand, cooperative learning is mainly applied in higher-level classes and capstone courses where a project is defined as a common goal and all team members have assigned roles to apply their knowledge and develop a final product. In this form of teamwork, less emphasis is on learning but more is on applying what has been learned to achieve a common goal. Cooperative learning is more structured and usually, team members have assigned tasks and roles. It is more suitable for advanced and challenging topics where tasks are distributed among the team members. We name this type of teamwork 'high-stake' teams since teamwork has a high contribution to students' final grades.

Whether students work in low-stake or high-stake teams, measuring their performance is critical. In low-stake teams, the performance evaluation can help in providing timely feedback to improve students' learning, and in the high-stake team, it helps in a fair assessment of individuals besides providing feedback to them. In the following section, we introduce existing methods in the literature to measure team performance.

2.3. Team performance measurement

Effective teamwork is vital, since it creates knowledge, promotes innovation, enhances productivity, and ensures success [3]. A key component to good team experience is performance measurement. However, the dynamic nature of teams makes it a complex and challenging task. Researchers have proposed different tools and insight on evaluating team performance over the past 30 years. In measuring teamwork, we need to collect data from different resources [3]. As mentioned by [46], getting the necessary information from all perspectives of teamwork requires a group of team observers. The review on the literature of teamwork assessment shows there are mainly four ways to collect data on team performance which are: 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 [3].

Evaluating and grading teams is difficult because both individuals and the team as a whole need to be considered. Grading schemes for assessing individual students either have students share one grade that was assigned to each team member, have their individual contributions evaluated, take quizzes to assess individual competencies, or have a cross-validating approach that combines more than one of these schemes [47]. As an example, Michaelsen suggested using an individual readiness assurance test as a way to ensure each student was developing individual competency and cross-validate the individual's contribution to the team [31].

In general, it is not easy for instructors to evaluate team dynamics inside and outside the classroom to verify fair and successful team experience for every individual. For this reason, peer and self-evaluation are common methods of assessing individuals' contributions in teamwork. Surveys are employed for this kind of assessment and these surveys can include Likert scales, partner ranking, descriptive word matching, short answers about peers, and journaling about their effort and experiences [48]. Finally, the weight of the grade based on each component is either provided by the instructor in the form of a standard rubric, or is negotiated between the instructor and the students [31].

The most common way to measure success of teams is to evaluate the quality of the artifact generated by the teams, such as in a capstone course. Examples of these artifacts in CS education are group presentations, written code, documentation, and project demonstrations. These artifacts are one important aspect of team-based learning because in the industry these artifacts are highly valued. In teams that collaborate on the final artifact(s), the artifact represents a significant portion of each team member's grade. Therefore, teams should be chosen as fairly as possible. This can be challenging because IRAT (individual readiness assessment tests) and other individual performance metrics such as GPA only represent one aspect of team performance [31]. Positive interdependence and individual accountability, which are not accounted for in GPA, are also essential components of collaborative learning [36]. Finally, because teams can only deliver one final artifact, they must be able to come to a consensus. For this reason, conflict resolution styles [49], personality traits [50], and leadership styles [51] are sometimes considered when forming teams for CS education.

Although a robust body of the literature talks about different methods for measuring team performance, it is not a perfect science and some gaps still exist [3]. As discussed, team performance evaluation can be complex and even contradictory depending on the intended learning outcomes and the associated problems that team-based learning is attempting to address.

In evaluating team performance, the attributes that may contribute are team types, context, the goal of teamwork, and the type of team outcome. The performance measurement goal may also differ in different team settings. Team performance measurement can be done to assign a grade to teams, to provide timely feedback or to modify team formation.

In the last section of this chapter, we discuss more on the existing gap in team performance measurement methods and the need for more research on evaluating individuals in low-stake teams.

2.4. Evaluating Performance of Individuals in Low-Stake Teams

Teams have outcomes at both teams and individual levels, therefore measuring teamwork at the individual level is as important. The team-level outcome is the final product which is the result of the effort of all team members, while at the individual-level outcome can be the team member's attitude in the teamwork process which is related to team performance [3].

Traditionally team performance evaluation is based on the outcome rather than the process. Focusing on the process can identify issues that serve as a guide for feedback, while the outcome provides information about the bottom-line performance [3].

Lack of quantitative and objective measures of teamwork at the individual level has slowed down the knowledge of evaluating teams' performance and assembling effective teams [52]. Most metrics to evaluate individuals in teams rely on experts who observe and rate teams by using rubrics or based on quantitatively vague dimensions like leadership and team structure [52].

As discussed in the earlier section, 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 activity as expected. This scenario has severe consequences in an active learning setting since students' will not either 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 and process of teamwork, in order to provide timely feedback to students.

Here the question is what data needs to be collected and what factors should be used for evaluating individuals in low-stake teams. There is a large body of recent research around attitudinal components in teamwork and how they determine team effectiveness [53]. 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 [54]. The most common form of measuring attitude is having students fill out surveys and by using a Likert scale expressing their feelings [3]. However, the drawback of such surveys and self-report is 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 status at the moment. More advanced tools such as EEG technology have been applied by researchers to retrieve emotional information through brain signals or by observing facial expressions, gestures, posture, and periods of silence to capture behavioral information [55, 56, 57, 52]. Although these tools provide better results on capturing attitude compared to individual surveys and self-reports, still they are not error-free, and deploying these tools and using experts as observers requires many resources. This can make it less practical in real life particularly in educational settings and classrooms, because of the possible distractions they can cause to the learning process.

The existing gap in research calls for quantifying other dimensions beyond cognition in evaluating teamwork at the individual level. Formative evaluation of individuals in different team settings without direct involvement and overhead for students can determine teamwork effectiveness. It is best to apply evaluation methods that require minimum student involvement since involving students to answer surveys or even having rubrics and observers in the evaluation process causes either distraction or takes class time.

In the next chapter, we explore attitudinal dimensions applied in the educational setting, such as affect, personality traits, and self-efficacy, and discuss how they contribute to team effectiveness.

CHAPTER 3: ATTITUDE DIMENSIONS

Attitude is a complex subject that has multiple properties and dimensions [58]. Generally, it is defined as a tendency to have certain beliefs and feelings in diverse situations [58]. It involves behavior for doing what one wants to achieve and therefore defines personality through the quality of performance [58].

Attitudinal attributes play an important role in students' learning in educational settings [59]. It is reported that students' attitude is observable through their behavior in class and how they engage in class activity [58]. Many studies have been conducted in diverse fields to identify how students' attitude towards a specific subject can impact their performance in the given context [60, 61].

One aspect of attitude is the emotion that can influence students' behavior in collaborative environments [62]. Research shows emotional obstacles can hinder students' learning process while feelings of joy, happiness, and satisfaction about the given subject positively influence students' learning [63]. Recent studies claim that students who experience emotional and behavioral difficulties are not identified early and therefore may not receive appropriate feedback and intervention [64].

Capturing attitude information and emotional awareness is important for both students and instructors [62]. From the students' perspective, emotional awareness besides the cognitive process involves the acquisition of skills to manage emotions and establish positive relationships with teammates and learn how to handle challenging situations [65]. This is most important in collaborative active learning where the learning process occurs during teamwork [62]. Affective information gets even more value for instructors since by identifying students' emotional state they can provide effective feedback and find ways to deal with students' negative emotions.

Affective information can lead to both cognitive and affective scaffolding, which will improve active learning and collaborative knowledge construction [62].

Self-efficacy and personality are other components of the attitude domain [58, 12]. Self-efficacy has been mentioned as a component of attitude which has an important role in students' performance in CS education [12, 13]. Research shows personality is also one of the most powerful factors which influence academic performance [66].

As instructors are looking for ways to understand and cope with the students' challenges during the learning process, theoretically supported measurement tools are required to quantify and analyze students' emotions and behavior in the educational setting [63, 67]. To date, most of the existing tools to identify attitude components use students' self-report instruments such as the Achievement Emotions Questionnaire (AEQ) tool to measure emotions or analyze students' textual data and collaborative writings in the chats or asynchronous forums. [62, 67]. Others use self-reports to measure attitude on a more abstract level such as attitude towards computing, technology, etc. [68].

In the rest of this chapter, we introduce the attitudinal components that play an important role in students' learning processes in academic settings. These components are affect, self-efficacy, and personality traits. Operationalizing and measuring these components can help in a better formative evaluation of individuals in low-stake teams. In the next sections, we discuss affect (emotion) and its role in the educational setting.

3.1. Affect (emotion)

Psychologists believe a person can be known based on three domains called the ABCs of attitude which are: Affect, Behavior, and Cognition [69]. Cognition can be measured by learning outcome and behavior can be observed, but affect is not directly observable and more challenging to measure [70]. The affective domain shows how much a person values the learning process, his willingness to contribute to learning new things, his ability to make a decision, and how generally one behaves in different situations [70, 65]. Multiple studies have reported that affective states also impact the teamwork process and interpersonal relationships in the educational domain [65].

Although there are difficulties in measuring the affective domain, however, it is an important factor to evaluate performance in the educational system [70]. The shortage of soft skills among employees in the workplace is another evidence that there is a lack of focus on individual affective states in the educational setting which needs to be integrated into curricula [70].

Researchers propose diverse methods of affect recognition and measurement. The methods vary from psychological measures such as heart rate, diagnostic tasks, self-reports, facial expressions, and knowledge-based methods to derive affective state in a given context like time of day and length of the task and individual histories like past failures and success [71]. One of the standard tools which have been applied for student self-report in the educational setting is the AEQ tool which measures a range of achievement emotions such as enjoyment, hope, pride, relief, anger, anxiety, shame, hopelessness, and boredom [67] ([72] provides sample self-report questionnaire). Other researchers consider personality as a metric to identify the affective state [71].

Fewer 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 [73]. 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 [73]. Although sentiment analysis is a more promising way to identify the affective domain, however, 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 corpus [63,73].

A review of the effective attributes that have been applied in the educational setting is listed in [71] which are: anger, fear, disgust, joy, sorrow, surprise, frustration, speech tone, head position, facial expression, pitch intensity, body movement, nervousness, and deception. Others suggest different affective attributes be considered in the educational setting such as grit, mindset, identity, and belonging [74].

Although more advanced tools are being introduced for affective assessment, reliable assessment requires concurrent use of multiple methods and not relying only on one approach [3, 71]. The selection of a suitable method depends on the type of emotions needed to be recognized, the required resources to collect data, and the context and setting in which the task is performed [71]. For example, in some contexts using self-reports may be more appropriate than using sensors, since sensors can cause interference with the given task [71]. 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 has not been much practiced [71]. Most of the existing research addresses certain aspects of affect sensing and recognition in the educational domain, however much research is required to measure students' affect to develop interventions [70, 71].

In summary, the most practiced method for measuring affect in existing research is using students' self-report data and surveys [3, 67]. These tools may not necessarily reveal reliable information about the affective state of the students. More reliable methods such as recognition of affect in speech can help in effective affect measurement. In the next section, we discuss the role of self-efficacy, another attitude component, and its impact on students' learning in the educational setting.

3.2 Self-efficacy

Studying self-efficacy in educational settings is important since it plays an important role in the learning process. By understanding what factors influence self-efficacy instructors can apply required interventions [12].

Self-efficacy is an attitudinal construct that is defined as the perception that one can successfully accomplish a given task [12]. It is inherited in self-regulation and reflects the motivation to apply learned skills [12, 13]. Self-efficacy is integral to social cognitive theory and provides a framework for analyzing human thought, motivation, and action [75]. Research claims that individuals' successful performance and accomplishment are directly affected by self-efficacy, and can also increase self-efficacy [12, 13].

Robust findings show that high self-efficacy positively predicts the learning process in academic settings [12, 13, 76, 77]. Self-efficacy stems from mastery experience, vicarious experience, social persuasion, and physical and emotional states, therefore it can change over time [12]. The dynamic nature of self-efficacy highlights the importance of capturing it to improve performance by applying interventions. There is a two-way positive correlation between self-efficacy and performance, as a result, they feed each other, and improvement in one lead to improvement in the other one.

Self-efficacy has been used in CS education to predict students' performance [66, 13]. Studies show combinations of self-efficacy and gender show different patterns in learning computer programming [13]. One of the most common methods to measure self-efficacy has been having students fill out standard tools such as Motivated Strategies for Learning Questionnaire (MSLQ)

[13]. The MSLQ is a widely used self-report tool to measure student motivation and learning strategies [13]. In [78] the authors develop an instrument to measure students' self-efficacy in STEM areas.

In the last section of this chapter, we discuss personality as one of the important factors which play a big role in team effectiveness.

3.3. Personality

It is mentioned that personality is the most powerful factor which influences academic performance [66]. In the field of computing and programming, in particular, there are tremendous differences between individual's performances [79]. Different personality traits have been studied as factors that can explain these differences [79]. Psychologists believe that people's personality type influences their involvement in the learning process [80]. The studies also show the correlation between certain personality traits and computer programming performance [79]. For example, an extensive review of the literature concludes that internal locus of control as a personality trait is positively correlated with the academic performance [79], and that conscientiousness and extraversion lead to individual achievement and also impact project success in teamwork [66, 81]. Current techniques in developing personalized learning systems target learners' personality traits and their effect on learning style to improve student learning experience [80].

Another trait of personality is being either introvert or extrovert [79]. Many studies have been conducted in identifying the correlation between being an introvert and computing programming success. However, there are controversial results to support this hypothesis and more research needs to be done before any conclusions can be made in this regard [79].

Aside from the impact of personality on individual performance in computing, multiple research has been conducted about the impact of personality on team performance. As claimed in [66], personality is one of the elements that can influence students' in collaborative learning. It impacts the way students form teams, how they behave in teams and apply their knowledge in practical applications [66]. Research on the impact of personality traits in teamwork shows that homogeneous teams have more positive teamwork experience, however, much cohesion has certain drawbacks too such as slowing down the critical thinking behavior [81].

Myers-Briggs Type Indicator (MBTI) is an instrument that has been widely used by educational psychologists to define personality types [80]. The MBTI questionnaire defines personality traits in four spectrums of 1) Extraverted/Introverted, 2) Sensing/Intuitive, 3) Thinking/Feeling, and 4) Judging/Perceiving. Besides the popularity and vast application of MBTI in the educational and business domain, major criticism has been made on it. The pitfalls of MBTI are that it does not acknowledge that some people may have neutral preferences on some dimensions, and also the forced-choice ipsative response format is another drawback that can yield to negative correlation among items [82].

The "big five personality traits" is a reliable instrument, developed by Elshout and Akkerman (1975), which is the first published personality questionnaire that measures the 5 personality factors: extraversion, sociability (or agreeableness), conscientiousness, neuroticism and culture (or openness to experience) [83, 84]. These 5 dimensions reflect individuals' patterns of thought, emotion, and behaviors which can be assessed by either self-report or other-report [84].

In summary in this chapter, we discussed different aspects of the attitude domain and how they impact the learning process, teamwork, and performance in the academic setting.

Researchers have applied multiple methods to recognize and quantify these attributes. The pros and cons of existing methods in identifying the attitude constructs were discussed.

In the next chapter, we propose our methodology to operationalize these attributes in low-stake teams to identify how they correlate with the individual's performance. Data collection protocol and how we analyze the data to identify attitude components will be discussed. Finally, data analysis results will be presented.

CHAPTER 4: METHODOLOGY TO OPERATIONALIZE ATTITUDE CONSTRUCTS

In this research, we analyze attitudinal components by extracting affective states from students' speech during the class activities and applying standard self-report tools to measure self-efficacy and personality types. We further analyze the correlation between attitude dimensions and students' performance.

The rest of this chapter explains our methodology to operationalize attitude constructs and our data collection approach.

4.1. Affect

For measuring affect (sentiment) from speech we record students' conversations once they start working on the activity during the class. The recording process is done every session that students work in teams. The recorded audio files are transcribed for conducting text mining and sentiment analysis. We employ different methods to analyze students' sentiments in their speech and identify their correlation with performance. First, we measure the polarity, frequency and intensity of the sentiments in three classes of positive, negative, and neutral sentiments. Next, we extend the number of sentiment classes by multi-class sentiment analysis to determine which classes of sentiment can be applied as predictors of performance. Finally, we perform topic modeling and thematic analysis of the sentiments (aspect-based sentiment analysis) to explore the context and themes in which students expressed more positive sentiment as they speak in the class.

Audio transcription was done on the speech corpora to perform text mining methods. For this purpose, 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. Since each team was recorded by one recorder (with a paired microphone), their speech utterance was stored into one file. Moreover, because the teams were sitting close to each other some speech from the nearby team was captured in the audio of each team. On some other occasions when teams had questions from the instructor or TAs, we had the speech of the third person in the audio. These caused transcriptions to include 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 the speech related to each ID (i.e. student) into a separate dataset. This resulted in 28 datasets each containing the transcription of speech of one individual in multiple sessions of the class. At this point, the textual corpora of the speech were ready for running text mining and sentiment analysis algorithms.

In the next section, we elaborate our text-mining algorithm for polarity and intensity sentiment analysis.

4.1.1. Intensity and polarity sentiment analysis

The first step of the text mining algorithm is segmentation where we segmented the speech in each dataset and created vectors based on speech initiation points 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). The text-mining algorithm for polarity and intensity sentiment analysis is shown in Figure 2.

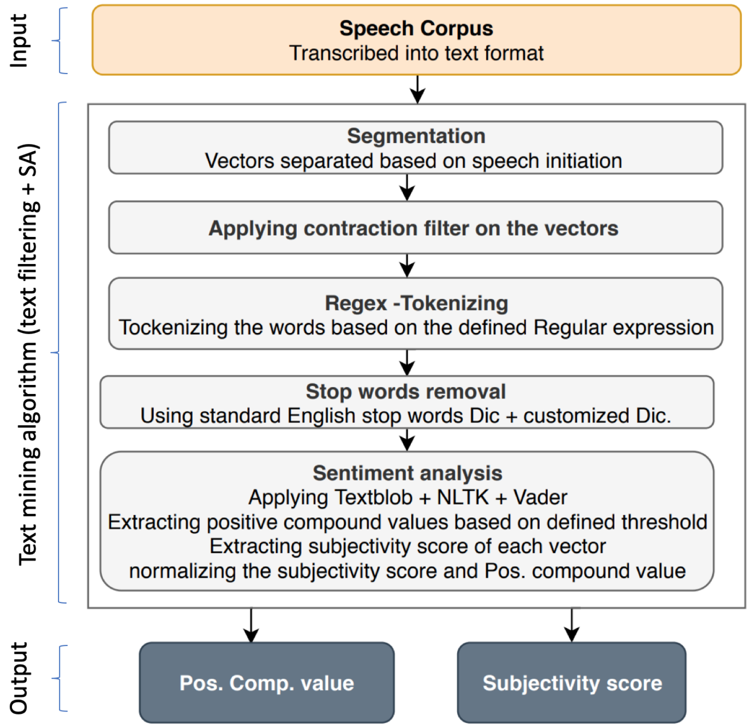


Figure 2. Text-mining algorithm for polarity and intensity sentiment analysis

Next, we applied contraction filtering on the vectors 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 our 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 created a custom dictionary to remove the redundant words that speakers used habitually which did not have any impact in the context of this study. For creating the customized dictionary, we measured the frequency of uni-grams in each dataset and defined a dictionary based on the most frequent common uni-grams and so removed those words from the text. (i.e. yeah, ok).

In the last step of the algorithm, we did sentiment analysis on the parsed text by applying lexicon-based and rule-based NLP python tools; TextBlob, Natural Language Toolkit (NKLT) and Valence Aware Dictionary for sEntiment Reasoning (VADER) to measure positive sentiments, and subjectivity score of all vectors in each dataset.

NLTK is an open-source Natural language processing platform that provides over 50 lexicon resources (i.e. SentiWordNet) with a set of libraries for word frequency, classification, tokenization, and semantic reasoning [85]. VADER is also an open-source lexicon and rule-based sentiment analysis tool. While VADER has the features of traditional sentiment analysis tools such as LIWC (Linguistic Inquiry andWord Count) it holds gold-standard quality validated by humans. VADER's advantage over LIWC is that it is more sensitive to sentiment expressions in social media as well as other domains [85]. We apply the VADER sentiment analysis algorithm due to its higher precision and accuracy in particular on short tokens at string-level compared to the other known sentiment analysis tools [86, 87, 88]. Most sentiment analysis tools have either a polarity-based or valence-based approach. Polarity determines if a part of the text is positive or negative, while valence-based approaches determine the intensity of each sentiment class as well. VADER provides both approaches to measure both the polarity and valence of text input.

The VADER algorithm 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 [89]:

$$Compound\ value = \frac{sum\_val}{\sqrt{(sum\_val)^2 + 15}} \qquad (1)$$

Where sum_val is the sum of the sentiment arguments passed to the score_valence() function in the Vadar algorithm.

The compound score is a metric for a unidimensional measure of sentiment. Depending on the context, the threshold for classifying positive, negative, and neutral sentiments can be defined. The typical threshold values for each sentiment class are [2, 85]:

Positive sentiment: compound value>=0.05

Neutral sentiment: compound value > -0.05 and <0.05

Negative compound: compound value <=0.05

In this study, to determine the threshold for neutral compound value, we did a k-means clustering on the compound values. To find the optimum number of clusters we apply the Elbow method. The Elbow method explains the percentage of the variance as a function of the number of clusters [90]. The principle of this method is that the optimum number of clusters is such that adding one more cluster does not provide better modeling of the data. The best number of clusters is chosen based on a certain point, called the 'Elbow criterion', which shows an angle in the graph [90]. In Figure 3 the percentage of the variance based on clusters is plotted against the number of clusters. The Elbow criterion indicates the optimum number of clusters is 3 as marked in Figure 3.

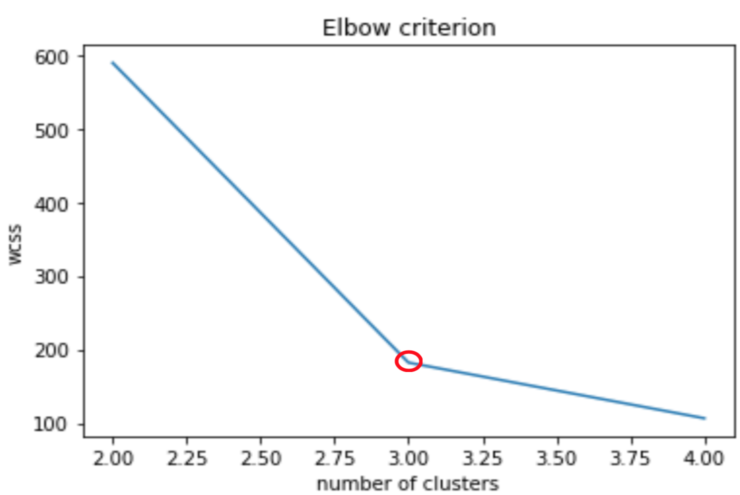


Figure 3. Elbow criterion for identifying the optimum number of clusters

The 3-means clustering of data shows we have the most density of compound values in the cluster where compound score equals 0 as shown in Figure 4.

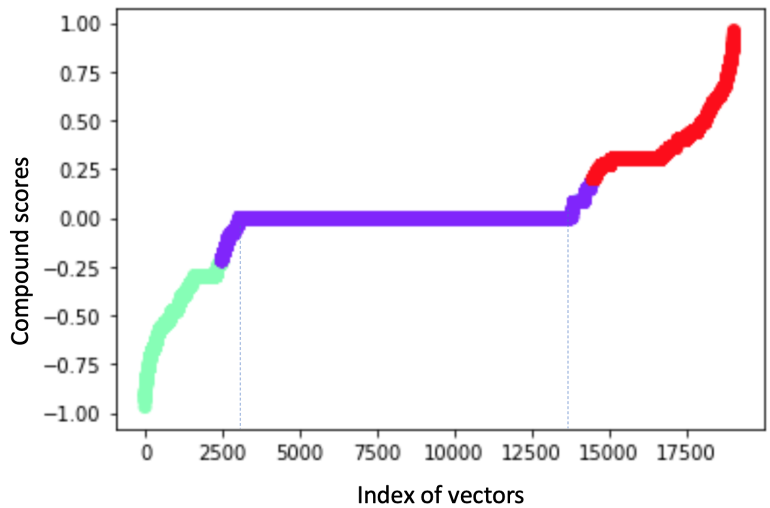


Figure 4. 3-Means clustering of compound values

herefore, we consider zero as the threshold for classifying sentences into positive, neutral, and negative. This means any vector with a negative compound score is labeled as negative sentiment, vectors with positive compound scores are considered as positive sentiment, and vectors with zero compound scores are treated as neutral sentiment class.

We further extract multiple classes of sentiment from individuals' speech which is described in the following.

4.1.2. Multi-class sentiment analysis

We extract multiple classes of sentiments from students' speech to identify the emotions that most contribute to their individual performance. For this purpose, we apply the LIWC (Linguistic Inquiry and Word Count) text analysis tool. LIWC is an effective and efficient tool for studying diverse emotional and cognitive components in verbal or written speech corpora [91].

The core of the LIWC tool for text analysis is the LIWC2015 dictionary. This dictionary includes almost 6400 word-stems and selected emotions. In this tool, a word that carries emotion such as “cry” can belong to different categories of emotions such as sadness, negative emotion, and overall affect. If the word ‘cry’ is identified in the target text the scale of each sub-dictionary category will increase [91]. The hierarchical structure of emotional categories and subcategories of the LIWC2015 is presented in Figure 5. Each category of LIWC2015 includes a list of dictionary words that define the relevant scale. A complete list of scales and scale words are presented in the LIWC2015 Development manual [91].

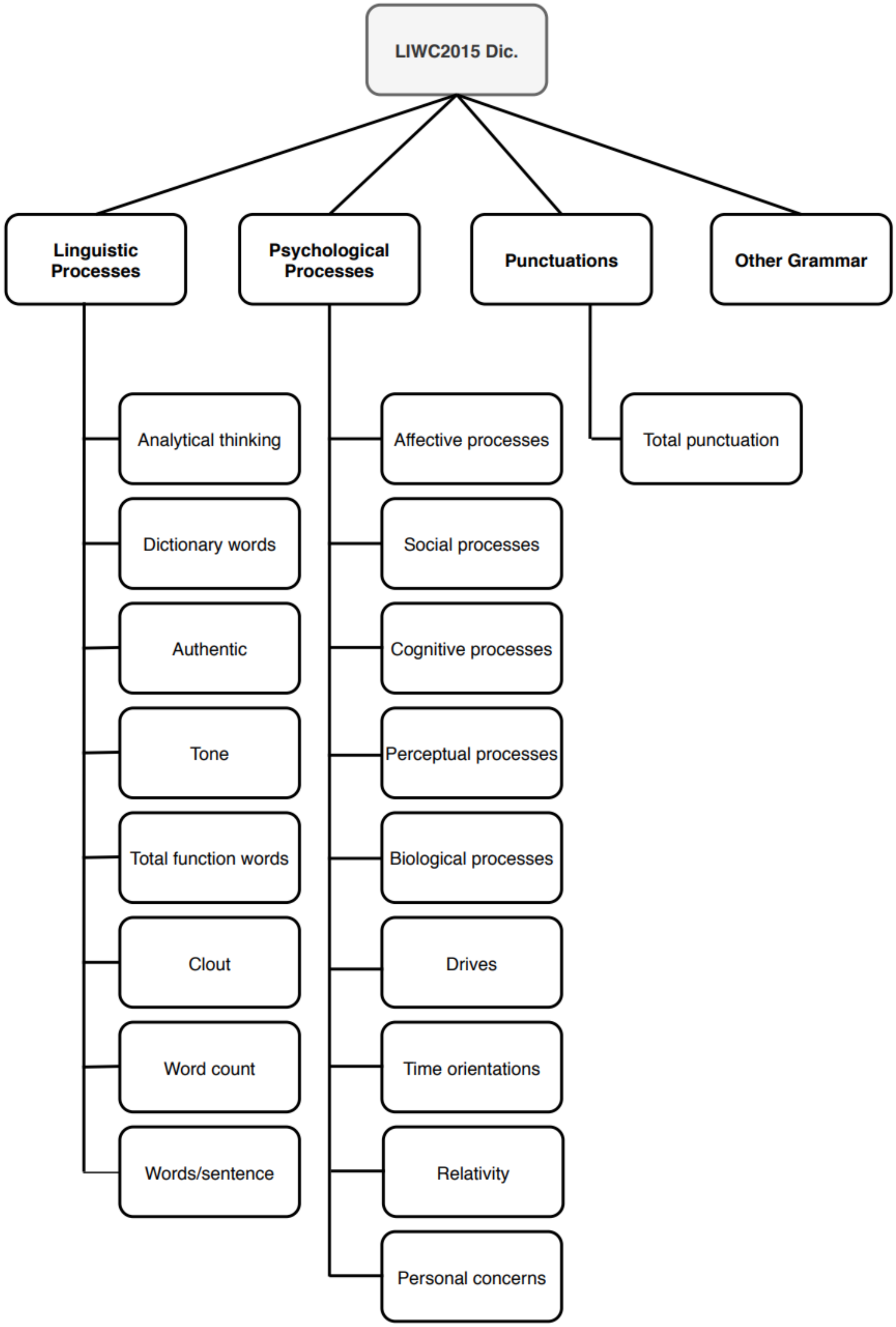


Figure 5. LIWC2015 dictionary categories

In this study, we focus on specific dictionary dimensions to analyze and identify which classes of emotions have a higher contribution to individuals' performance. There are a total number of 93 dimensions in all 4 sub-dictionaries of the LIWC2015, from which we have identified 63 dimensions based on the context of the categories and how they are related to the attitude, affective domain, social aspects and cognition process. Table 1 presents the 63 dimensions, their corresponding label in the LIWC framework, and which sub-dictionaries each belong to. The list of sub-dictionaries we identified for this study are A) Linguistic processes, B) psychological processes, C) punctuation, D) other grammar.

Table 1: LWIC dimensions and output labels

Category	Output label	LIWC Dimension	Category	Output label	LIWC Dimension
A	WC	Word count	B	cause	Causation
A	Analytic	Analytical thinking	B	discrep	Discrepancy
A	Clout	Clout	B	tentat	Tentative
A	Authentic	Authentic	B	certain	Certainty
A	Tone	Emotional tone	B	differ	Differentiation
A	WPS	Words/sentence	B	percept	Perceptual process
A	i	1st person singular	B	see	See

A	we	1st person plural	B	hear	Hear
A	you	2nd person	B	feel	Feel
A	shehe	3rd person singular	B	bio	Biological process
A	they	3rd person plural	B	body	Body
A	prep	Prepositions	B	health	Health
A	auxverb	Auxiliary verbs	B	sexual	Sexual
A	adverb	Common adverbs	B	ingest	Ingestion
A	negate	Negations	B	drives	Drives
D	verb	Common verbs	B	affiliation	affiliation
D	adj	Common adjectives	B	achieve	Achievement
D	number	Numbers	B	power	Power
D	quant	Quantifiers	B	reward	Reward
B	affect	Affective processes	B	risk	Risk
B	posemo	Positive emotion	B	time	Time
B	negemo	Negative emotion	B	work	Work

B	anx	Anxiety	B	leisure	Leisure
B	anger	Anger	B	informal	Informal language
B	sad	Sadness	B	swear	Swear words
B	social	Social processes	B	netspeak	Netspeak
B	family	Family	B	assent	Assent
B	friend	Friends	B	nonflu	Nonfluencies
B	female	Female references	B	filler	Fillers
B	male	Male	C	QMark	Question marks
B	cogproc	Cognitive processes	C	Exclam	Exclamation marks
B	insight	Insight			

The proportion of selected features for analysis is 67% psychological processes, 24% linguistic processes, 6% other grammar, and 3% punctuations as shown in Figure 6.

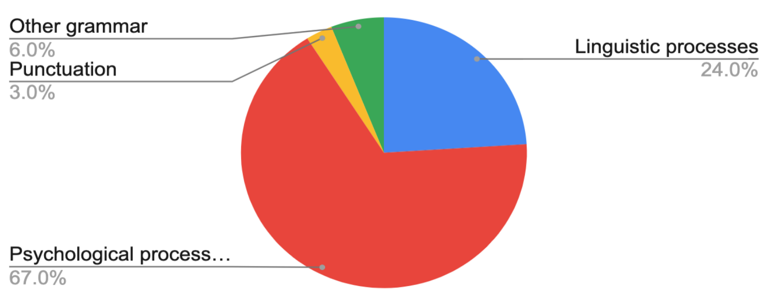


Figure 6: The proportion of selected features for analysis

We measure the scale score of each 63 dimensions by using the LIWC framework. The outcome is a numerical value for each feature.

In the next step, we conduct factor analysis by applying the Principal Component Analysis (PCA) technique which is an unsupervised statistical method to reduce the dimensionality of complex data [92]. PCA identifies the interrelation between features and by reducing the number of features it creates a new feature (component) which is the linear combination of initial features. Listed below are the main steps of performing PCA:

1. Removing the target feature and class labels. (creating a *d-dimensional* matrix)
2. Calculate the covariance matrix of the whole dataset
3. Compute eigenvectors ($e_1, e_2, \ldots, e_d$) and corresponding eigenvalues ($\lambda_1, \lambda_2, .., \lambda_d$)
4. Sort the eigenvectors by decreasing eigenvalues and choose k eighteen vectors with the larger eigenvalues to form a $d \times k$ dimensional matrix. In this study, we considered the threshold of 1 such that we choose the number of components (k) with eigenvalues higher than 1.
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subset.

One important application of the PCA method is developing predictive models. After dimensionality reduction, the dataset is divided to test and train datasets to fit the data into the training algorithm. Since this is an exploratory study the goal is to identify the correlation of sentiment classes with performance, and the development of predictive models is not the focus of this study. Predictive methods will be developed in future work after identifying the correlations of the attitude domain with performance and generating larger corpora to validate the model.

4.1.3 Thematic Analysis and Topic Modeling

The research question is whether positive sentiments have a correlation with students' performance. We take a further step to identify the direction of positive sentiments to get a better sense of the nature of positive sentiments. In the final step of the sentiment analysis, we explore the themes in which students expressed more positive emotion as they spoke in teams.

In order to do so, we apply two methods; 1) word frequency analysis and 2) topic modeling.

In the word frequency analysis, we measure the frequency of the word tokens (unigrams and bi-grams) in the positive sentiment corpora. We visualize the word cloud of the most frequent word tokens and identify the proportion of course-related tokens.

Another way to extract themes from a document is topic modeling. Topic modeling is an unsupervised machine learning technique which analyses the textual corpora and extracts word clusters (topics). Specific topics can represent themes that are not easily measurable otherwise [93]. One of the common applications of topic modeling is in the field of sentiment analysis [93]. There are diverse methods to perform top modeling. Latent Dirichlet Allocation (LDA) is one of the algorithms which has been widely practiced in the literature [94].

LDA algorithm considers the documents as bags of words and assumes that the collection of the words determines the topics. The topics are assigned to the n-grams (arrangement of words) based on the probability of belonging to each topic, and the list of topics are assigned to the documents. Basically, the learned topics are multinomial distribution over words. To identify the subject area by convention the top 10 topics are selected [95].

To get a more coherent topic modeling from LDA on the corpora with shorter text, linguistic "cleaning" is suggested complementary to other practices [95, 96]. Mehrotra et. al. [95] suggest improving the performance of LDA on microblogs of tweet data, by merging relevant tweets together which is called "tweet pooling" and applying LDA on them as a single document [97, 98].

To improve the topic modeling accuracy in our study we applied the same approach. After parsing the speech corpora using the text mining algorithm explained earlier in this chapter (Figure 2), we created vectors of speech based on the initiation point. This resulted in having short blocks of text in each vector. In order to improve the performance of LDA we merged the vectors of each speech corpora and treated them as one document, and then fitted the data into the algorithm. The emergent topics from the LDA algorithm and their analysis are presented in detail in the data analysis section.

4.1.4. Self-efficacy

For measuring self-efficacy, we employed a standard tool named "Student Attitudes Toward STEM (S-STEM) Survey" [99, 100]. This is a 5-point Likert Scale tool to measures student's self-efficacy in five categories of 'Math', 'Science', 'engineering and technology', '21st-century learning' which includes questions about teamwork skills and 'about yourself' section which asks how well the individuals expect themselves to perform in the given domain.

For the purpose of this study, we adopted the 'science', 'the 21st-century learning', and 'about yourself' components of the self-efficacy tool. The survey includes 23 questions attached in Appendix A. We conduct a pre and post self-efficacy test at the beginning and at the end of the semester to observe any correlation between students' level of self-efficacy and their performance score. Students' responses to two sections of 'Computer Science' and 'learning and social skills' are considered for analysis in this study.

We observed that all the participants filled out the pre (initial) self-efficacy survey however many of them did not fill out the post-survey. As mentioned earlier this is one of the great disadvantages of using self-report tools in attitude analysis, this led us to discard the post surveys and just consider the initial self-efficacy response for analysis. The analysis result is presented in the data analysis section.

4.1.5. Personality

For identifying the students' personality type in this study, we apply the “big five personality traits” instrument (Appendix B). This tool has 50 questions that measure the 5 personality factors: extraversion, sociability (or agreeableness), conscientiousness, neuroticism, and culture (or openness to experience). This is a 5-point Likert scale tool (1=disagree, 2=slightly disagree, 3=neutral, 4=slightly agree, 5=agree), with the output values between 0 to 40 for each personality type.

By identifying students’ personality types (i.e. being extraversion, sociability, conscientiousness, neuroticism, and openness to experience) we find the correlation between the combination of students’ personality types in a team and individual students’ performance.

The reason that we analyze combinations of personality types rather than considering an individual student’s personality type is that we do not see personality as a feature that can be easily adjusted. Thus, identifying the effect of an individual's personality types on their performance may not lead to effective cognitive interventions. Nevertheless, by analyzing how the combination of personality types impacts students’ performance we will be able to adjust team formation algorithms based on personality types.

In the following, we present our study design and data collection protocol.

4.2. Data collection

The data collection process involves two main approaches: 1) recording students' speech as they talk in teams during the class and 2) self-report tools. We conducted one phase of pilot data collection during Fall 2018 and had the second data collection phase during Spring 2019.

During data collection we encountered multiple challenges such as protocols to organize and form teams, recording students' speech, and encouraging them to participate in providing responses to self-report tools. The challenges during the first phase (Fall 2018) and takeaways helped us to have a more promising and effective data collection approach during Spring 2019.

In order to characterize the teams for our study, we applied the teamwork dimensions of the object-based design pattern model (Figure 1) to structure to various teams participating in this study. In the following, the details of each data collection phase and the lessons learned are described.

4.2.1. The First Phase of Data Collection

The pilot phase of data collection was conducted during Fall 2018. The goal of this phase was to do a feasibility study on recording students' talk in the class setting and to develop the research hypothesis based on the domain knowledge and observing how students behave in teams with different characteristics.

Based on research and our empirical evidence we believe that the most challenging components of teamwork in active learning are team formation and evaluating students' participation in teamwork. In order to come up with a promising study design, we observed teams with different characteristics in different courses that varied fundamentally.

In order to observe how students, behave in diverse team settings and identify ways to evaluate their performance, we focused on three teamwork attributes of the object-based design pattern model (Figure 1). These teamwork attributes are 1) formation, 2) size, and 3) composition.

Next, we identified four courses in CCI College who taught CS courses at different levels applying diverse team settings (i.e. having different values for each team attribute) in their active learning classes. The information about each class and their team setting is provided in Table 2.

Table 2: Selected classes and team settings in the first phase of data collection

	ITIS 8011 (HST)	**ITCS 4155 (HST)**	**ITSC 1212 (LST)**	**ITCS 6520/8520 (LST/HST)**
Formation	Background-based	Algorithmically assigned (self-efficacy score)	Random	Subjective
Composition	Complementary skills	Complementary skills	Random	Performance-based
Size	3-4	4-5	4-5	2
Diversity	Mix (F & M)	Mix (F & M)	Mix (F & M)	Mix (F & M)

Sources of data collection from these classes were: 1) self-efficacy pre/post surveys: using the same tool at the beginning and at the end of the semester, 2) recording in-class team discourses by using one recorder with built-in microphone for each team, 3) weekly reflections on teamwork experiences on how students perceive their teamwork experience (i.e. positive, negative, and neutral).

During the pilot data collection phase, we faced multiple unpredicted challenges. These challenges are listed below:

1- Multiple revisions of IRB approval caused delays in the data collection process which led to the loss of data from multiple sessions of the classes.

2- Encouraging students to participate in the research was a major challenge. At times students expressed if there was no extra credit for participation, they were not willing to participate (which was the case for this study). We also observed that the instructors' interest in conducting this research in their class had a great impact on the rate of students' participation.

3- Finding the appropriate recorder device for recording the speech with minimal environmental noise while being cost-effective was a challenge.

3- Diverse classroom structures with table and chair arrangements had a great impact on the quality of recorded data.

4- Many of the students who participated in the study did not fill out the self-report tools in a timely and committed manner.

5- At times in capstone courses when there was no obligation to stay in the class after the introduction (mini-lecture), the teams chose to work outside the class which made the recording impossible for that whole session.

In Table 3, we show the summary of these classes and how many sessions could be recorded during Fall 2018.

Table 3: Summary of the recorded session during fall 2018

Course Number	**8011**	**1212**	**4155**	**6520/8520**
Number of Teams	2	0	2	3
Team Formation	Background	Random	Self-efficacy	Performance
Size	4	4-5	4	2
Recorded Sessions	6	0	2	1
Recording Quality	Medium	N/A	Low	Good
Class Location	CEI-WWD	Kennedy	WWD-130	WWD-150
LST/HST*	HST	LST	HST	HST
Team Gender	T_1: 4 M T_2: 2F/2M	T_1: 2F T_2: 2F	T_1: 4M T_2: 2F/2M	T_1: 2M T_2: 1F/1M T_3: 1F/1M
Course Elective/Required	Elective	Required	Elective	Elective

** Low-Stake Teams (LST) - High-Stake Teams (HST)*

As shown in Table 3, the number of teams in each class that participated in the study was so limited, and even among those teams, the number of recorded sessions was not sufficient to perform the analysis.

The only class that had the highest number of recorded sessions was 8011, but due to the noise level and classroom setting, the quality of data was not acceptable for analysis.

The high level of environmental noise in the recorded audios for multiple sessions prompted us to provide additional microphones for recording the teams. However, preparing the appropriate microphones for our recorder devices caused a delay for the good quality recording of the teams during that semester. In these classes, the participants also skipped filling out the self-efficacy and teamwork reflection surveys.

Another observation that we had on the collected data was that in classes where students worked in pairs (i.e. 6520/8520) we had better voice quality and less crosstalk.

We further noticed the impact of team size on students' level of involvement in teamwork depending on the team type. In high-stake teams regardless of the team size students are mainly involved in teamwork, since everyone has normally assigned roles and tasks. As opposed to low-stake teams where students worked in teams of more than 2, we observed less engagement from some students in the team.

The lessons we learned in this phase of data collection prepared us for the second phase of data collection during Spring 2019. In the next section, we explain the data collection procedure during Spring 2019.

4.2.2. The Second Phase of Data Collection

Based on the lessons learned during the first phase of data collection we conducted a second round during Spring 2019. At this point, our hypothesis was developed and the goal was to focus on the low-stake teams where we had higher student participation in class teamwork. Based on the literature [3] and our empirical evidence we decided to minimize the size of low-stake teams to improve students' engagement level. The information about the selected class and team setting is listed below:

Course: CS1-ITSC 1212 (introduction to Java programming)

Team information:

- Team formation: Subjective
- Team composition: Random
- Team size: 2
- Team diversity: Female and male

The formation of the teams was conducted at the beginning of the semester by a gamified activity. The whole class was divided into three clusters. The clusters were formed based on the registered lab sections so that the students have the same teammate in both class and lab. In each cluster, we paired the students based on their month of birth. Since students were involved in the process of team formation, we did not have any complaints from students about their teammates. The target class was CS1 (introductory programming course) with 63 students. Among 63 students 48 students participated in the study which led to having 24 participating teams. Among the 24 teams, we could collect the recordings of 14 teams (i.e. 28 participants) who showed a more consistent pattern of attendance in the class.

In this round of data collection, we decided to add a personality type measurement tool as well. The resources for the data collection were: 1) self-efficacy pre/post self-report tools: by using the same tool at the beginning and at the end of the semester, 2) recordings in-class team discourses by using one recorder and a pair of microphones for each team, 3) personality type measurement self-report tool.

For this active learning course every week two sessions of the class were held for 75 minutes. The first 15-20 minutes of the class was dedicated to the questions and answers from the prep work in the form of clicker quizzes. Next, we had about 15 minutes mini-lecture on students' misconceptions about the material. The rest 40 minutes was dedicated to class activities in the form of low-stake teamwork. Students had to sit with their assigned teammates to do the class activity. We paired the students whose teammates were absent together so that all students had the chance to work in teams. However, if one student was absent that team was excluded for recording in that class session.

During the recording process, some environmental, technical, or human errors happened which made the data unavailable for analysis. For example, occasionally the TAs who were assigned to set up recorders misplaced the microphone cord that made the data of that session unavailable. In other cases, students accidentally pressed the stop button on the recorder which again led to the loss of data. On top of all these challenges, we had a 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 extensive research on the recording devices. The recorders we required for the study were expected to have bidirectional paired microphones, have built-in noise cancellation features, lasting battery, be user friendly, and cost-effective. After trying three different types of recorders we identified the one that best matched our criteria. Figure 7 shows the three different types of recorders we used for the study.

Figure 7: Three recorder types used in the study

A sample view of the class arrangement, chair types, and how recorders were set up is shown in Figure 8.

Figure 8: Sample view of class arrangement and recorder setup

Table 4 shows the summary of data collected from all teams during Spring 2019. In Table 4, the “R/T” denotes the sessions in which we could successfully record and transcribe students’ speech without any technical issues (count = 160). Label “R” refers to the sessions in which we recorded students’ speech however the quality of the voice was not acceptable for transcription (count = 16). Finally, the “NR” labels refer to the session in which the recording was not accomplished in the class either due to technical issues or the fact that one of the team members was absent in that session of the class (count = 48).

Table 4: Overview of data during second phase of data collection

Team ID	Participant ID	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8
4	1	R/T	NR	R/T	NR	R/T	R/T	R/T	R/T
4	2	R/T	NR	R/T	NR	R/T	R/T	R/T	R/T
3	3	R/T	R/T	R/T	R/T	R/T	R/T	R/T	R/T
3	4	R/T	R/T	R/T	R/T	R/T	R/T	R/T	R/T
6	5	R/T	NR	NR	R/T	R/T	R/T	R	R/T
6	6	R/T	NR	NR	R/T	R/T	R/T	R	R/T
2	7	R/T	R/T	R/T	NR	R/T	R/T	R/T	NR
2	8	R/T	R/T	R/T	NR	R/T	R/T	R/T	NR
7	11	NR	R	R/T	NR	NR	R/T	R/T	R
7	12	NR	R	R/T	NR	NR	R/T	R/T	R
10	13	R	R/T	R/T	R/T	R/T	R/T	NR	R/T
10	14	R	R/T	R/T	R/T	R/T	R/T	NR	R/T
8	15	R/T	R/T	NR	R/T	R	NR	NR	R
8	16	R/T	R/T	NR	R/T	R	NR	NR	R
34	17	R/T	R/T	NR	NR	R/T	R/T	R/T	R/T
34	18	R/T	R/T	NR	NR	R/T	R/T	R/T	R/T
13	21	R/T	NR	R/T	R/T	R/T	R/T	R/T	R/T
13	22	R/T	NR	R/T	R/T	R/T	R/T	R/T	R/T
18	23	R/T	NR	R/T	NR	NR	R/T	NR	R/T
18	24	R/T	NR	R/T	NR	NR	R/T	NR	R/T
21	33	R/T	R/T	R/T	NR	R/T	R/T	R/T	R/T
21	34	R/T	R/T	R/T	NR	R/T	R/T	R/T	R/T
23	35	R/T	R/T	R/T	R/T	R/T	R/T	R/T	R/T
23	36	R/T	R/T	R/T	R/T	R/T	R/T	R/T	R/T
39	41	NR	R/T	R/T	R/T	R/T	R/T	R	R/T
39	42	NR	R/T	R/T	R/T	R/T	R/T	R	R/T
28	47	NR	R/T	R/T	R	R/T	R/T	NR	R/T
28	48	NR	R/T	R/T	R	R/T	R/T	NR	R/T

As mentioned earlier one of our biggest challenges in collecting data using self-report tools was the lack of students' commitment to filling out the surveys. Table 5 shows the response rate out of total 28 participants for each self-report tool.

Table 5: Response rate to self-report tools during spring 2019

Self-Report tool	Participation Rate
Pre-self-efficacy	100%
Post-self-efficacy	82%
Personality	86%

In the next section, the result of data analysis on both students' speech and self-report tools are presented and the research hypotheses are validated.

CHAPTER 5. DATA ANALYSIS

We aim to identify the correlation between attitude constructs and students' performance and explore which of these constructs can serve as predictive metrics. For operationalizing the attitude constructs we applied both quantitative algorithmic analysis as well as qualitative analysis methods.

Students' final grade in the course is considered as the performance metric. In this course, we have multiple milestones for individual assessments which are: three major assignments, three lecture tests, and three lab tests. The contribution of each milestone to the final grade is Assignment 20%, Lecture test 30% and Lab test 30% of the final grade.

The grade distribution of participants is presented in Figure 9.

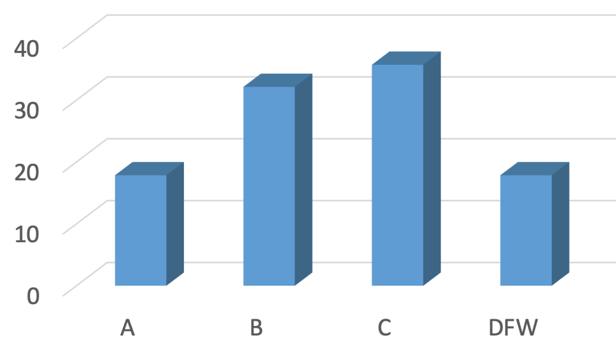


Figure 9: Grade distribution of the participants

Based on the linear regression analysis of the performance scores over the course of the semester we measure the slope of trendlines. The slope shows the ratio of grade change over time. Based on the trends we categorized the students into three groups as follow:

Upward trend (U): The trend shows increasing grades over time (Slope >0)

Downward (D): The trend shows increasing grades over time (Slope <0)

Same (S): No significant changes observed in trendline (Slope = 0)

Data shows 57% of the participants had an upward trend in their performance, 28% of the students had downward trends while 14% of the students did not have any significant change of performance during the semester. Figure 10 plots the breakdown of grades in each trend category.

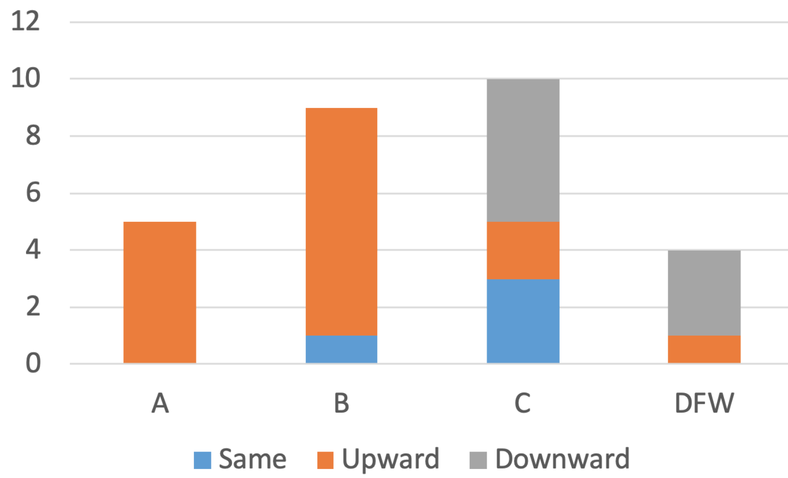


Figure 10: Trends of grades during the semester

The rest of this section is dedicated to the analysis of three attitude constructs of affect, self-efficacy, and personality to identify their correlation with students' performance and testing the research hypothesis.

5.1. Affect (Polarity and Intensity Sentiment Analysis)

To measure the intensity and frequency of affective states we do sentiment analysis by applying NLTK and VADER algorithms on speech segments [107]. The NLTK/VADER outputs the segments into vectors with a trouble of values (i.e. sentiment classes) as shown in Figure 11.

Each vector contains four features of Positive (Pos), Neutral (Neu). Negative (Neg) and Compound (Comp). In Figure 12, R is the total recorded speech of each participant during the semester and **S** is part of speech segmented based on the initiation of talks, and V is the speech vectors including four features of Pos, Nue, Neg, and Comp, where:

Corpora of all participants = {R1, R2, R2 … R28}

Total speech of each participant= {S1, S2, S3, S4 … Sn}

n = number of vectors in corpus

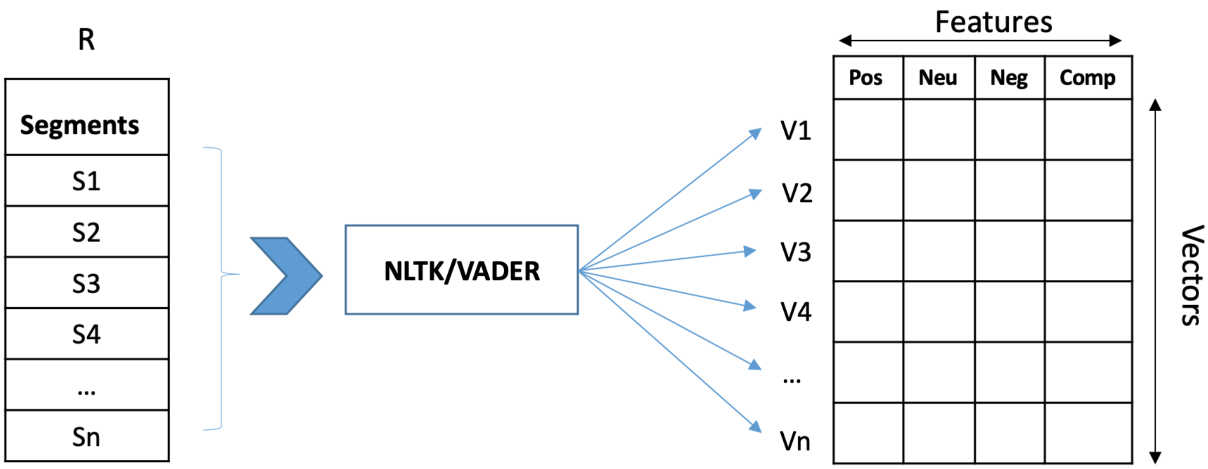
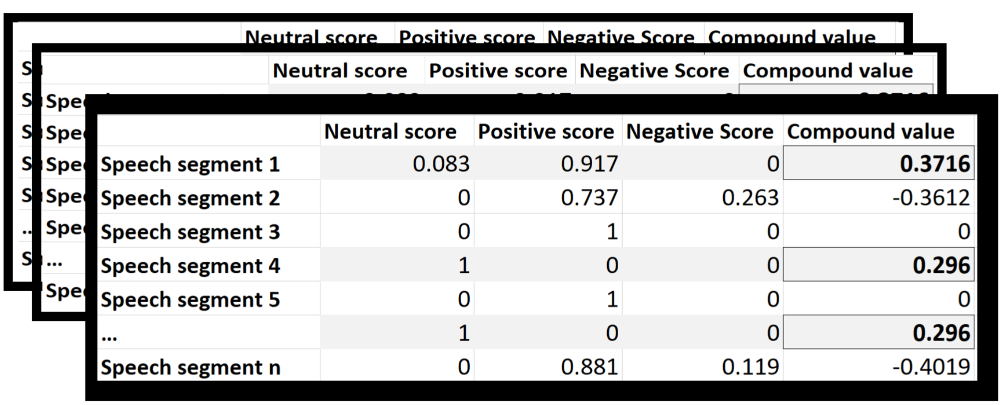


Figure 11: Applying NLTK/VADER to create vector features

We have 28 datasets each containing the speech of individual participants during the semester. Figure 12 is a sample representation of each dataset. Each dataset denotes total speech of a participant, segmented based on the part of speech (i.e. segments).



	Neutral score	Positive score	Negative Score	Compound value
Speech segment 1	0.083	0.917	0	**0.3716**
Speech segment 2	0	0.737	0.263	-0.3612
Speech segment 3	0	1	0	0
Speech segment 4	1	0	0	**0.296**
Speech segment 5	0	1	0	0
…	1	0	0	**0.296**
Speech segment n	0	0.881	0.119	-0.4019

Figure 12: Dataset representation sample

As described in chapter 4, the vectors with compound values greater than zero are considered positive sentiment. We normalize the intensity and frequency of positive sentiments based on the amount of speech. Frequency refers to the number of vectors that have positive compound value (Comp>0) and intensity refers to the actual value of compound scores of each vector.

Equations (2) and (3) shows how the mean frequency and intensity values for each participant is calculated.

$$\textit{Frequency} = \frac{m}{n} \tag{2}$$

$$\textit{Intensity} = \frac{\sum_{x=1}^{m}(comp\_value_x)}{n} \tag{3}$$

Where:

n= total number of vectors in each dataset

Vectors with positive comp_value= {1, 2, …m}

Linear regression analysis of frequency and intensity of positive compound values shows a homogeneous pattern such that students with a higher frequency of positive sentiment had higher levels of intensity in positive sentiments in their speech. Figure 13 shows the regression plot of the frequency and intensity of positive sentiments for the 28 participants.

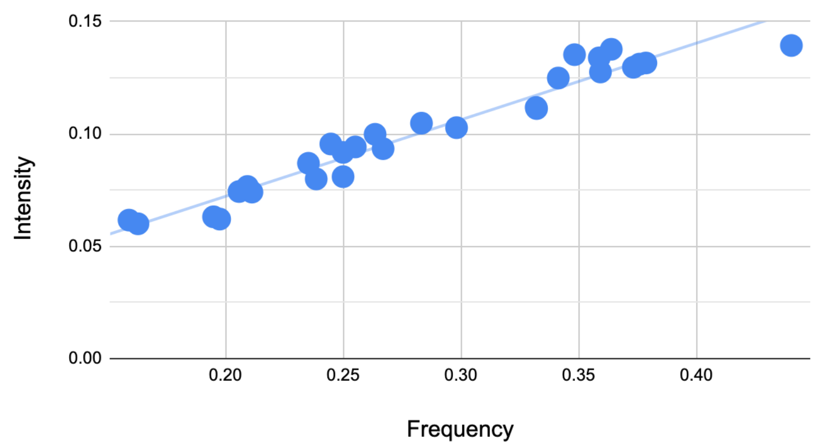


Figure 13: Regression plot of frequency and intensity of positive sentiments

In Figure 14 we present the scatter plot of frequency and intensity of positive sentiments with the normalized value of students' performance.

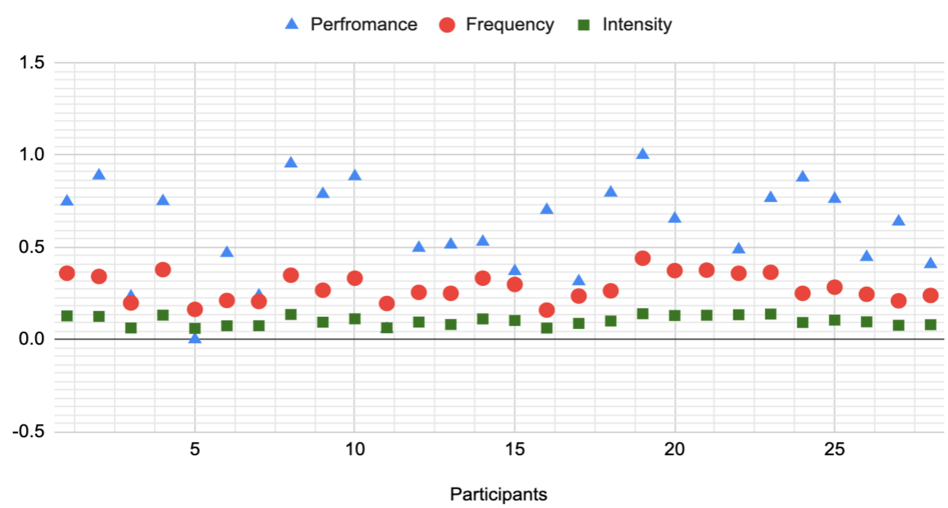


Figure 14: Frequency, intensity, and performance scatter plot of the participants

The trend of all compound values (not just positive compound scores) of the four performance categories (i.e. A, B, C, DFW) is analyzed by measuring the kernel density of compound values for each grade category as shown in Figure 15. In these plots, the horizontal axis shows the compound value and the vertical axis denotes the density of each value.

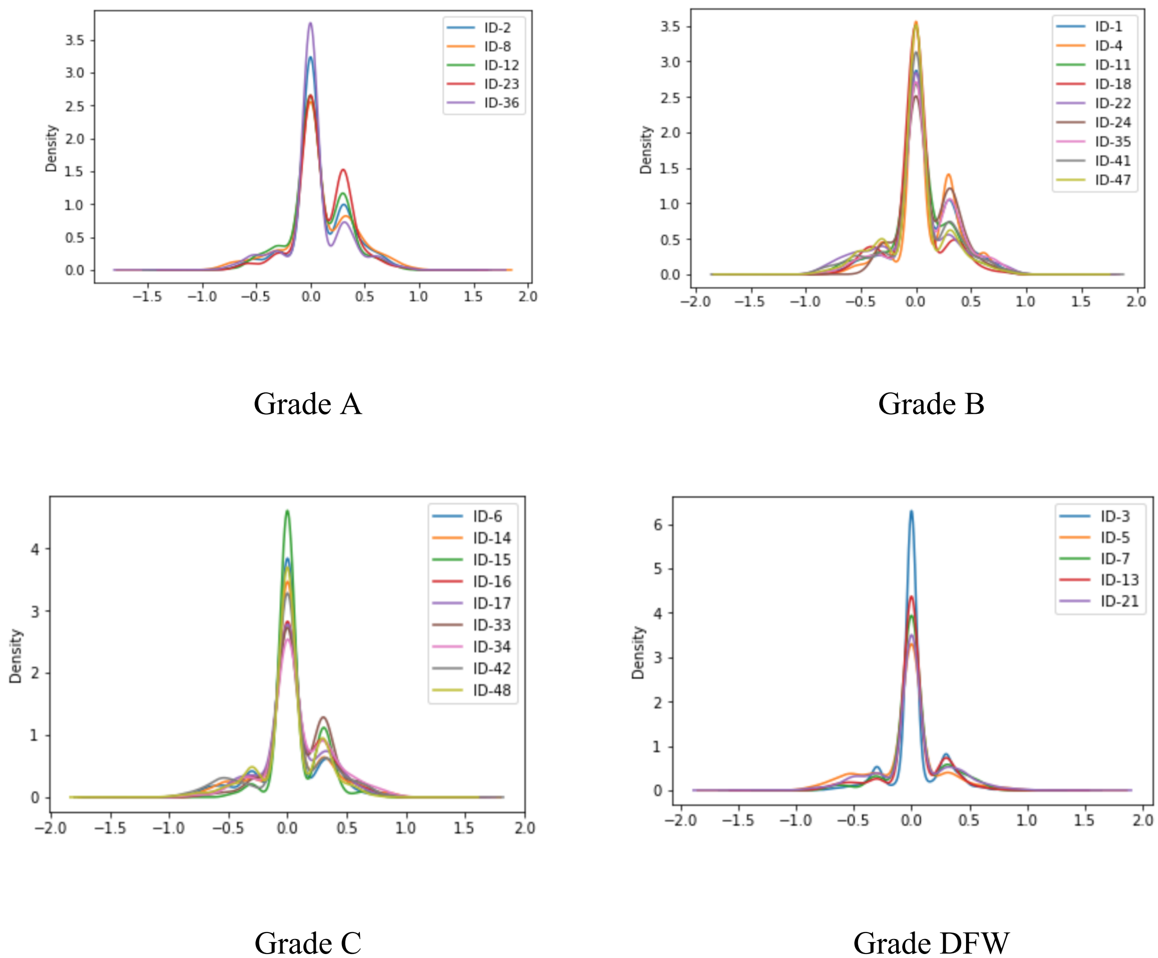


Figure 15: Kernel density plots of compound values of different grade categories

The plots show more density of positive compound values for students with higher grades, and as the students' grades decrease the density of positive compound values decrease. In the plot of grade A, we can see the most density is related to compound values between 0 to 0.5, but in grade B plot we observe the density shift from positive sentiments towards zero and we observe more density in negative values compared to grade A. Accordingly in the plot of grade C, we observe a significant decrease in the density of positive compound values as the density of neutral compound values increases compared to the A and B grades. Finally, for the DFW grade category, the density of positive sentiment decreases compared to the grade categories of A, B, and C, and the neutral values increase. In summary, the general trend in the kernel density plots shows we have less density in positive compound values as the grades decrease.

To answer the research question, we measure if there is any positive relationship between positive sentiments (intensity and frequency) and the performance score by applying Spearman's rank correlation coefficient. Spearman's rank correlation coefficient is a nonparametric (distribution-free) rank statistic for measuring the strength of an association between two variables [101]. 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 [101]. 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 [102]; r_s = 00-.19 "very weak", r_s =.20-.39 "weak", r_s = .40-.59 "moderate", r_s = .60-.79 "strong", r_s = .80-1.0 "very strong".

The r_s value is calculated using equation (4): (n = number of cases)

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad (4)$$

Where d_i is the difference in ranks for variables

By using Spearman's rank correlation coefficient equation, we measured the correlation between positive compound values (intensity and frequency) with the performance scores. The coefficient values (r_s) are presented in Table 6.

Table 6: Coefficient values of positive sentiment, subjectivity, and performance

	Coefficient value (r_s)	
Frequency and performance	0.61	This is a strong positive correlation
Intensity and performance	0.64	This is a strong positive correlation

The coefficient values (r_s) related to intensity of positive sentiments and frequency of positive sentiments shows that both have a strong positive correlation with performance.

Figures 16 and 17 visualize the linear regression analysis of intensity and frequency of positive sentiments vs performance score. In these plots, the horizontal axis specifies the intensity and frequency of positive sentiments and the vertical axis shows the performance score of each participant.

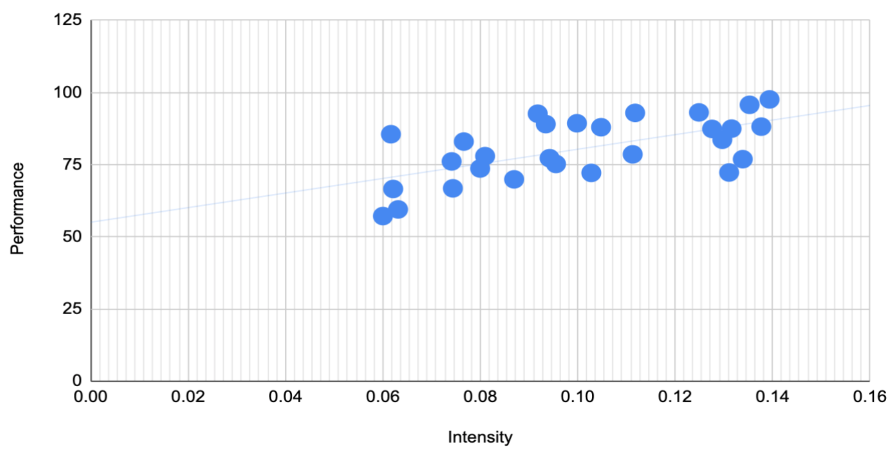


Figure 16: Intensity of positive sentiment vs performance regression plot

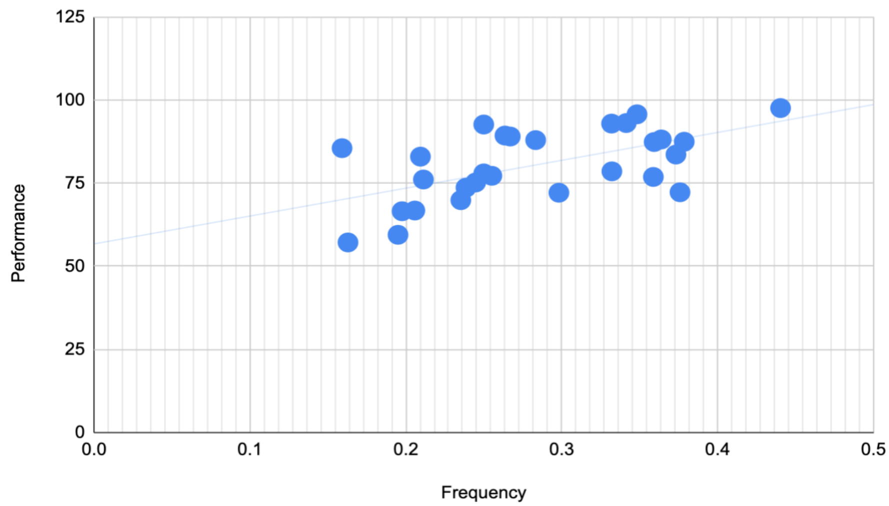


Figure 17: Frequency of positive sentiment vs performance regression plot

To test the first hypothesis (H1), we applied a chi-square test and evaluated the two-tailed *p* value with the confidence level of 0.05.

The Null hypothesis states that**:** There is no correlation between students' positive sentiment (frequency and intensity of compound values) in low-stake teams and their individual performance (grade).

The calculated *p* value is compared to the confidence level of 0.05. If the *p* value is less than 0.05 the null hypothesis is rejected which indicates there is a statistically significant relationship between positive sentiments and students' individual performance. The two-tailed *p* value is presented in Table 7.

Table 7: *p* values of frequency and intensity of positive sentiment, and performance

	two-tailed P-value	
Frequency and performance	0.001	Statistically significant correlations
Intensity and performance	0.002	Statistically significant correlation

The calculated *p* value for the intensity of positive compound value is 0.002 and the frequency of compound value is 0.001. Since the *p* value for both the intensity and frequency of positive sentiments are less than the confidence level, therefore the null hypothesis is rejected which confirms that there is a statistically significant correlation between students' positive sentiment (frequency and intensity) and their performance.

By answering the research question, which confirms the positive correlation between students' sentiment and their performance, we take one step further and classify students' sentiments into multiple classes beyond positive, negative and neutral to capture more psychological constructs such as affect, cognition, drives, etc. and investigate which classes can be used as predictive metrics for performance as the target value. In the following, we present our results on multiclass sentiment analysis.

5.1.1. Multiclass Sentiment Analysis Using LIWC Dimensions

We applied the LIWC text analysis tool to extract sentiment class from the corpora.

As explained in chapter 4, 63 features out of 93 features were chosen for this study. The selected features are in psychological, affective, and cognitive domains.

The output features in the LIWC platform scale differently depending on the variable. For example, the range of word count (WC) is greater than the affective variables such as negative emotion (negemo). To standardize the distribution of the data for analysis we applied the StandardScaler() algorithm from the SKLEARN library. This algorithm standardizes the features by removing the mean and scaling to the unit variance [103].

The standard score of sample x is calculated with equation (5):

$z = (x-u)/s$ (5)

Where u is the mean and s is the standard deviation of the training sample. Figure 18 plots the distribution of scaled features for the 28 participants.

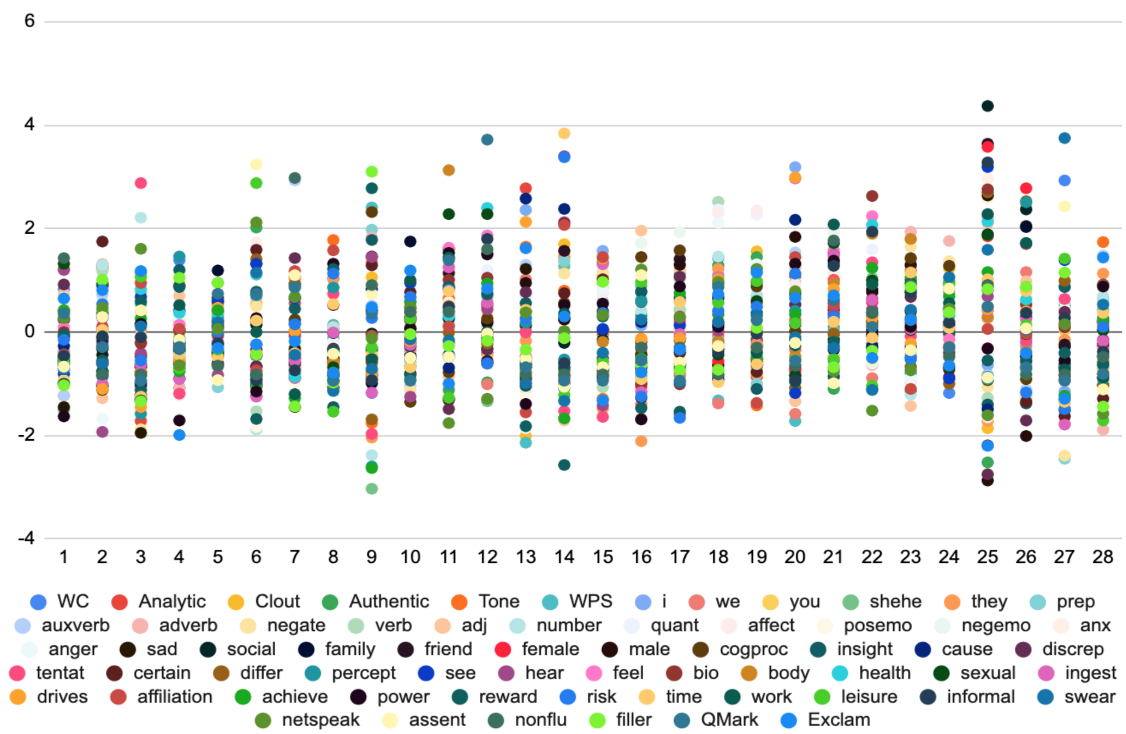


Figure 18: The distribution of scaled features

One common practice when having a high dimension of data is dimensionality reduction. With high dimensionality of the data it is challenging to interpret the features and use them for developing predictive models. tIn order to reduce the feature space dimension to most critical ones, we use the Principal Component Analysis (PCA) method to minimize the dimensionality while preserving as much 'variability' (i.e. statistical information) as possible.

Here the question is identifying the best number of components to maintain the originality of the dataset. For this purpose, there are two approaches to find the optimum number of components. We used both methods to ensure we get consistency in the number of principal components. The steps of the first approach described in Figure 19.

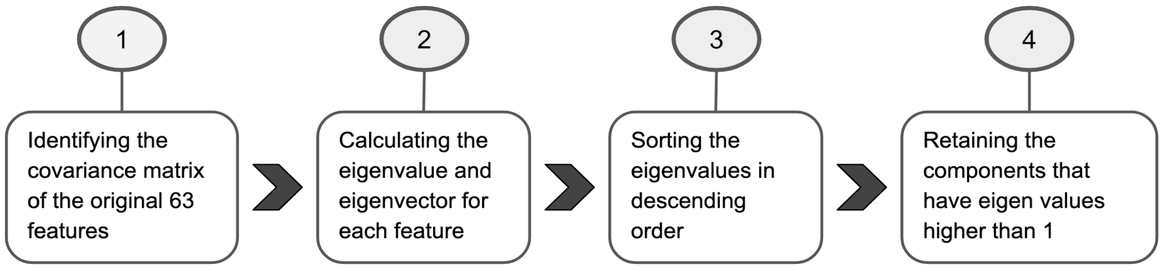


Figure 19: Component identification process

The eigenvalues in descending order (step 3 in Figure 19) are presented in figure 20. As highlighted in Figure 20, the last eigenvalue with a score higher than 1 conjunct with the PC-14 which implies we need 14 components to represent the sample data.

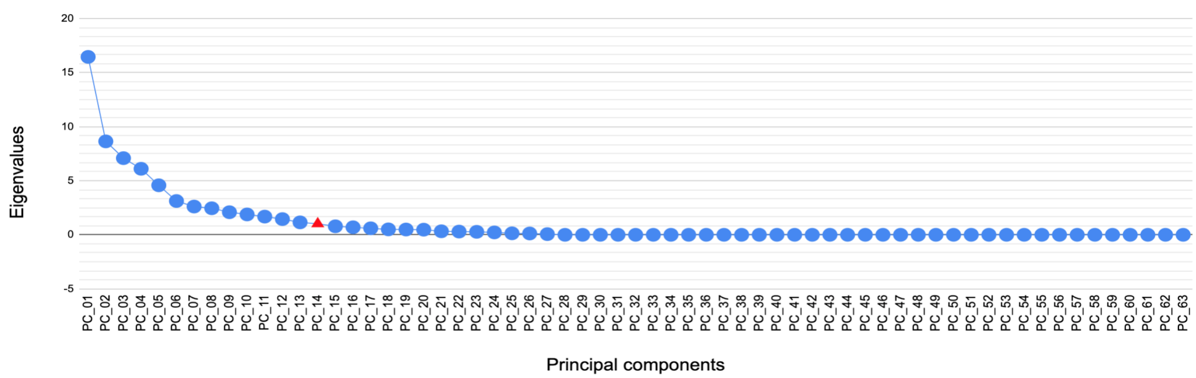


Figure 20: Eigenvalues of the principal components

Another way to determine the minimum number of principal components is to calculate the cumulative proportion of variance that the components explain. The acceptance level depending on the application can vary from 80% to 95%. For descriptive purposes, 80% would be sufficient. With the 14 principal components retained from the previous method, we have 92% variance which is an acceptable number to represent data. The Scree plot of the LIWC features is presented in Figure 21.

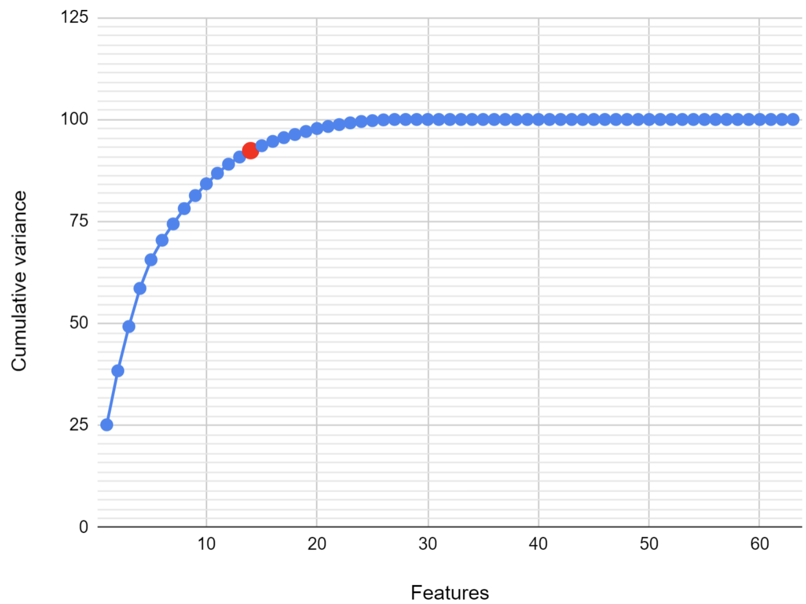


Figure 21: Scree plot of LIWC dimensions

The following heatmap (Figure 22) visualizes 14 principal components and the 63 LIWC features. The color bar represents the correlation between the original features and the principal components. The lighter is the color the more positive correlation exists between the features and the principal component whereas the darker colors represent the negative correlation.

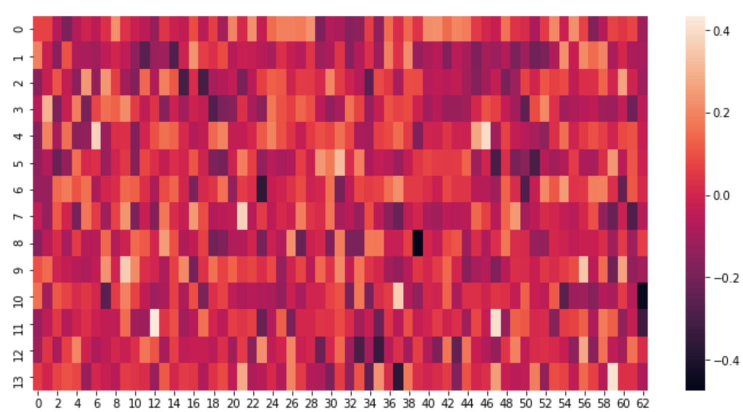


Figure 22: The heatmap visualization of principal components vs. features

The 14 principal components corresponding to combinations of original features are stored as the attributes of the fitted PCA object which are not easy to interpret and understand. In order to interpret each principal component in terms of the original variables, we examine the magnitude of the coefficients for the original variables. Larger absolute values of coefficients show that corresponding variables have more importance in calculating the component. Table 8. maps the maximum absolute values of the coefficients with the original LIWC features. In other words, it lists the most relevant features in terms of principal components that have higher levels of importance in determining the target value which is students' performance.

Table 8: Mapping of principal components to the original LIWC features

Principal Components	LIWC Features	Principal Components	LIWC Features
PC_1	anx	PC_8	negemo
PC_2	negate	PC_9	hear
PC_3	verb	PC_10	shehe
PC_4	Analytic	PC_11	Exclam
PC_5	drives	PC_12	auxverb
PC_6	insight	PC_13	certain
PC_7	anger	PC_14	nonflu

The result of this analysis helps to train the predictive algorithms on larger samples to predict student's performance based on their psychological, affective, and cognitive behaviors. This is beyond the scope of this study to fit the sample data into predictive models. This will be considered in future work.

In the next section, we explain our approach to thematic analysis (aspect analysis) and topic modeling on students' positive sentiments.

5.1.2. Thematic analysis (word frequency and topic modeling)

The analysis of the data so far shows there is a correlation between students' positive sentiments and their performance. The question here is what is the subject or context of positive emotion in students' speech? Their positive emotion could be about the course content, which is what we expect it to be, or can be about their friends, or any other experience.

To answer this question, we dig more into the positive sentiments to identify themes by applying word frequency and topic modeling algorithms.

First, we adopt the word frequency approach in which the positive sentiment vectors are parsed to extract unigrams and bigrams. The frequency of unigrams and bi-grams are measured to decide the dominant theme in students' positive sentiments.

The normalized frequency of the top 30 unigrams and bi-grams of positive sentiment vectors (Comp > 0) are plotted in the following graphs (Figures 23 and 24).

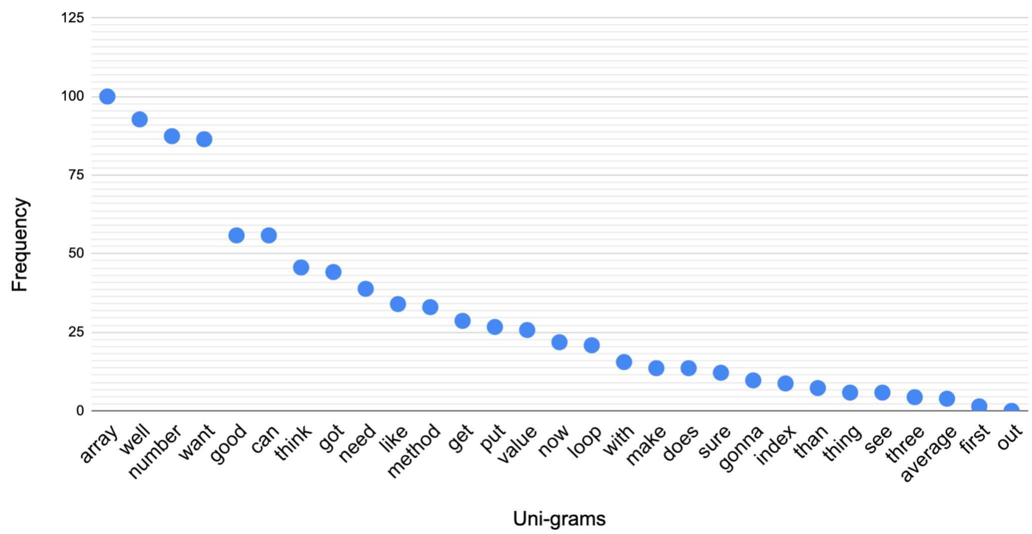


Figure 23: Frequency of Uni-grams in positive sentiment vectors (comp>0)

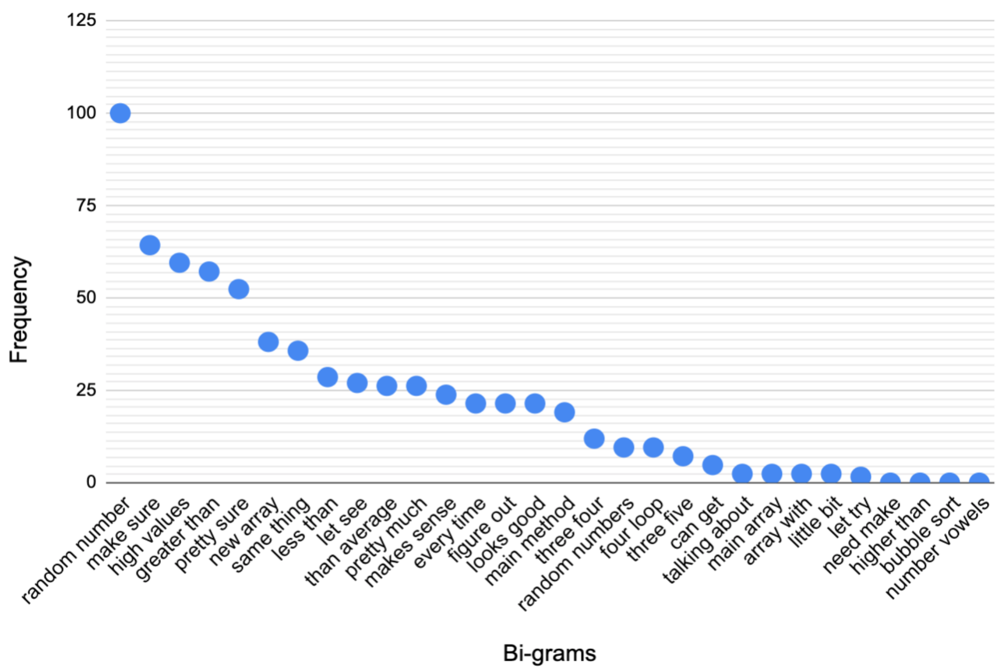


Figure 24: Frequency of Bi-Grams in positive sentiment vectors (comp>0)

By using “wordcloud” and “sklearn” libraries in python the word cloud visualization of positive sentiments vectors is presented in Figure 25.

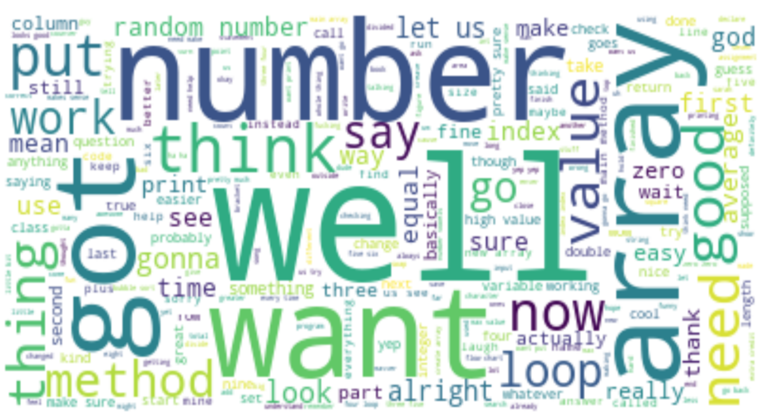


Figure 25: Word cloud visualization of positive sentiment vectors

The plots in figures 23 and 24 as well as the word cloud (Figure 25) show that the most dominant theme of word tokens frequently used in the positive sentiment vectors are specifically about the course content (i.e. Array, random number).

In the following pie charts (Figures 26 and 27), the proportion of course-specific unigrams and bi-grams in the positive sentiment vectors are presented.

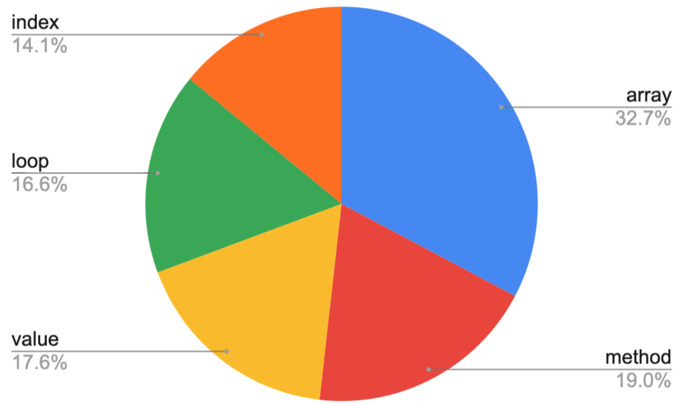


Figure 26: Proportion of course-specific uni-grams in the positive sentiment vectors

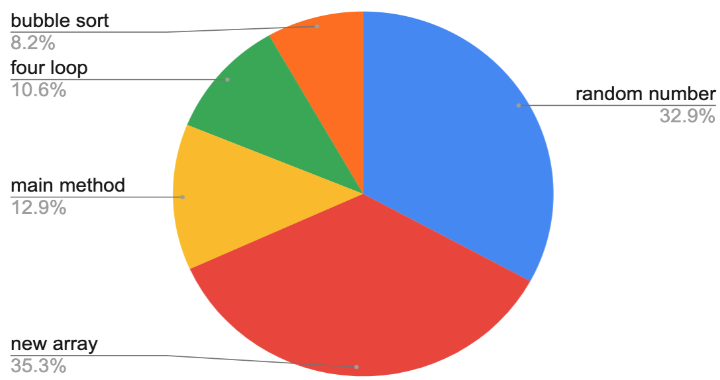


Figure 27: The proportion of course-specific bi-grams in the positive sentiment vectors

In summary, the result of frequency analysis shows top frequent unigrams and bi-grams in the positive sentiment vectors (Comp > 0) are specifically in the context of the course. This implies students' positive sentiments were mainly centered on the course topic.

Another evidence of this is that the class activities were designed in such a way that students who were well prepared could finish them within 35-40 minutes. This means they did not have extra time to discuss the topics outside the course content.

Next, we apply a topic modeling approach by using the LDA algorithm on the positive sentiments vectors to extract the top 10 dominant topics in the positive sentiment samples. In this algorithm, a list of words from the document are mapped to the topics. In Table 9. the list of ten topics and the top 3 keywords related to each topic is presented:

Table 9: Extracted topics and keywords by LDA algorithm

Topic No.	Keywords	Topic No.	Keywords
1	[Array, value, index]	6	[Like, values, got]
2	[Array, plus, yep]	7	[Alright, really, wait]
3	[Want, method, think]	8	[loop, does, fine]
4	[number, random, cool]	9	[god, zero, true]
5	[Good, sure, thank]	10	[shall, work, easier]

By applying the pyLDAvis library visualization of topics and the inter-topic distance map is presented in Figure 28. Figure 28 shows the overall term frequency of all topics. The estimated term frequency within each topic is presented in the Appendix C. In figure 28 the topics marked by green color are the ones that include at least one course specific keyword and the red ones are those that contain general words or action verbs.

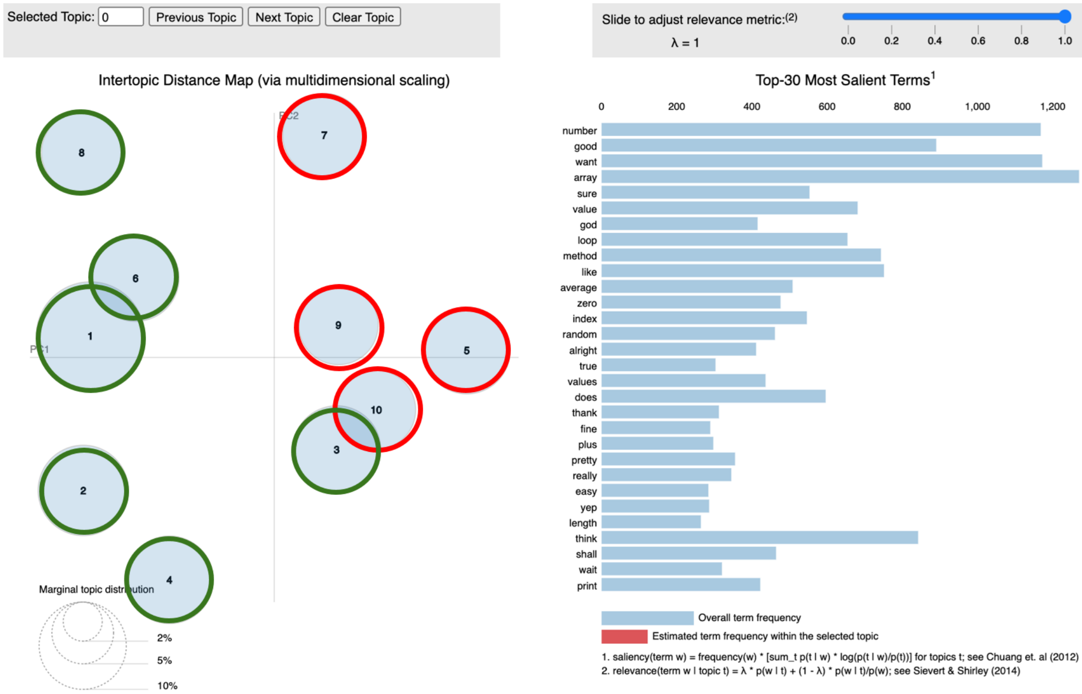


Figure 28: Inter-topic distance map

Based on the inter-topic distance map shown in Figure 28 we visualize the topics into 5 clusters such that the topics that are overlapped or have relatively closer distance are located into one cluster as shown in Figure 29. We observe in all clusters except the one highlighted in red color, there is at least one keyword [bolded] that is highly related to the course content.

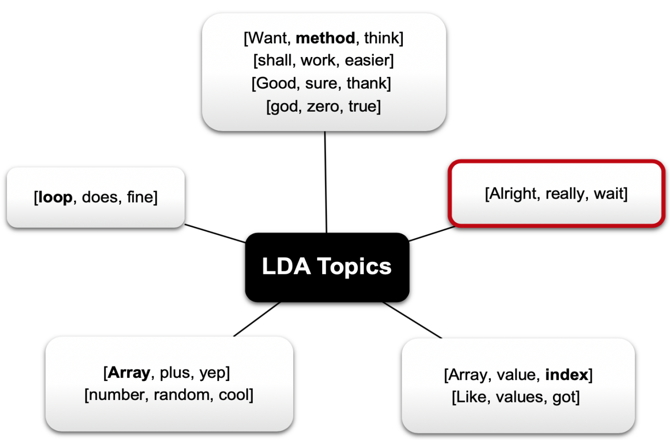


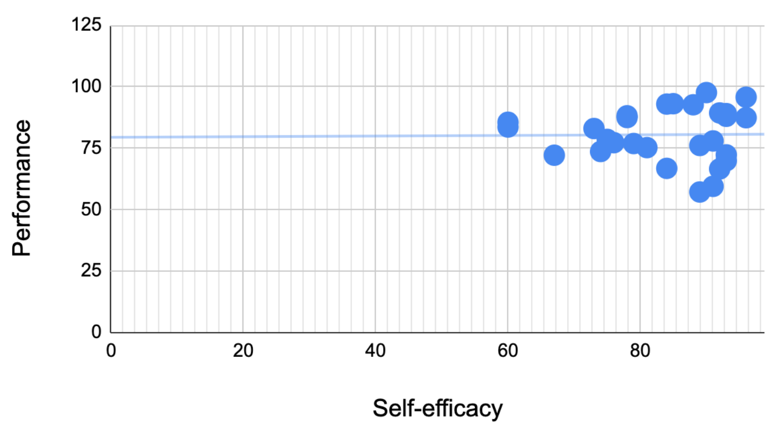
Figure 29: Topic clusters based on the inter-topic distance map

In summary, the thematic analysis and topic modeling of the positive sentiment vectors confirm that students' positive sentiment was mostly about the course content. In the next section, we present an analysis of the self-efficacy self-report tool.

5.2. Self-efficacy

We measure students' self-efficacy with a standard tool called "Motivated Strategies for Learning Questionnaire" (MSLQ). For this, we asked the participants to self-report on a 5-point Likert-scaled survey including 20 questions. The answers ranged from strongly disagree= 1 to strongly agree = 5). These 20 questions are grouped into 2 main categories of computer science, land earning/social skills. The questionnaire is provided in Appendix A.

There are some debates in the literature on whether we can apply parametric or nonparametric methods to the ordinal data. Multiple studies such as the research conducted by Norman G. (2010) supports both parametric and non-parametric tests are applicable to the Likert scaled data [104]. By running Spearman's rank correlation coefficient, the calculated value of $r_s = 0.11$ which denotes a positive but very weak correlation between the two variables of self-efficacy and performance. Figure 30, Shows the regression plot of self-efficacy and performance. In this plot, the horizontal axis is the sum of the 20 data points of the self-efficacy tool for each participant.



Figures 30: Regression plot of self-efficacy vs performance score

In the H2 Null hypothesis states, there is no correlation between students' initial self-efficacy and individual performance. The alternative hypothesis occurs when there is a correlation between students' initial self-efficacy and their individual performance. The calculated *p* value is .61 which is higher than the confidence level of 0.05, and the Null value cannot be rejected. The p-value confirms that there is no statistically significant correlation between the initial self-efficacy and the performance.

In H3 the Null hypothesis states that there is no correlation between students' initial self-efficacy and positive sentiments (frequency and intensity of vectors with Comp > 0) in low-stake teams. By running Spearman's rank correlation coefficient, the r_s and *p* values indicate there is a weak positive correlation between the self-efficacy and positive sentiment however the correlation is not statistically significant. Therefore, the Null hypothesis cannot be rejected. The calculated r_s and *p* values of frequency and intensity of positive sentiments with self-efficacy are presented in Table 10.

Table 10: The result of spearman's correlation coefficient test

	Self-efficacy	
Frequency of Pos_Sen	$r_s = .06$	By normal standards the association between the two variables is not statistically significant.
	$P = .77$	The result is not significant at $p < 0.05$
Intensity of Pos_Sen	$r_s = .06$	By normal standards the association between the two variables is not statistically significant.
	$P = .75$	The result is not significant at $p < 0.05$

Next, we investigate which category of self-efficacy questions have more impact on students' performance. (i.e. computer science or learning and social skills).

For this purpose, we applied Principal Component Analysis (PCA) method to identify the main components that better represent the dataset with students' performance as the target value.

We determine the number of components based on the value of eigenvalues such that the components that have eigenvalues higher than 1 are retained. The plot of eighteen values (Figure 31) determines we need only 7 component values (highlighted in red color) to represent the sample.

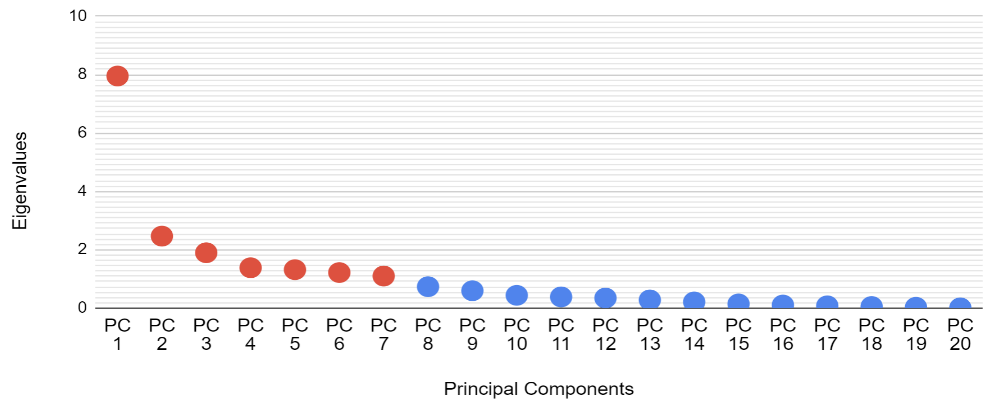


Figure 31: Scatter plot of the eigenvalues of the principal components

To interpret each principal component in terms of the original variables, we examine the magnitude of the coefficients for the original variables. Larger absolute values of coefficients show that corresponding variables have more importance in calculating the component. Table 11 shows the self-efficacy questions related to each principal component. The maximum coefficient values of the two PCs were related to one question therefore the total number of 6 questions is listed based on the 7 principal components.

Table 11: Self-efficacy questions with highest coefficient values

ID	PC	Self-Efficacy Questions	Category
Q1	PC_4 & PC_5	Computer science will be important to me in my life, and work.	Computer science
Q2	PC_1	I am confident I can encourage others to do their best.	Learning/social skills
Q3	PC_6	I am confident I can respect the differences of my peers.	Learning/social skills
Q4	PC_3	I am confident I can manage my time wisely when working on my own.	Learning/social skills
Q5	PC_2	When I have many assignments, I can choose which ones need to be done first.	Learning/social skills
Q6	PC_1	I am confident I can work well with students from different backgrounds.	Learning/social skills

As shown in Table 11, out of 6 questions (each represented by a principal component), 5 of them are explicitly about learning and social skills. This shows that students' level of self-efficacy about their learning and social skills plays an important role in their performance. We further investigated the range of responses in these 6 self-efficacy questions and plot the divergent stacked bar chart in Figure 32.

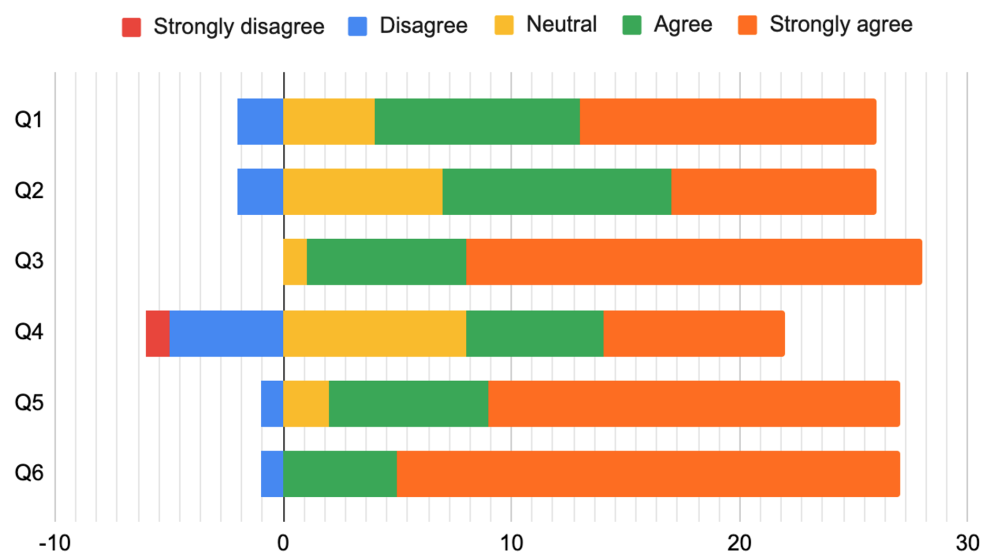


Figure 32. Divergent stacked bar of the self-efficacy questions

As shown the question which has the most negative score (disagree and strongly disagree) is the Q4, which asks if students can manage their time when they work on their own. This implies that teamwork can help students manage their time wisely and reduce the procrastination rate.

We argue that lack of correlation between self-efficacy and students' performance can be due to multiple reasons. One is that students' self-efficacy can be changed over time and thus the initial self-efficacy might not be a good metric to predict performance. Moreover, we only considered the initial self-efficacy for analysis since we have a very low response rate in the post self-efficacy tool. The other reason could be that since sample data is from a CS1 course which is the first course the students take in the major, they might not have much domain knowledge about the CS to provide precise answers to the self-efficacy tool. Finally, self-report tools generally cannot be as reliable as methods that observe the behavior directly.

In the following, we present the analysis of the personality test.

5.3. Personality

In this study in order to identify students' personality types, the "big five personality traits" instrument was applied. Students were asked to self-report on the personality tool. From the total number of 28 participants, 24 students provided answers to the questionnaire.

This tool measures the five personality factors of Extraversion (E), Agreeableness (A) (aka sociability), Conscientiousness(C), Neuroticism (N), and Openness to experience (O). The factors reflect individuals' patterns of thought, emotion, and behaviors. The quantified measure of each of these factors is a number between 0 to 40 which are presented in Figure 33. In this figure, only the dominant trait of personality for each individual (i.e. the maximum value) is plotted.

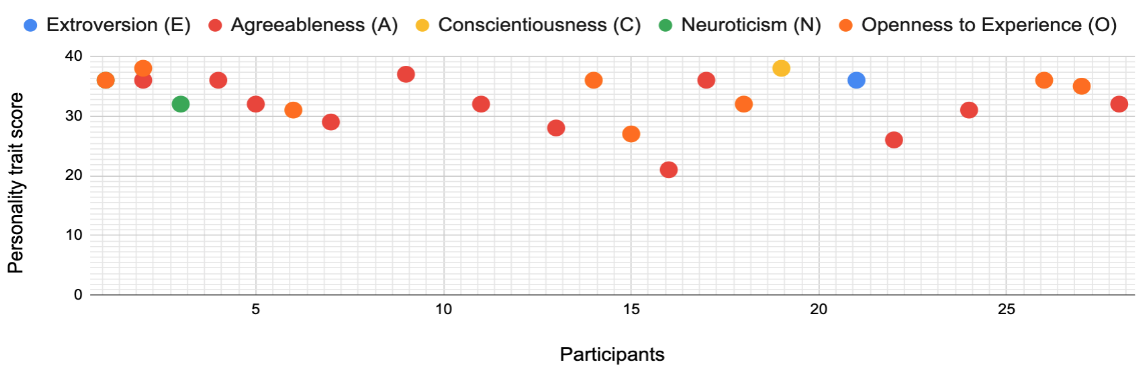


Figure 33: Dominant personality traits of the participants

Since 4 out of 24 participants in this study did not provide answers to the personality test it caused excluding 4 teams out of a total of 14 teams for the analysis.

The goal of this analysis is to identify if combinations of personality types in teams impact the individuals' performance. For this purpose, we create a 5D contingency matrix (A) to quantify unique combinations of personality types as shown in Figure 34.

$$A = \begin{array}{cc} & \begin{array}{ccccc} \mathbf{E} & \mathbf{A} & \mathbf{C} & \mathbf{N} & \mathbf{O} \end{array} \\ \begin{array}{c} \mathbf{E} \\ \mathbf{A} \\ \mathbf{C} \\ \mathbf{N} \\ \mathbf{O} \end{array} & \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Figure 34: Contingency matrix of the combination of personality types

To validate H4 we need to identify the correlation between the combination of personality types and performance. Each unique combination of personality type is coded with an integer value.   Figure 35 shows the mean of performance for each unique combination of personality type, as well as the code of each personality combination.

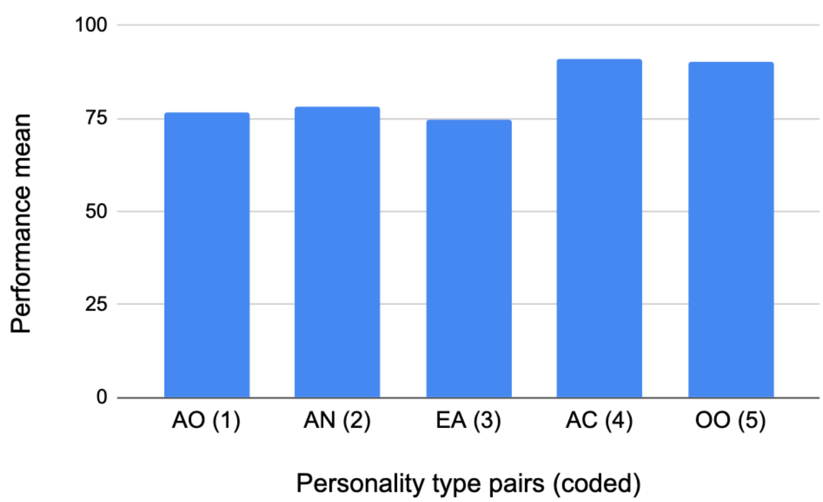


Figure 35: The combination of personality vs performance

By applying the Spearman's rank correlation coefficient, we calculated the r_s and p value for personality pairs and the performance of the individuals as well as the team. The result is presented in table 12.

The r_s values show a very weak positive correlation between personality types and the performance. The *p* value which is greater than .05 (confidence level) does not show a statistically significant correlation. Thus, in H4, the Null hypothesis cannot be rejected which means there is not a statically significant correlation between personality types and performance.

Table 12: Spearman's rank correlation coefficient test result

	p value	r_s
Personality and team performance	0.91	.04
Personality and Individual performance	0.98	.00

In the following section, we discuss the results and present more findings from data analysis.

5.4. Discussion

In this section, we present further analysis on students' speech. We identify the correlation of the neutral sentiments and the level of subjectivity in individuals' speech with their performance. We also conduct a thematic analysis of students' negative sentiments to identify the themes in which students expressed more negative emotions. Finally, we investigate if the combination of personality types impacts the team experience in terms of students' positive sentiments. The finding of these analyses can cue the development of new research questions and identifying ways to address them.

First is identifying the correlation between the frequency of neutral sentiments and students' performance. As discussed earlier in Chapter 4 by applying the VADER tool we extracted the sentiments in four classes of positive, neutral, negative, and compound. The calculated threshold for this study is zero meaning any vector with the compound value of zero is considered to be neutral.

Figure 36 plots the mean frequency of neutral sentiments vectors vs students' individual performance. The data points show a homogeneous pattern such that by decreasing the performance the frequency of neutral sentiments increases.

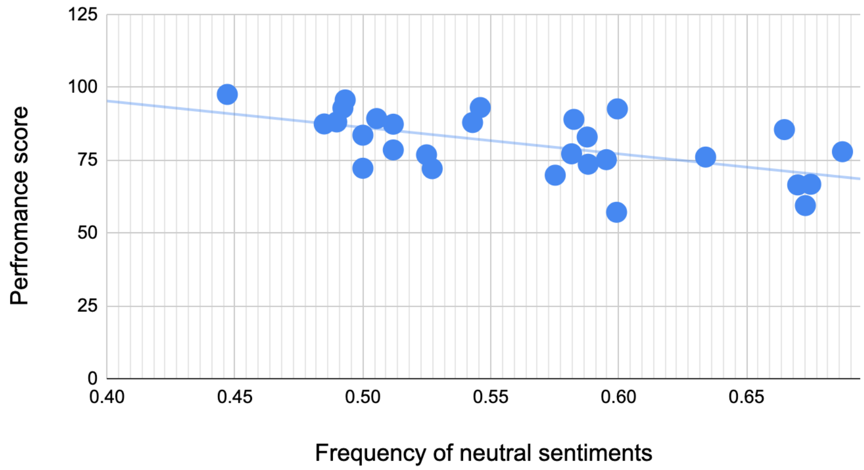


Figure 36: Regression plot of the frequency of neural sentiments vs performance score

By applying the Spearman's rank correlation coefficient test the calculated r_s value for the two variables is r_s = -0.61. This shows the association between variables is statistically significant. We conclude there is a strong negative correlation between the variables and there is a tendency for high-performance scores to go with a low frequency of neutral score and vice versa.

Next, we analyze the correlation between students' level of subjectivity in their speech and their performance. For extracting the subjectivity metric, we apply TextBlob which is a rule-based sentiment analysis tool that has the essential component for the basics of natural-language processing to analyze the subjectivity level in the speech [105]. 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 [106]. The subjectivity score of each participant is calculated with equation (4)

$$\text{Subjectivity score} = \sum_{x=1}^{n} Sx \qquad (4)$$

Where, Sx is the subjectivity score of each vector (0<=Sx<=1), and n is the total number of vectors in each dataset. Figure 37 plots the subjectivity score vs. performance score.

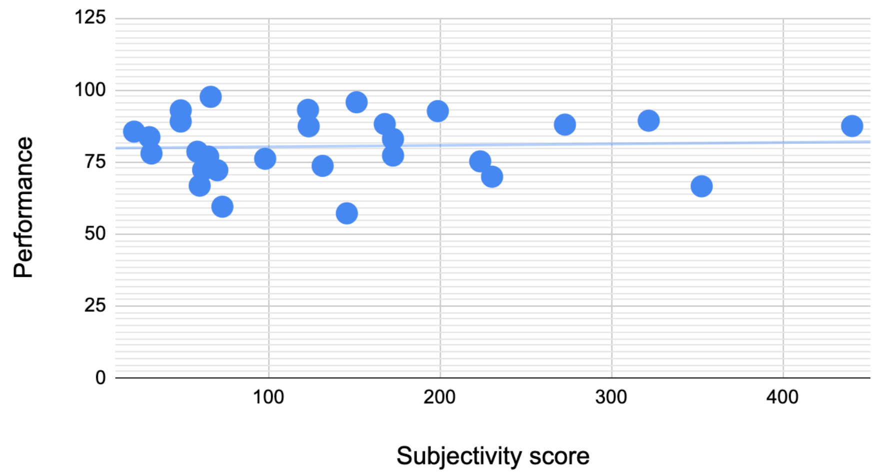


Figure 37. The regression plot of Subjectivity vs Performance Score

The result of the Spearman's rank correlation coefficient shows $r_s = -.02$, which means technically very weak negative correlation exists between the level of subjectivity and students' performance.

Next is the thematic analysis of negative sentiment vectors (Comp < 0). By running the wordcloud library in Python the wordcloud of the negative sentiment corpora is presented in Figure 38. It shows that students expressed more negative sentiments when they made mistakes and faced a problem.

For the purpose of this study, we have replaced the swear words to "shift" and "freak".

Figure 38. The word cloud of the negative sentiments

Figure 39. visualizes the frequency of the unigrams in the negative sentiment vectors. It shows the most frequent word token expressed in negative emotions is the word “wrong”.

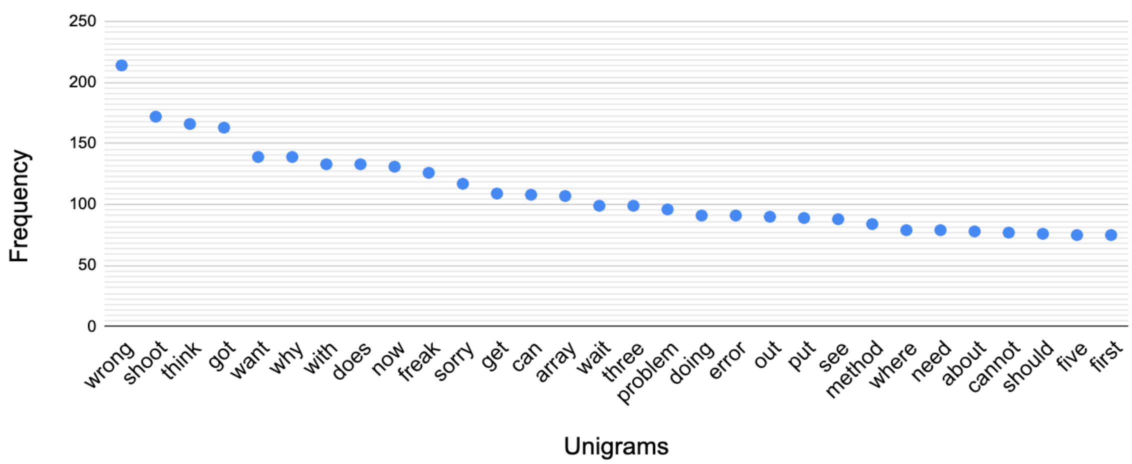


Figure 39: Frequency of negative sentiment unigrams

It's worth noting that we did not find any correlation between students’ negative sentiments and their performance in this study.

Finally, we analyze the correlation between the combination of personality types in teams and the positive sentiments (sum of positive compound scores). The result of Spearman’s rank correlation coefficient shows $r_s = .91$ which means there is a strong positive correlation between the combination of personality types and students’ positive sentiments in teams. Figure 40 plots the bar chart of positive sentiment scores in each team vs their personality types.

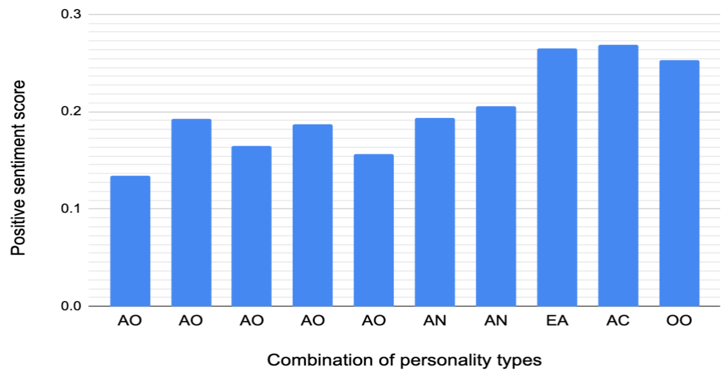


Figure 40: Positive sentiment scores vs. personality combinations

It is worth noting that since 4 participants did not provide a response to the personality test it caused the elimination of 4 teams from this analysis. In order to generate more concrete conclusions, more samples should be analyzed.

In the next chapter, we conclude this study by discussing the contribution of this study, takeaways and the path towards future work.

CHAPTER 6: CONCLUSION AND FUTURE WORK

In this chapter, we summarize the research question, methodology, and data analysis result. We discuss the limitations we faced in this research, the takeaways, and finally, we present the future research plan.

6.1. Summary

Teamwork and collaboration are important aspects of cognition and learning. However, evaluating the performance of students in low-stake teams is a challenge due to low grade contribution of teamwork which causes some students to undergo the load of teamwork more than their peers.

Research suggests attitudinal components are indicators of students' learning in educational settings. Measuring and quantifying attitude is mainly done in the form of surveys and self-reports which do not always provide accurate and consistent results. Another main drawback of self-report tool is the lack of commitment from participants to provide responses. One approach that has been practiced for measuring attitude constructs is applying text analysis methods on students' textual conversations on discussion forums or blogs. The drawback of these approaches is that they are conducted on asynchronous conversations and make it hard to assess students' affect in real-time situations. Studies show speech is the best source to capture the attitude and affective states. However, recording and analyzing speech is not cost-effective in terms of time, effort, and resources. Due to the environmental noise level and the distraction caused by setting up equipment this method of data collection has been less practical in educational settings.

In this research, we propose a novel approach to capture and operationalize students' affects in teams by recording their real-time speech and conducting sentiment analysis on them. Finding the correlation between the affective components and students' performance can help in developing models to predict students' performance at earlier stages of the semester and provide timely feedback to at-risk students.

In this study, we analyze three attitude components of affect, self-efficacy, and personality. The research question is to examine whether these constructs have any correlation with students' final grades as the performance metric. If the correlation exists, we claim some interventions can be made earlier in the course to predict students' performance and provide timely feedback to them based on their behaviors in teams and the attitudes they express.

The novelty of our research is capturing students' affective states based on their speech in the teams. This method has minimum distraction for students and does not require them to spend extra time in providing data. Furthermore, we operationalize self-efficacy and personality constructs by using standard self-report tools.

We experimented this method in a CS1 class during one semester with 28 participants. By implementing our text mining algorithms and using NLP methods we analyze the speech corpus to extract the frequency and intensity of students' emotions as well as diverse psychometric and language features. We asked students to self-report on Self-efficacy and Personality tools at the beginning of the semester.

The result of our analysis shows a strong positive correlation exists between students' positive emotions and their performance. It means the students who had higher scores in frequency and intensity in their positive sentiments earned higher grades in the course. On the other hand, the students who had more frequency in their neutral sentiments had lower performance scores in the course. However, we did not identify a statistically significant correlation between negative sentiments and students' performance.

The result of thematic analysis of the positive sentiments shows the subject areas and topics in which students showed positive emotions were the course-related keywords.

Analysis of the self-efficacy self-report tool does not reveal a statistically significant correlation with students' performance. We argue this can be due to limitations of self-report tools which is elaborated in this research such as lack of awareness or commitment to provide a precise answer.

The principal component analysis on the self-efficacy survey reveals the questions about 'learning and social skills' have more contribution in determining the target value which is performance rather than the 'computer science' questions. We argue the reason can lie in the fact that freshmen might not have much domain knowledge about computer science to rate their self-efficacy in the field.

Team formation has many challenges and includes diverse factors that can impact team experience. In this study, we analyze the combination of personality types in a team to see if collaboration between specific personality types can result in a better teamwork experience. The result of the analysis did not show a statistically significant correlation between the combination of personality types and individual performances in the team. However, our findings show a positive correlation between the combination of personality types and the intensity and frequency score of positive sentiments. It means certain combinations of personality types led to happier teamwork experience. It is worth mentioning that this result cannot be generalized science other than the mentioned shortcomings of self-reports tools, for the personality test 4 students did not participate which causes loss of data of four teams in the analysis.

In the personality analysis, out of 25 possible combinations we had 5 unique pairs of personality types. In the future, we need to have more diverse pairs of personalities to conduct a more in-depth analysis of this.

This study was conducted on a limited sample size of 28 participants. The results are based on the data collected and there needs to be more robust data collection to substantiate the conclusions from the data. Furthermore, the outcome of this study may not be generalized to other courses or the population of students. However, the early result encourages future research to collect more samples to develop generic conclusions.

One major contribution of this research is demonstrating that the novel method of recording students' in class, besides its many challenges, is practical and is an ideal way to observe their behavior and analyze students' affective states without causing distraction in the learning process. Another takeaway is that direct observation of students' behavior in teams is a better indicator of the attitude components than using self-report tools. The finding of the research also confirms that the team formation protocol impacts teamwork experience. This calls for implementing well-studied team formation strategies.

6.2. Limitations

The major limitation of this study was the data collection. The available data in the educational domain is very limited and since we analyzed human subjects it has specific challenges. Given the novelty of this research method (i.e. recording students in every class session), one of the challenges was the IRB approval process which was very time-consuming. We had so many iterations to come up with a study design that was aligned with the IRB protocol. Moreover, we were not able to collect data unless all participants signed and returned the consent forms, which caused delays in the data collection process.

The study design and team formation were other challenges that we faced. Since we formed teams at the beginning of the semester, after add/drop period some team formations had to be adjusted which caused the loss of some recorded sessions. Moreover, since we studied teams, the absence of one participant led to the elimination of that team in the recording of that class session.

Besides these challenges, we had technical limitations as well. Some students randomly pressed the stop button on the device during the teamwork which made the audio unavailable. Some TA's misplaced the microphone cords which caused the loss of data. For transcribing the speech, we used a human service for more accuracy. Due to some noise level and lack of voice quality some audios could not be transcribed. Finally, the process of providing recorders, ensuring they all are fully charged for every class session, the recording process in class, and unloading the audio from all recorders was very time consuming and required human resources. All these limitations led to the loss of some data in our study.

6.3. Future Work

Future research includes in-depth exploration of the role of peer communication in cognitive processes by collecting more samples to analyze data and come up with generic conclusions about the correlation of affective states with performance. In the current study, we identified the principal components from multiple sentiment classes that determine the performance. We will fit the principal components into a machine-learning algorithm to predict students' performance as we run the algorithms on the new samples.

The next step would be applying interventions based on the prediction result. One possible intervention would be the development of a system that provides emotional feedback to instructors and students as they work in teams. This would help students adjust their behavior and interaction with peers. The emotional feedback system can cue the instructor to provide learning opportunities and timely feedback to students.

By collecting more samples, we consider studying the impact of personality combinations in teamwork experience and students' happiness. The findings will help in incorporating the identified correlations into team formation strategies to optimize team performance and generating algorithms for team formation.

We will investigate the possibility of developing models to automate the transcriptions of the audio so that we can provide real-time analysis as students work in class. This would provide real-time feedback to instructors to observe the climate of different teams and provide help when needed. Tone analysis directly on the audio files is another research path we will consider in the future work. Finally, we will apply AI to design intelligent systems that support our understanding of team behavior beyond educational settings by studying teams in industry.

In summary, the result of this research has the potential to benefit both students and instructors. Instructors can adjust team settings and also identify at-risk students in earlier stages of the semester and help them to collaborate more effectively and learn better by applying cognitive interventions. The real-time analysis platform in particular has the potential to provide effective feedback and help improve students' learning and social skills.

REFERENCES

[1] Salomon, G., & Perkins, D. N. (1998). Chapter 1: Individual and social aspects of learning. Review of research in education, 23(1), 1-24.

[2] Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J. D. (2016). A tough nut to crack: Measuring collaborative problem solving. In Handbook of research on technology tools for real-world skill development (pp. 344-359). IGI Global.

[3] Salas, E., Reyes, D. L., & Woods, A. L. (2017). The assessment of team performance: Observations and needs. In Innovative assessment of collaboration (pp. 21-36). Springer, Cham.

[4] Peterson, E., Mitchell, T. R., Thompson, L., & Burr, R. (2000). Collective efficacy and aspects of shared mental models as predictors of performance over time in work groups. Group Processes & Intergroup Relations, 3(3), 296-316.

[5] Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. Annu. Rev. Psychol., 56, 517-543.

[6] Kozlowski, S. W. J., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), Handbook of psychology: Industrial and organizational psychology (Vol. 12, pp. 333–375). London, England: Wiley.

[7] Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. Psychological science in the public interest, 7(3), 77-124.

[8] Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. Journal of management, 34(3), 410-476.

[9] DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The cognitive underpinnings of effective teamwork: a meta-analysis. Journal of applied psychology, 95(1), 32.

[10] Hackman, J. R. (1990). Groups that work and those that don't (No. E10 H123). Jossey-Bass.

[11] Mcgrath, J. E., Arrow, H., Gruenfeld, D. H., Hollingshead, A. B., & O'Connor, K. M. (1993). Groups, tasks, and technology: The effects of experience and change. Small group research, 24(3), 406-420.

[12] Wilson, K., & Narayan, A. (2016). Relationships among individual task self-efficacy, self-regulated learning strategy use and academic performance in a computer-supported collaborative learning environment. Educational Psychology, 36(2), 236-253.

[13] Lishinski, A., Yadav, A., Good, J., & Enbody, R. (2016, August). Learning to program: Gender differences and interactive effects of students' motivation, goals, and self-efficacy on

performance. In Proceedings of the 2016 ACM Conference on International Computing Education Research (pp. 211-220).

[14] Mazni, O., Syed-Abdullah, S. L., & Hussin, N. M. (2010, December). Analyzing personality types to predict team performance. In 2010 International Conference on Science and Social Research (CSSR 2010) (pp. 624-628). IEEE.

[15] Latulipe, C., Long, N. B., & Seminario, C. E. (2015, February). Structuring flipped classes with lightweight teams and gamification. In Proceedings of the 46th ACM Technical Symposium on Computer Science Education (pp. 392-397).

[16] Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T., & Haynes, J. D. (2011). Flow of affective information between communicating brains. Neuroimage, 54(1), 439-446.

[17] Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. Proceedings of the National Academy of Sciences, 107(20), 9388-9393.

[18] Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance, 29(2), 326.

[19] Stevens, R., Galloway, T., Lamb, J., Steed, R., & Lamb, C. (2017). Linking team neurodynamic organizations with observational ratings of team performance. In Innovative assessment of collaboration (pp. 315-330). Springer, Cham.

[20] Arguedas, M., Daradoumis, A., & Xhafa Xhafa, F. (2016). Analyzing how emotion awareness influences students' motivation, engagement, self-regulation and learning outcome. Educational technology and society, 19(2), 87-103.

[21] Dehbozorgi, N., MacNeil, S., Maher, M. L., & Dorodchi, M. (2018, October). A comparison of lecture-based and active learning design patterns in CS education. In 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1-8). IEEE.

[22] Bonwell, C. C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183.

[23] Prince, M. (2004). Does active learning work? A review of the research. Journal of engineering education, 93(3), 223-231.

[24] Köppe, C., & Portier, M. (2014, July). Lecture design patterns: improving the beginning of a lecture. In Proceedings of the 19th European Conference on Pattern Languages of Programs (pp. 1-12).

[25] Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. Science, 323(5910), 122-124.

[26] Millis, B. J., & Cottell Jr, P. G. (1997). Cooperative Learning for Higher Education Faculty. Series on Higher Education. Oryx Press, PO Box 33889, Phoenix, AZ 85067-3889.

[27] Feden, P. D., & Vogel, R. M. (2003). Methods of teaching: Applying cognitive science to promote student learning. McGraw-Hill Humanities, Social Sciences & World Languages.

[28] Narayanan, N. H., Hundhausen, C., Hendrix, D., & Crosby, M. (2012, February). Transforming the CS classroom with studio-based learning. In Proceedings of the 43rd ACM technical symposium on Computer Science Education (pp. 165-166).

[29] Dehbozorgi, N. (2017, August). Active learning design patterns for CS education. In Proceedings of the 2017 ACM Conference on International Computing Education Research (pp. 291-292).

[30] "Active Learning Patterns." THE CONNECTED LEARNER, UNC Charlotte - College of Computing and Informatics, 3 Nov. 2019, innovationsteched.com/active-learning-design-patterns/.

[31] Michaelsen, L. K., & Sweet, M. (2008). The essential elements of team-based learning. New directions for teaching and learning, 2008(116), 7-27.

[32] Adams, S. G. (2003). Building successful student teams in the engineering classroom. Journal of STEM Education: Innovations and Research, 4(3).

[33] LeJeune, N. (2003). Critical components for successful collaborative learning in CS1. Journal of Computing Sciences in Colleges, 19(1), 275-285.

[34] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual review of sociology, 27(1), 415-444.

[35] Layton, R. A., Loughry, M. L., Ohland, M. W., & Ricco, G. D. (2010). Design and validation of a web-based system for assigning members to teams using instructor-specified criteria. Advances in Engineering Education, 2(1), n1.

[36] Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). Active learning: Cooperation in the College Classroom", Interaction Book Company, Edina, MN.

[37] Mennecke, B., & Bradley, J. (1998). Making project groups work: The impact of structuring group roles on the performance and perception of information systems project teams. Journal of Computer Information Systems, 39(1), 30-36.

[38] Cruz, S. S., da Silva, F. Q., Monteiro, C. V., Santos, P., Rossilei, I., & dos Santos, M. T. (2011, April). Personality in software engineering: Preliminary findings from a systematic

literature review. In 15th annual conference on Evaluation & assessment in software engineering (EASE 2011) (pp. 1-10). IET.

[39] Heidy Maldonado, Scott R. Klemmer, Roy D. Pea. 2009. When is collaborating with friends a good idea? insights from design education. In Proceedings of the 9th international conference on Computer supported collaborative learning - Volume 1 (CSCL'09), International Society of the Learning Sciences 227-231.

[40] Barker, L. J., McDowell, C., & Kalahar, K. (2009). Exploring factors that influence computer science introductory course students to persist in the major. ACM Sigcse Bulletin, 41(1), 153-157.

[41] DeBerard, M. S., Spielmans, G. I., & Julka, D. L. (2004). Predictors of academic achievement and retention among college freshmen: A longitudinal study. College student journal, 38(1), 66-81.

[42] Smith, B. L., & MacGregor, J. T. (1992). What is collaborative learning.

[43] Bruffee, K. A. (1995). Sharing our toys: Cooperative learning versus collaborative learning. Change: The Magazine of Higher Learning, 27(1), 12-18.

[44] Panitz, T. (1999). Collaborative versus Cooperative Learning: A Comparison of the Two Concepts Which Will Help Us Understand the Underlying Nature of Interactive Learning.

[45] Stahl, R. J. (1994). The Essential Elements of Cooperative Learning in the Classroom. ERIC Digest.

[46] Dickinson, T. L., & McIntyre, R. M. (1997). A conceptual framework for teamwork measurement. Team performance assessment and measurement, 19-43.

[47] Hayes, J. H., Lethbridge, T. C., & Port, D. (2003, May). Evaluating individual contribution toward group software engineering projects. In 25th International Conference on Software Engineering, 2003. Proceedings. (pp. 622-627). IEEE.

[48] Wilkins, D. E., & Lawhead, P. B. (2000). Evaluating individuals in team projects. ACM SIGCSE Bulletin, 32(1), 172-175.

[49] Forrester, W. R., & Tashchian, A. (2013). Effects of personality on conflict resolution in student teams: A structural equation modeling approach. Journal of College Teaching & Learning (TLC), 10(1), 39-46.

[50] Peslak, A. R. (2006, April). The impact of personality on information technology team projects. In Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future (pp. 273-279).

[51] Shen, S. T., Prior, S. D., White, A. S., & Karamanoglu, M. (2007). Using personality type differences to form engineering design teams. Engineering education, 2(2), 54-66.

[52] Stevens, R., Galloway, T., Lamb, J., Steed, R., & Lamb, C. (2017). Linking team neurodynamic organizations with observational ratings of team performance. In Innovative assessment of collaboration (pp. 315-330). Springer, Cham.

[53] Salas, E., Burke, C. S., & Fowlkes, J. E. (2006). Measuring team performance "in the wild": Challenges and tips. Performance measurement: Current perspectives and future challenges, 245-272.

[54] Driskell, J. E., Salas, E., & Hughes, S. (2010). Collective orientation and team performance: Development of an individual differences measure. Human factors, 52(2), 316-328.

[55] Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T., & Haynes, J. D. (2011). Flow of affective information between communicating brains. Neuroimage, 54(1), 439-446.

[56] Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L., & Keysers, C. (2010). Mapping the information flow from one brain to another during gestural communication. Proceedings of the National Academy of Sciences, 107(20), 9388-9393.

[57] Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance, 29(2), 326.

[58] Laguador, J. M. (2013). Developing students' attitude leading towards a life-changing career. Educational Research International, 1(3), 28-33.

[59] Phye, G. D. (2007). Emotion in education (Vol. 10). P. A. Schutz, & R. Pekrun (Eds.). San Diego, CA: Academic Press.

[60] Bacay, T. E., Dotong, C. I., & Laguador, J. M. (2015). Attitude of Marine Engineering Students on Some School-Related Factors and their Academic Performance in Electro Technology 1 and 2. Studies in Social Sciences and Humanities, 2(4), 239-249.

[61] Cohn, E., Cohn, S., Balch, D. C., & Bradley Jr, J. (2004). The relation between student attitudes toward graphs and performance in economics. The American Economist, 48(2), 41-52.

[62] Arguedas, M., Daradoumis, T., & Xhafa, F. (2014, July). Towards an emotion labeling model to detect emotions in educational discourse. In 2014 Eighth International Conference on Complex, Intelligent and Software Intensive Systems (pp. 72-78). IEEE.

[63] Munezero, M., Montero, C. S., Mozgovoy, M., & Sutinen, E. (2013, November). Exploiting sentiment analysis to track emotions in students' learning diaries. In Proceedings of the 13th Koli Calling International Conference on Computing Education Research (pp. 145-152).

[64] Maras, P., & Kutnick, P. (1999). Emotional and behavioural difficulties in schools: Consideration of relationships between theory and practice. Social Psychology of Education, 3(3), 135-153.

[65] Araújo-Simões, A. C., & Guedes-Gondim, S. M. (2016). Performance and affects in group problem-solving. Revista de Psicología del Trabajo y de las Organizaciones, 32(1), 47-54.

[66] Hórreo, V. S., & Carro, R. M. (2007, September). Studying the impact of personality and group formation on learner performance. In International Conference on Collaboration and Technology (pp. 287-294). Springer, Berlin, Heidelberg.

[67] Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). Contemporary educational psychology, 36(1), 36-48.

[69] Mizokawa, D. T., & Hansen-Krening, N. (2000). The ABCs of attitudes toward reading: Inquiring about the reader's response. Journal of Adolescent & Adult Literacy, 44(1), 72-79.

[70] McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization. Handbook of research on mathematics teaching and learning, 1, 575-596.

[71] Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. International journal of human-computer studies, 59(1-2), 1-32.

[72] Peixoto, F., Mata, L., Monteiro, V., Sanches, C., & Pekrun, R. (2015). The achievement emotions questionnaire: Validation for pre-adolescent students. European Journal of Developmental Psychology, 12(4), 472-481.

[74] Rynearson, A. M., & Rynearson, L. (2018, October). Embedded Affective Assessment. In 2018 IEEE Frontiers in Education Conference (FIE) (pp. 1-4). IEEE.

[75] Bartimote-Aufflick, K., Bridgeman, A., Walker, R., Sharma, M., & Smith, L. (2016). The study, evaluation, and improvement of university student self-efficacy. Studies in Higher Education, 41(11), 1918-1942.

[76] DiBenedetto, M. K., & Bembenutty, H. (2013). Within the pipeline: Self-regulated learning, self-efficacy, and socialization among college students in science courses. Learning and Individual Differences, 23, 218-224.

[77] Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. Psychological bulletin, 138(2), 353.

[78] Unfried, A., Faber, M., Stanhope, D. S., & Wiebe, E. (2015). The development and validation of a measure of student attitudes toward science, technology, engineering, and math (S-STEM). Journal of Psychoeducational Assessment, 33(7), 622-639.

[79] Bishop-Clark, C. (1995). Cognitive style, personality, and computer programming. Computers in human behavior, 11(2), 241-260.

[80] Kim, J., Lee, A., & Ryu, H. (2013). Personality and its effects on learning performance: Design guidelines for an adaptive e-learning system based on a user model. International Journal of Industrial Ergonomics, 43(5), 450-461.

[81] Mazni, O., Syed-Abdullah, S. L., & Hussin, N. M. (2010, December). Analyzing personality types to predict team performance. In 2010 International Conference on Science and Social Research (CSSR 2010) (pp. 624-628). IEEE.

[82] Arnau, R. C., Thompson, B., & Rosen, D. H. (1999). Alternative measures of Jungian personality constructs. Measurement and evaluation in counseling and development, 32(2), 90-104.

[83] Busato, V. V., Prins, F. J., Elshout, J. J., & Hamaker, C. (1998). The relation between learning styles, the Big Five personality traits and achievement motivation in higher education. Personality and individual differences, 26(1), 129-140.

[84] Furnham, A., Monsen, J., & Ahmetoglu, G. (2009). Typical intellectual engagement, Big Five personality traits, approaches to learning and cognitive ability predictors of academic performance. British Journal of Educational Psychology, 79(4), 769-782.

[89] Tarmazdi, H., Vivian, R., Szabo, C., Falkner, K., & Falkner, N. (2015, June). Using learning analytics to visualise computer science teamwork. In Proceedings of the 2015 ACM Conference on Innovation and technology in computer science education (pp. 165-170).

[68] Durndell, A., & Haag, Z. (2002). Computer self efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. Computers in human behavior, 18(5), 521-535.

[85] Bonta, V., & Janardhan, N. K. N. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. Asian Journal of Computer Science and Technology, 8(S2), 1-6.

[86] Ahmed, T., Bosu, A., Iqbal, A., & Rahimi, S. (2017, October). SentiCR: a customized sentiment analysis tool for code review interactions. In 2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE) (pp. 106-111). IEEE.

[87] Hood, K., & Kuiper, P. K. (2018, January). Improving student surveys with natural language processing. In 2018 Second IEEE International Conference on Robotic Computing (IRC) (pp. 383-386). IEEE.

[88] Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Eighth international AAAI conference on weblogs and social media.

[89] Amin, A., Hossain, I., Akther, A., & Alam, K. M. (2019, February). Bengali vader: A sentiment analysis approach using modified vader. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE.

[90] Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

[91] Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

[92] Hess, A. S., & Hess, J. R. (2018). Principal component analysis. Transfusion, 58(7), 1580.

[93] Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models, volume 11 of Foundations and Trends in Information Retrieval.

[94] Chen, Q., Yao, L., & Yang, J. (2016, July). Short text classification based on LDA topic model. In 2016 International Conference on Audio, Language and Image Processing (ICALIP) (pp. 749-753). IEEE.

[95] Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013, July). Improving lda topic models for microblogs via tweet pooling and automatic labeling. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (pp. 889-892).

[96] Han, B., Cook, P., & Baldwin, T. (2012, July). Automatically constructing a normalisation dictionary for microblogs. In Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (pp. 421-432).

[97] Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010, February). Twitterrank: finding topic-sensitive influential twitterers. In Proceedings of the third ACM international conference on Web search and data mining (pp. 261-270).

[98] Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88).

[99] Erkut, S., & Marx, F. (2005). 4 Schools for WIE. Evaluation Report. Wellesley Centers for Women.

[100] Friday Institute for Educational Innovation (2012). Student Attitudes toward STEM Survey-Middle and High School Students, Raleigh, NC: Author.

[101] Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones geographicae, 30(2), 87-93.

[102] Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. Malawi medical journal, 24(3), 69-71.

[103] "Learn: Machine Learning in Python - Scikit-Learn 0.16.1 Documentation." Scikit, scikit-learn.org/.

[104] Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. Advances in health sciences education, 15(5), 625-632.

[105] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for twitter accounts. Mathematical and Computational Applications, 23(1), 11.

[106] Loria, S. (2018). textblob Documentation (pp. 1-73). Technical report.

[107] Dehbozorgi N., Maher M.L., and Dorodchi M., (2020). Sentiment Analysis on Verbal Collaborative Conversations in Active Learning as Early Predictor of Performance. In 2020 IEEE Frontiers in Education Conference (FIE). IEEE

APPENDIX A: SELF-EFFICACY SURVEY

"Student Attitudes Toward STEM (S-STEM) Survey"

In the questions below, for each statement mark how much you agree with on the scale 1-5, where 1=Strongly disagree, 2= disagree, 3=neutral, 4= agree and 5= Strongly agree

CS-1: I am sure of myself when I do computer science.

CS-2: I would consider a career in computer science.

CS-3: I expect to use computer science when I get out of College.

CS-4: Knowing computer science will help me earn a living.

CS-5: I will need computer science for my future work.

CS-6: I know I can do well in computer science.

CS-7: Computer science will be important to me in my life's work.

CS-8: I can handle most subjects well, but I cannot do a good job with computer science.

CS-9: I am sure I could do advanced work in computer science

L&S-1: I am confident I can lead others to accomplish a goal.

L&S-2: I am confident I can encourage others to do their best.

L&S-3: I am confident I can produce high-quality work.

L&S-4: I am confident I can respect the differences of my peers.

L&S-5: I am confident I can help my peers.

L&S-6: I am confident I can include others' perspectives when making decisions.

L&S-7: I am confident I can make changes when things do not go as planned.

L&S-8: I am confident I can set my own learning goals.

L&S-9: I am confident I can manage my time wisely when working on my own.

L&S-10: When I have many assignments, I can choose which ones need to be done first.

L&S-11: I am confident I can work well with students from different backgrounds.

A-1: How well do you expect to do in this course?

A-2: In the future do you plan to take advanced computer science classes?

A-3: What best describes you?

APPENDIX B: PERSONALITY TYPE SURVEY “BIG FIVE PERSONALITY TRAITS”

Introduction

This is a personality test, it will help you understand why you act the way that you do and how your personality is structured. Please follow the instructions below, scoring and results are on the next page.

Instructions

In the table below, for each statement 1-50 mark how much you agree with on the scale 1-5, where 1=disagree, 2=slightly disagree, 3=neutral, 4=slightly agree and 5=agree, in the box to the left of it.

Test

Rating	I....	Rating	I.....
	1. Am the life of the party.		26. Have little to say.
	2. Feel little concern for others.		27. Have a soft heart.
	3. Am always prepared.		28. Often forget to put things back in their proper place.
	4. Get stressed out easily.		29. Get upset easily.
	5. Have a rich vocabulary.		30. Do not have a good imagination.
	6. Don't talk a lot.		31. Talk to a lot of different people at parties.
	7. Am interested in people.		32. Am not really interested in others.
	8. Leave my belongings around.		33. Like order.
	9. Am relaxed most of the time.		34. Change my mood a lot.
	10. Have difficulty understanding abstract ideas.		35. Am quick to understand things.
	11. Feel comfortable around people.		36. Don't like to draw attention to myself.
	12. Insult people.		37. Take time out for others.
	13. Pay attention to details.		38. Shirk my duties.
	14. Worry about things.		39. Have frequent mood swings.
	15. Have a vivid imagination.		40. Use difficult words.
	16. Keep in the background.		41. Don't mind being the center of attention.
	17. Sympathize with others' feelings.		42. Feel others' emotions.
	18. Make a mess of things.		43. Follow a schedule.
	19. Seldom feel blue.		44. Get irritated easily.
	20. Am not interested in abstract ideas.		45. Spend time reflecting on things.
	21. Start conversations.		46. Am quiet around strangers.
	22. Am not interested in other people's problems.		47. Make people feel at ease.
	23. Get chores done right away.		48. Am exacting in my work.
	24. Am easily disturbed.		49. Often feel blue.
	25. Have excellent ideas.		50. Am full of ideas.

APPEMDIX C: INTER TOPIC DISTANCE MAP

The estimated term frequency within 10 topics based on LDA topic modeling.

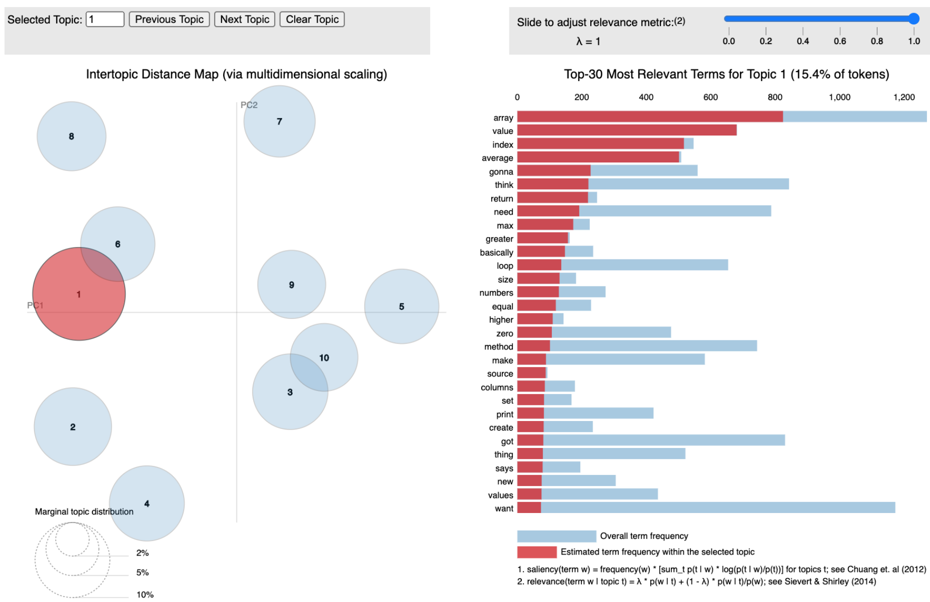


Figure C1: Estimated term frequency within topic 1

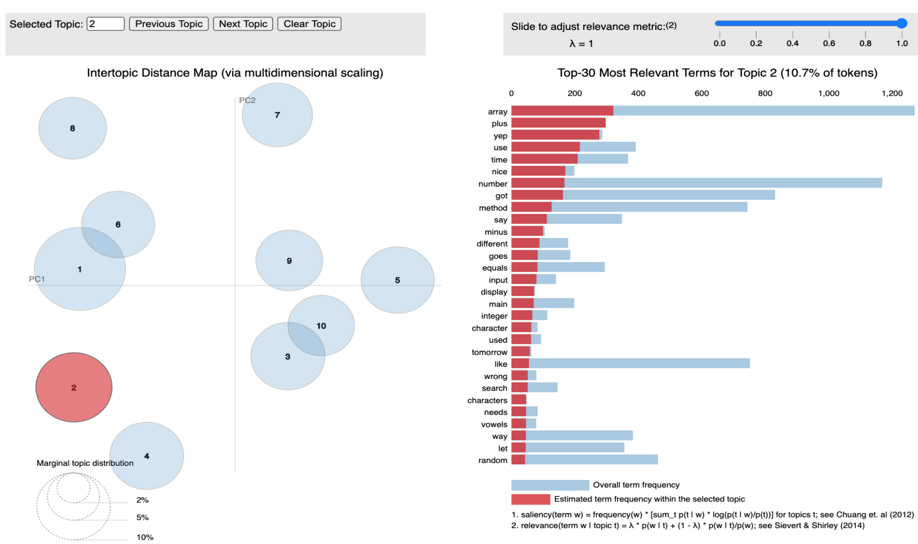


Figure C2: Estimated term frequency within topic 2

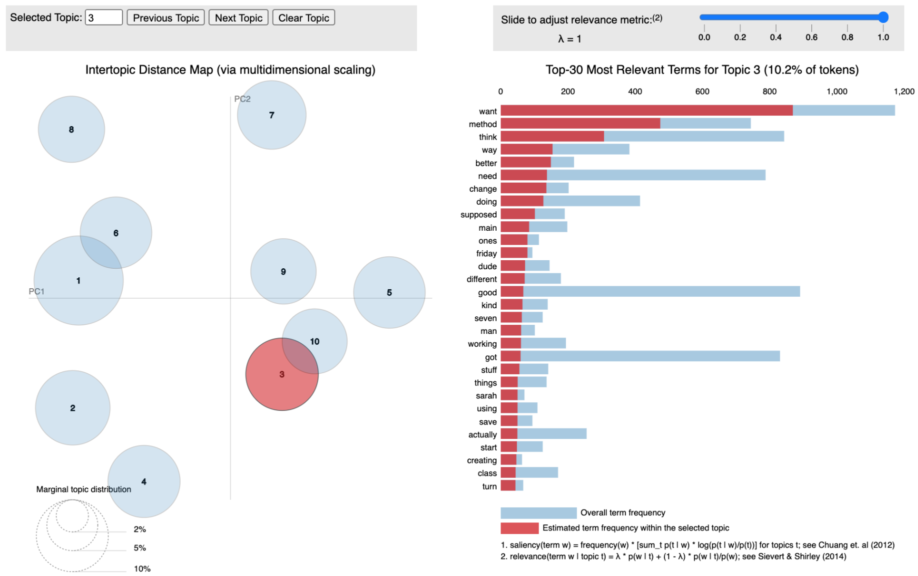


Figure C3: Estimated term frequency within topic 3

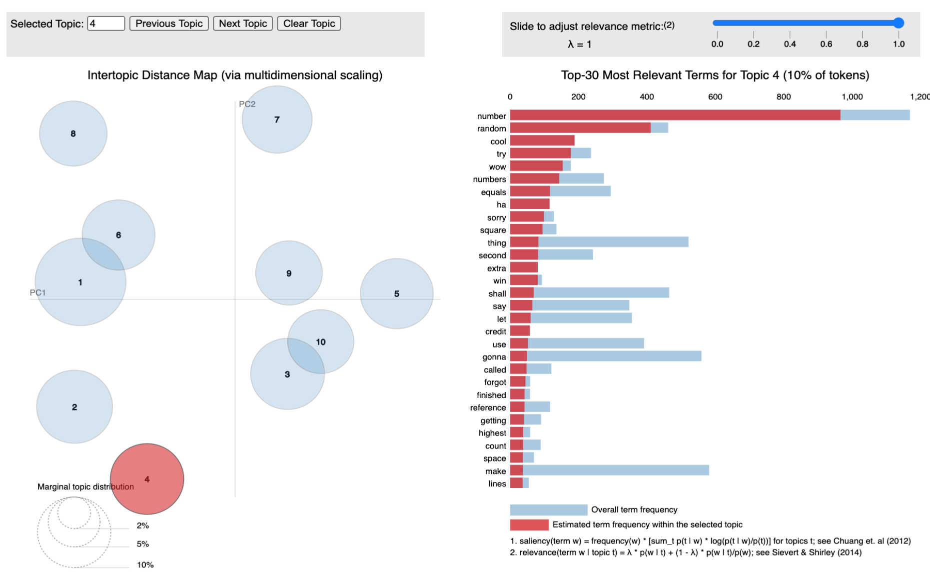


Figure C4: Estimated term frequency within topic 4

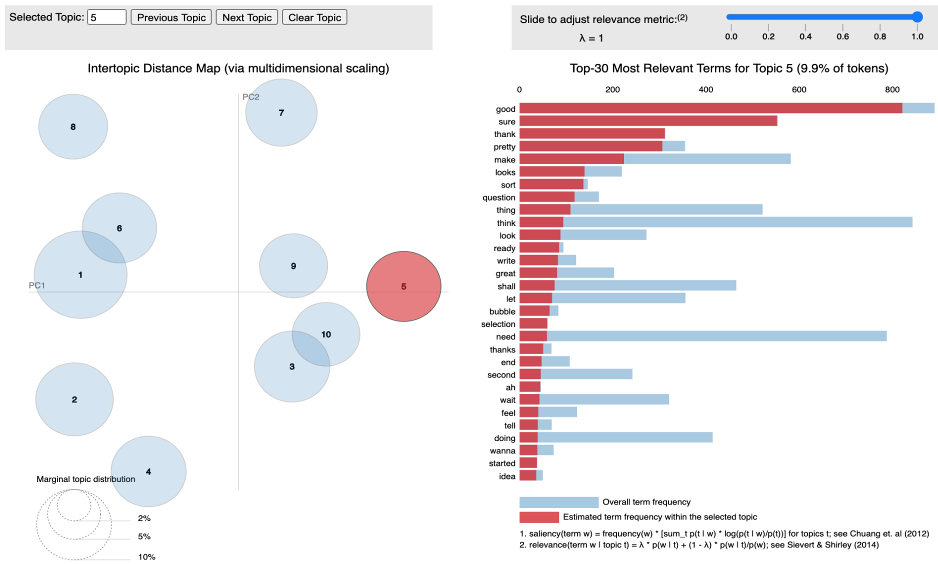


Figure C5: Estimated term frequency within topic 5

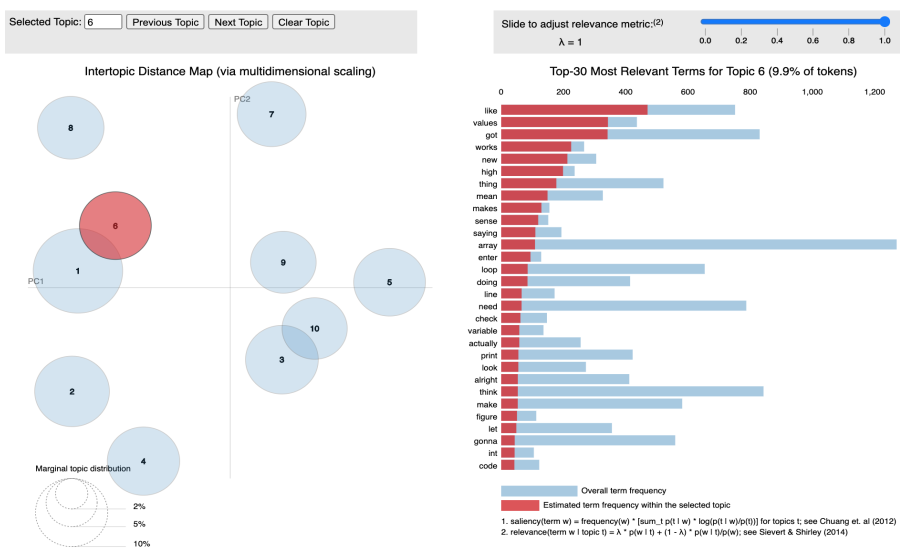


Figure C6: Estimated term frequency within topic 6

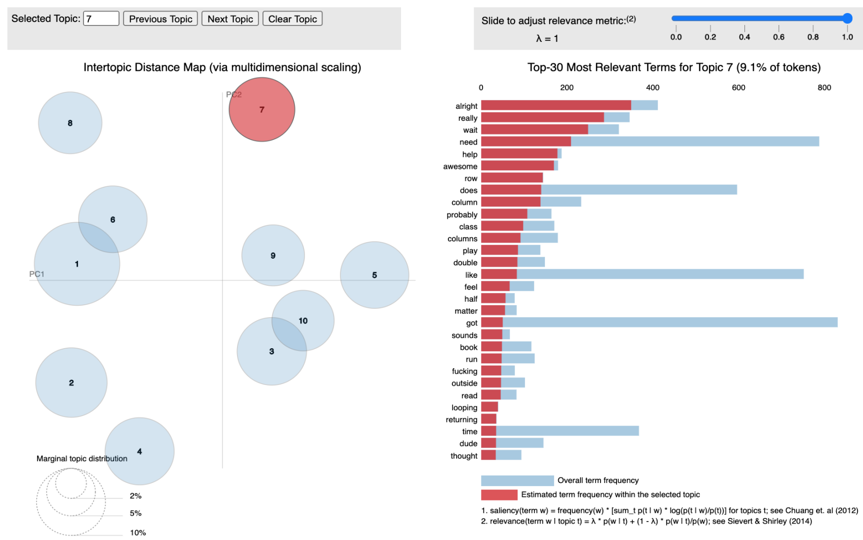


Figure C7: Estimated term frequency within topic 7

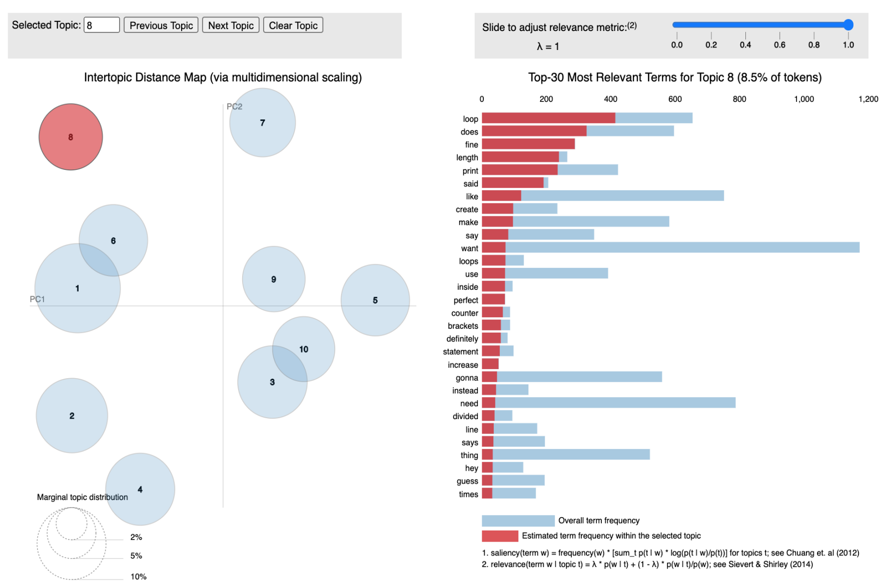


Figure C8: Estimated term frequency within topic 8

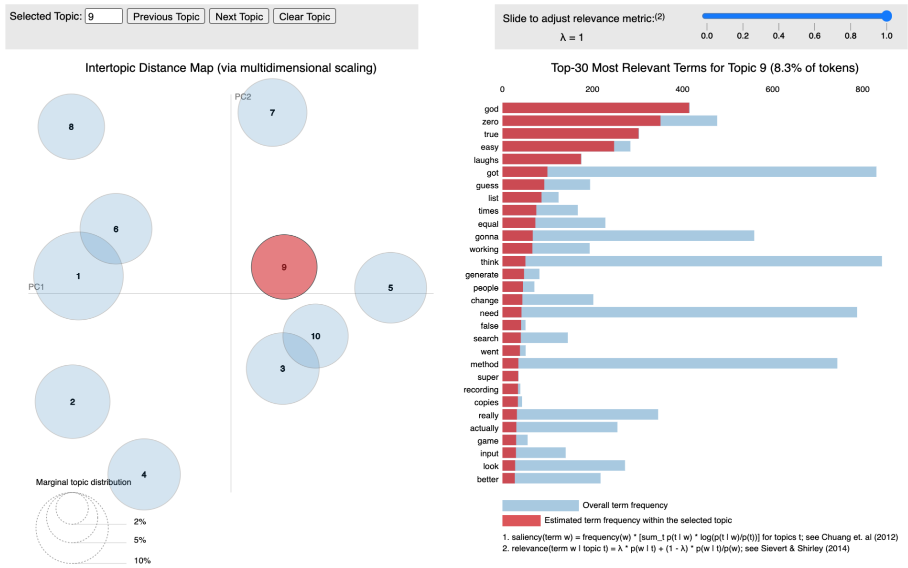


Figure C9: Estimated term frequency within topic 9

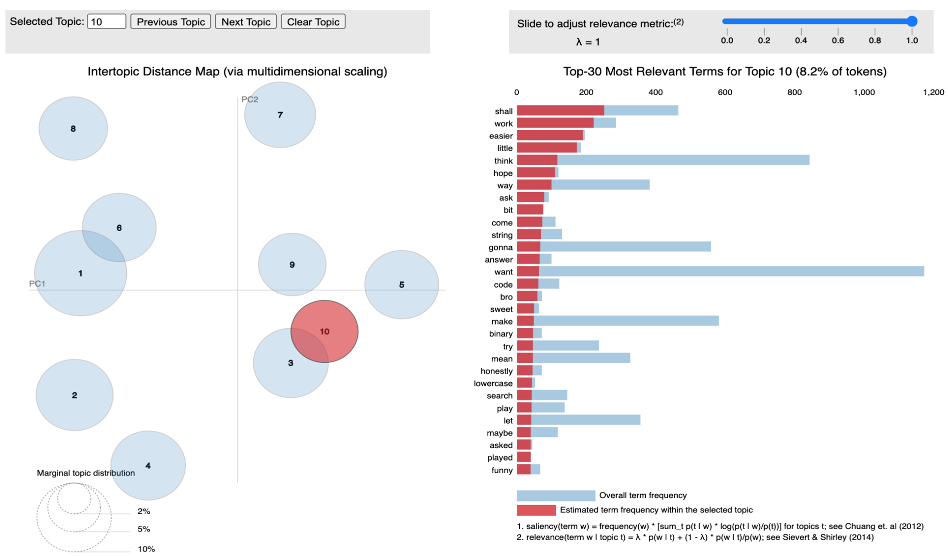


Figure C10: Estimated term frequency within topic 10