Working on weekends and cursing increases performance – measuring team success from WhatsApp messages

Hannah Apel[a], Lukas Dimdik[b], Matthias Wlcek[c], Alec Vayloyan[d], Rodrigo Gonzalez Alonso[e], Juan Garbajosa[f], Peter A. Gloor[g]

[a] University of Bamberg, hannah.apel@stud.uni-bamberg.de

[b] University of Bamberg, lukas-george.dimdik@stud.uni-bamberg.de

[c] University of Cologne, mwlcek@smail.uni-koeln.de

[d] HSLU Lucerne University of Applied Sciences and Arts, alec.vayloyan@stud.hslu.ch

[e] HSLU Lucerne University of Applied Sciences and Arts, rodrigo.gonzalesalonso@stud.hslu.ch

[f] Universidad Politécnica de Madrid, juan.garbajosa@upm.es

[g] MIT System Design and Management, pgloor@mit.edu

Abstract This study investigates the relationship between collaborative communication patterns and academic performance using WhatsApp group chat data from student teams working on university projects. By applying Natural Language Processing (NLP) and statistical methods, we analyze features such as sentiment, participation equality, dialogue acts, and topic modeling to assess their predictive power for final grades. Our findings indicate that communication styles, emotional tone, and structural engagement play a significant role in team performance.

1 Introduction

In an increasingly digitalized academic environment, group work has become a fundamental part of higher education (Laurillard, 2012, p.3). Digital communication platforms, such as WhatsApp, facilitate interaction among students but also introduce new challenges regarding collaboration quality and group dynamics (Hrastinski, 2008). The COVID-19 pandemic has underscored the indispensable relevance of digital communication in academia, as institutions rapidly transitioned to online platforms to maintain educational continuity (Eming & Philipowski, 2022).

While effective communication is widely recognized as a key factor in successful teamwork, research on how specific communication patterns in digital environments impact academic performance remains scarce (Amelia & Balqis, 2023). Roessler and Gloor (2020) highlight the need for empirical studies examining the nuances of digital interactions and their correlation with learning outcomes. Prior research suggests that aspects such as participation frequency, sentiment expression, and language style play a crucial role in shaping the effectiveness of digital teamwork (Davidson et al., 2017; Zhang et al., 2013). However, there is still a gap in understanding how these factors influence academic performance in digitally mediated collaborative learning environments (Zhu, 2012). This study seeks to bridge this gap by investigating the extent to which students' grades in group projects can be predicted based on their communication behavior in WhatsApp. By leveraging computational text analysis techniques, such as sentiment analysis and topic modeling, this research aims to identify key linguistic and interactional factors that correlate with academic success (Kim et al., 2012). Sentiment analysis has been widely used to assess emotional tone in digital communication and its effect on collaboration (Hutto & Gilbert, 2014), while topic modeling provides insights into thematic coherence and focus (Gloor, 2017, p.170). Additionally, conversational flow and engagement balance, measured through the Gini coefficient, have been found to influence the effectiveness of teamwork (van Mierlo et al., 2016).

This study builds on previous research in digital communication and team performance by integrating methods from computational linguistics, social network analysis, and machine learning (Hadyaoui & Cheniti-Belcadhi, 2023). Predictive modeling techniques, including feature selection and performance evaluation, are employed to identify key linguistic and interactional patterns associated with academic success (James et al., 2013, p.59-75; Hadyaoui & Cheniti-Belcadhi, 2023). Additionally, theoretical models of group development, such as Tuckman's (1965) framework, provide a foundation for interpreting communication structures in student collaboration.

To address the primary research question *To what extent can students' grades in group projects be predicted based on their communication in WhatsApp?* this study examines sentiment, coherence, participation balance, and message frequency as potential predictors of academic performance (van Mierlo et al., 2016). A structured analysis of linguistic and behavioral features is conducted, with results evaluated through statistical and machine learning techniques.

This study contributes to both the academic and practical discourse on digital teamwork by offering insights into how communication structures influence learning outcomes. The findings may provide valuable implications for educators and students alike, helping to optimize digital group work and foster more effective collaboration strategies. Understanding the key predictors of academic performance in digital communication can inform instructional design and collaborative learning frameworks in online and hybrid education settings.

This paper is structured as follows: Section 2 provides an overview of important literature, examining prior research on digital communication and its impact on

academic performance. Following this, section 3 outlines the methodological framework, detailing the data collection process and analytical techniques applied. The key findings derived from the analysis are then presented in section 4, offering insights into the relationship between communication patterns and student success. These results are subsequently discussed in section 5, where they are contextualized within existing research, and their implications are explored. Section 6 then addresses the study's limitations and identifies potential directions for future investigations. Lastly, section 7 synthesizes the main contributions of this research, concluding with reflections on its broader significance for digital collaboration and educational practice.

2 Related Work

Understanding communication patterns within teams and their impact on performance has been a central focus of organizational and academic research. Several key theoretical frameworks and empirical studies provide essential foundations for analyzing team dynamics through communication data. This section reviews relevant literature on group development, communication metrics, and the application of Natural Language Processing (NLP) and Machine Learning (ML) in predicting team success, highlighting their relevance to our study on WhatsApp-based team communication and academic performance prediction.

2.1 Tuckman's Model of Group Development

A fundamental theoretical model for understanding group processes is Tuckman´s (1965) model of group development, which identifies distinct stages: forming, storming, norming, performing, and adjourning. Each phase reflects specific team dynamics, where communication plays a crucial role in resolving conflicts and enhancing collaboration (Tuckman, 1965). In the forming stage, members establish relationships and communication styles, while in the storming phase, conflicts may arise due to differing opinions and personalities (Tuckman, 1965). The norming phase fosters the establishment of shared expectations, and in the performing phase, teams operate at peak efficiency with streamlined communication (Tuckman, 1965). Finally, the adjourning phase concerns the dissolution of the team after task completion (Tuckman, 1965).

For our study, Tuckman's model is essential in understanding how different communication styles emerge and stabilize over time in digital settings such as WhatsApp. We utilize this model to analyze whether teams that reach a norming and performing stage demonstrate more synchronized and effective communication patterns, which in turn correlates with academic success.

2.2 Empirical Studies on Team Cooperation and Performance

Beyond the core studies discussed, research by Pulgar et al. (2019) suggests that classroom cooperation significantly benefits real-world problem-solving but can hinder performance in purely theoretical tasks. Their study highlights that students working collaboratively in applied learning environments exhibit better problem-solving abilities and deeper knowledge retention (Pulgar et al., 2019). However, in subjects requiring individual cognitive processing, such as algebra-based learning, collaboration may lead to distractions and over-reliance on group members (Pulgar et al., 2019). This dual effect is relevant for our study, as we examine how WhatsApp communication fosters collaboration in project-based learning while identifying potential downsides in theoretical coursework.

Additionally, Lim et al. (2023) demonstrate that teams with high levels of agreeableness perform better under uncertain conditions. Their research indicates that group members who exhibit agreeableness are more likely to maintain a positive team atmosphere, resolve conflicts efficiently, and encourage collaborative problem-solving (Lim et al., 2023). This finding supports our hypothesis that sentiment-based cohesion within WhatsApp groups contributes to successful teamwork (Lim et al., 2023). By incorporating sentiment analysis into our model, we aim to determine whether emotionally supportive and cooperative digital interactions are positively associated with academic performance.

2.3 Communication Patterns and Predictive Modeling

Figueira (2015) demonstrated that communication patterns, such as message density and response time, can be used to predict team success. By analyzing interaction logs, Figueira's work illustrates how quantitative communication data can serve as an early indicator of group efficiency and cohesion. Hadyaoui and Cheniti-Belcadhi (2023) extend this approach by integrating ontology-based analytics with machine learning models to predict project group performance. Their framework leverages structured ontologies to classify and interpret communication behavior, offering a scalable method for performance prediction (Hadyaoui & Cheniti-Belcadhi, 2023). For our study, we adapt the methodologies proposed by Figueira and Hadyaoui & Cheniti-Belcadhi by implementing sentiment and response-time analyses on WhatsApp group conversations. This approach enables us to determine which communication features are most predictive of academic success, contributing to a deeper understanding of how digital interactions shape performance.

2.4 Happimetrics: A Framework for Measuring Communication Quality

Happimetrics, introduced by Gloor (2023), offers a quantitative approach to assessing team collaboration by analyzing communication data. This framework measures group success through three core metrics:

- *Sentiment Analysis:* Evaluates the emotional tone of messages, which serves as an indicator of group motivation and engagement (Gloor, 2023, p.73).
- *Synchrony*: Measures how well team members align in timing and thematic focus, signifying collaboration efficiency (Gloor, 2023, p.124-127).
- *Network Analysis:* Identifies key roles within a team, such as central figures or isolated members, to quantify group dynamics (Gloor, 2023, p.20-22).

By applying these metrics, Happimetrics provides a structured method for assessing communication effectiveness. In the context of WhatsApp-based team communication, this framework is valuable for determining whether sentiment shifts or message synchrony correlate with group performance. Our research builds upon Gloor's Happimetrics model, extending it to informal digital communication by assessing WhatsApp chat logs and evaluating whether the identified roles (e.g., central communicators, peripheral members) predict academic achievement.

2.5 Emotional Feedback and Team Performance

Emotional expressiveness has been demonstrated to play a crucial role in team collaboration (Schneider et al., 2024). The present study examined how real-time emotional feedback via a *virtual mirror* system enhances teamwork, and the results indicated that teams receiving such feedback perform better, showing higher engagement and improved coordination. A key result is that greater emotional variability - alternating between positive and negative expressions - correlates with better team performance. This finding extends Happimetrics (Gloor, 2023) by emphasizing emotional dynamics over static sentiment balance. Additionally, vocal arousal, expressive communication, and moderate interruptions positively impact collaboration, suggesting that strong emotional involvement fosters team cohesion rather than disrupting it. These insights provide a theoretical basis for analyzing sentiment variation and expressive language use in academic teamwork, highlighting the relevance of emotional dynamics in digital collaboration.

2.6 The Role of NLP and Machine Learning in Communication Analysis

Recent advances in NLP and machine learning have significantly enhanced the ability to analyze large-scale communication data (Devlin et al., 2018). Transformer-based models, such as BERT (Devlin et al., 2018), enable the precise extraction of thematic coherence and emotional nuances from text-based interactions. These models facilitate several key analytical techniques:

- *Emotion and Sentiment Analysis:* Helps in detecting underlying team sentiment, which can indicate morale and potential conflicts (Devlin et al., 2018).
- *Topic Modeling*: Identifies dominant themes in discussions, revealing the alignment between group communication and task objectives (Devlin et al., 2018).
- *Entity Recognition and Role Detection:* Discerns key actors within a team based on interaction frequency and influence (Devlin et al., 2018).

By leveraging NLP-driven methodologies, the present research integrates sentiment analysis and role detection to explore how WhatsApp conversations reflect team effectiveness. We extend the application of BERT-based models by fine-tuning them to classify collaborative efficiency indicators specific to student group work, adding a novel dimension to previous research.

2.6 Research Gaps and Contributions

While existing literature provides strong foundations for studying communication in teams, gaps remain in integrating multiple theoretical and methodological perspectives. Tuckman's model provides insights into team development but lacks predictive power regarding communication data. Happimetrics quantifies interaction quality but does not account for deep linguistic features, which NLP can address. Similarly, while studies by Figueira (2015) and Hadyaoui & Cheniti-Belcadhi (2023) demonstrate effective predictive models, their focus remains on structured team settings rather than informal digital communication spaces such as WhatsApp.

This research addresses these gaps by combining Tuckman's developmental model, Happimetrics, and NLP-based analysis to create a comprehensive framework for evaluating team communication and predicting academic performance. By doing so, we not only advance theoretical models but also introduce practical applications for digital team collaboration, offering insights into how educators and institutions can foster more effective team interactions.

3 Methods

To predict student performance, this study systematically analyzes WhatsApp group communication data using a structured methodology. The process involves data collection and description, followed by preprocessing, feature engineering, aggregation, and selection before evaluating the predictive models. Each stage is outlined in this section to ensure methodological transparency and replicability.

3.1 Data Collection and Description

The dataset, provided by *Universidad Politécnica de Madrid*, comprises WhatsApp communication records from 28 student groups, each consisting of 4 to 6 individuals, spanning two academic years, 2023 and 2024. In total, approximately 10,500 messages were collected from 72 participants. To preserve privacy, all names were replaced with unique hashed identifiers (Tenzer et al., 2014).

3.2 Data Preprocessing

To standardize the dataset and remove irrelevant noise, a structured preprocessing pipeline was applied, addressing text cleaning, normalization, and privacy protection.

First, message cleaning was performed by removing emojis, special characters, URLs, mentions, and timestamps to standardize the textual content and ensure consistency across the dataset (Davidson et al., 2017). Next, text normalization was applied, which involved converting all messages to lowercase and eliminating common stopwords to reduce noise in the text data (Jauhiainen et al., 2018). To preserve participant anonymity and ensure compliance with data privacy regulations, all participant identifiers were hashed, effectively anonymizing personal information (Tenzer et al., 2014). Additionally, any numbers (phone numbers) and words containing the symbol "@" (emails) were removed to maintain privacy in the subsequent analysis. Furthermore, a filtering mechanism was implemented to exclude messages containing fewer than three words, as shorter messages were unlikely to contribute meaningful information to the analysis (Hutto & Gilbert, 2014).

A structured diagram illustrating the data aggregation process is provided in the *Appendix Figure A1.*

3.3 Feature Engineering

Understanding how communication patterns relate to student performance requires the identification of meaningful linguistic and interaction-based features. Based on theoretical considerations and existing research, a set of relevant features was extracted from the dataset to capture key aspects of language use and engagement. These features serve as the foundation for the subsequent analysis and are described in detail below, including their definition, conceptual relevance, and role within this study. A comprehensive overview of the extracted features is provided in *Appendix Table A1.*

Language Detection identifies the dominant language used in messages, a crucial factor in analyzing communication consistency (Jauhiainen et al., 2018). Linguistic uniformity within a group can enhance efficiency and collaboration (Jauhiainen et al., 2018). In this study, language identification ensured that messages were

predominantly in a single language, allowing for an assessment of linguistic diversity within student teams.

Sentiment Analysis measures the emotional polarity of messages on a scale from -1 (negative) to 1 (positive) (Hutto & Gilbert, 2014). Emotional expression within a team can impact engagement and group performance (Hutto & Gilbert, 2014). The application of Vader and TextBlob allowed us to quantify the emotional tone of messages, investigating whether higher positive sentiment correlated with better team performance.

The **Berger Score** Calculation assesses linguistic style by analyzing the usage of personal pronouns and verb tenses (Boghrati et al., 2023). Prior research suggests that present-tense communication and personal language are associated with more engaging and effective interactions (Boghrati et al., 2023). By computing the ratios of personal versus impersonal words and present versus past-tense verbs, this feature helped determine whether linguistic engagement influenced academic success.

Cosine Similarity evaluates the coherence of discussions by comparing the semantic similarity of consecutive messages (Blei et al., 2003). A high similarity score suggests structured and focused discussions, while low similarity may indicate fragmented conversations (Blei et al., 2003). This feature was implemented using sentence embeddings to assess whether cohesive dialogues contributed to better performance.

Dialogue Act Tagging categorizes messages into different communicative functions, such as statements, questions, or directives (Stolcke et al., 2000). Well-structured conversations have been shown to enhance group coordination and effectiveness (Stolcke et al., 2000). The application of a pre-trained model allowed us to examine whether organized dialogue structures were linked to better student outcomes.

Sentiment Trajectory Analysis tracks the evolution of sentiment across messages to detect emotional fluctuations within a conversation (Zhang et al., 2013). Emotional variation can indicate stress levels or cooperation dynamics (Zhang et al., 2013). By analyzing sentiment over time, we explored whether shifts in emotional tone influenced group interactions and student performance.

Five-Factor Inventory (FFI) Trait Extraction applies personality analysis techniques to infer personality traits from textual communication (John & Srivastava, 1999). Different personality traits influence communication styles and teamwork efficiency (John & Srivastava, 1999). By leveraging lexicon-based personality mapping, we investigated whether certain personality traits within groups correlated with academic achievement.

Inappropriate Word Detection identifies toxic or offensive language within conversations (Davidson et al., 2017). Negative or hostile communication can hinder collaboration and group cohesion (Davidson et al., 2017). By applying a predefined lexicon and classification model, we assessed whether the presence of inappropriate language affected group dynamics and overall performance.

Gini Coefficient Calculation measures inequality in participation by analyzing word count distribution among participants (van Mierlo et al., 2016). Groups with

more balanced participation tend to exhibit stronger teamwork (van Mierlo et al., 2016). By computing the Gini coefficient, we examined whether equitable communication contributed to better student performance.

Focus on content over traits is applied due to access to detailed messages, but not the students as individuals themselves of the course. The focus of the feature extraction was on communication patterns rather than on personality traits of the students. Thus only one personality assessment feature is included in an attempt to control for personality of the members, which has been shown as important in business success in previous studies like Frank et. al (2000), John & Srivastava (1999) and Owens et. al (2013).

3.4 Feature Aggregation

For the final prediction, two data sources need to be combined. The previously generated message-level features, which capture various linguistic and structural attributes of WhatsApp messages (as described in the previous section), serve as input to the models. The course grades, provided at the group level, represent overall group performance and act as the target variable for predictive modeling, allowing for evaluation of the prediction results.

To integrate these sources, the feature dataset is merged with the label dataset using the team's anonymous identifier. Entries without corresponding grade information are removed to maintain dataset completeness and analytical consistency (Han et al., 2012).

Since the target variable represents group-level performance, the message-level features must be aggregated accordingly. As the optimal aggregation method is unknown, multiple statistical measures are computed, including minimum, maximum, mean, median, standard deviation, interquartile range, range, skewness, and kurtosis (Kelleher et al., 2015, p.9-14).

3.5 Feature Selection Strategy

Given that the dataset contains over 100 aggregated features and only 28 data points, feature selection is necessary to avoid severe over-fitting of the models. For optimal predictive performance, the analysis focused on selecting the most meaningful and representative features.

The selection process prioritized features that exhibited a strong correlation with final grades, as those were most likely to capture predictive patterns (Ramaswami & Bhaskaran, 2009). Additionally, features with substantial variation across different groups were preferred, as highly uniform variables provided limited discriminatory power (Chandrashekar & Sahin, 2014). To implement this, two separate lists were created: one containing the features with the highest linear Pearson correlation to final grades and another comprising those with the greatest variance. The final selection was derived from the intersection of these lists, ensuring that the chosen features were both strongly associated with the target

variable and sufficiently diverse for predictive modeling. To further refine the feature set, multicollinearity filtering was performed, eliminating variables that demonstrated excessive correlation with each other (Dormann et al., 2013). If two features exhibited a correlation coefficient above 0.8, the one with lower variability was excluded to prevent redundancy (Kuhn & Johnson, 2013, p.27-58). A detailed overview of all selected features is shown in *Appendix Table A2* and *Appendix Table A3*. Further analysis are evaluated in section 4.

By integrating correlation-based selection, variance analysis, and redundancy filtering, the final feature set was optimized to include only the most informative and statistically significant predictors (Guyon & Elisseeff, 2003; James, 2021, p.59-75).

3.6 Model Evaluation Framework

As part of the data preprocessing, steps were taken to ensure consistency and model stability before evaluation. Since the dataset contained only numerical labels and preselected features, preprocessing was limited to feature scaling and optional dimensionality reduction. Numerical features were standardized to zero mean and unit variance to prevent bias and improve algorithm performance (Jain et al., 2000). Additionally, Principal Component Analysis (PCA) was used to mitigate multicollinearity and enhance interpretability (Smith et al., 2023). With the data prepared, the following section outlines the framework used to evaluate model performance.

To establish the relationship between team communication and group performance, various regression models were trained using a customized parameter grid. Model evaluation was conducted via 5-fold cross-validation, ensuring robustness despite the limited dataset size of only 28 groups. Due to the small sample size, a separate test set was omitted to prevent excessive data reduction and potential overfitting. Instead, cross-validation results were averaged across folds, which typically approximates real-world test performance reliably (Kuhn & Johnson, 2013, p.27-58).

Cross-validation is a widely used technique in predictive modeling, as it mitigates overfitting by partitioning the dataset into multiple training and validation subsets. This approach ensures that models are exposed to different data distributions, enhancing their generalizability to unseen data (Kohavi, 1995). By applying this validation strategy, performance estimates become more reliable, leading to a more accurate assessment of the model's predictive strength (James, 2021, p.197-206).

Model performance was assessed using Mean Squared Error (MSE) and the coefficient of determination (R^2), both standard measures for evaluating regression models (Chai & Draxler, 2014). MSE measures how much, on average, the squared differences among observed and predicted values deviate, making it a useful metric for assessing model accuracy. Meanwhile, R^2 evaluates the extent to which the model accounts for variability in the data, reflecting its explanatory strength (James, 2021, p.59-75). The best-performing model was selected based on the lowest MSE

score, ensuring optimal accuracy and generalizability. These evaluation steps provided a rigorous assessment framework, enhancing the model's reliability for real-world applications.

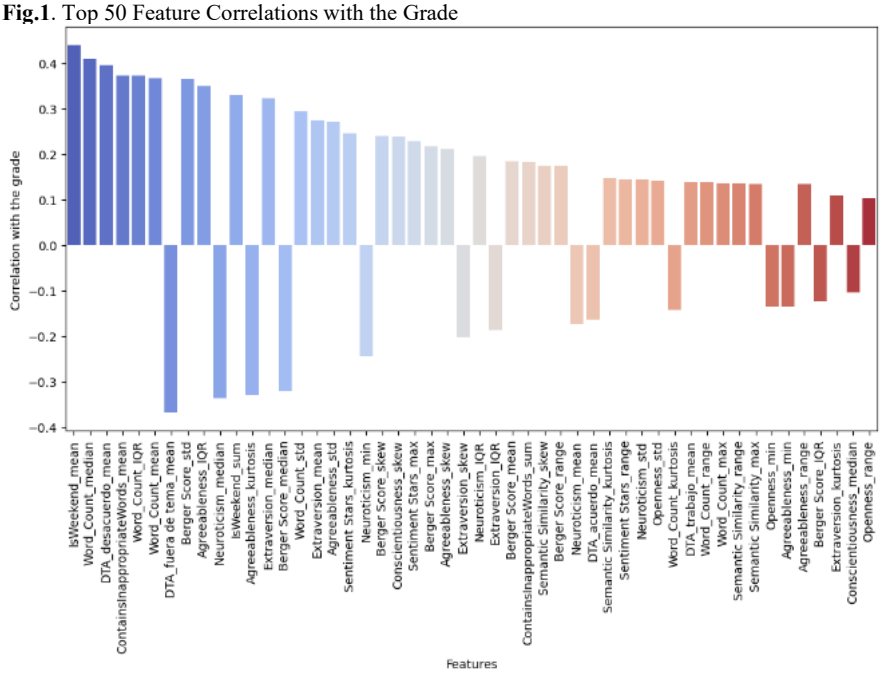
4 Results

The final results are presented in two sections: the selection of relevant features for the predictive model and the evaluation of model performance. This section first analyzes the key features contributing to grade prediction, followed by the results of the model evaluation.

4.1 Feature Selection and Correlation Analysis

Feature selection was conducted with the objective of identifying characteristics that exhibited a high degree of correlation with the final grades, which served as the target variable, and variability across different teams, measured by the coefficient of variation. *Figure 1* provides a visual representation of the correlation between individual features and final grades. The strongest positive correlations were observed in variables related to participation levels, sentiment balance, and linguistic diversity.

Fig.1. Top 50 Feature Correlations with the Grade



A salient finding of the study was the strong correlation between **word count**, specifically its mean, median, and Interquartile Range (IQR), and final grades. The findings suggest that teams engaging in longer message exchanges demonstrated higher performance, potentially attributable to enhanced collaboration and content exchange. Additionally, measures of dispersion, such as the IQR and standard deviation, indicated that teams with more varied message lengths exhibited improved performance. This suggests that balanced contributions from all team members may be advantageous.

A noteworthy correlation was identified in the examination of the use of **inappropriate language**. Contrary to initial suppositions, its presence exhibited a positive correlation with final grades. This observation suggests two potential implications. Firstly, the use of inappropriate language may serve as an indication of a high level of engagement and critical discussion, thereby fostering collaboration and facilitating more in-depth conversations. Alternatively, it could be a sign of a relaxed group dynamic, which, in turn, might enhance team cohesion and improve the quality of interaction. These findings are consistent with the work of Schneider et al. (2024), which highlights the positive effects of emotional intensity and variability on virtual team collaboration (see section 2.5). Their research suggests that highly engaged teams, characterized by expressive emotional exchanges, tend to perform better. The present results support this hypothesis by demonstrating a positive correlation between the use of strong language and final grades.

The present study sought to ascertain the relationship between disagreement, as measured by **"desacuerdo" dialogue sections**, and final grades. The findings indicated a positive correlation between the two variables. A certain level of disagreement appeared to be beneficial, as it encouraged critical discussions and active participation, both of which contribute to productive collaboration. Conversely, off-topic messages were negatively correlated with final grades, suggesting that staying focused on relevant discussions is important for academic success.

The personality traits derived from the **FFI** also played a role in team performance. Agreeableness demonstrated a strong correlation with final grades, particularly its measures of dispersion (i.e., IQR and standard deviation), indicating that a combination of cooperative and critical viewpoints within a team can be advantageous. Conversely, neuroticism exhibited a negative correlation, particularly with range and standard deviation, suggesting that emotional instability within teams may impede collaboration and overall academic performance. Finally, extraversion, measured by its median and range, demonstrated that groups with more extroverted members engaged in more discussions, which positively impacted grades.

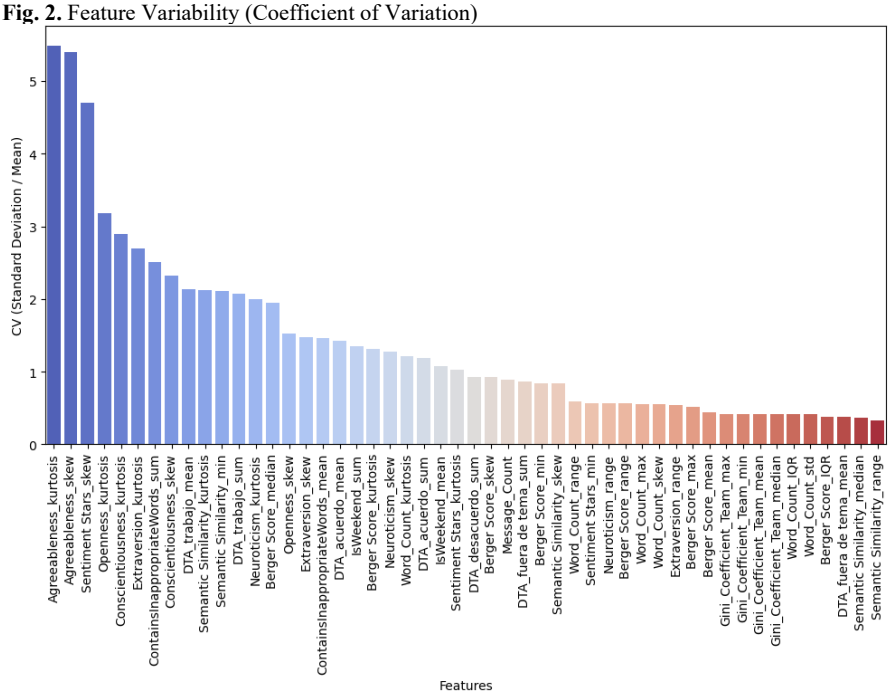
The results of the **sentiment analysis** indicated that teams with a wide range of emotional expressions, as measured by sentiment standard deviation and skewness, tended to demonstrate superior performance. This finding suggests that discussions incorporating both positive and negative sentiments were associated with dynamic and engaging conversations. However, extreme sentiment shifts (e.g., highly

negative or overly positive discussions) might indicate stress or conflict, which could potentially affect group cohesion and efficiency.

Interestingly, **semantic similarity** and **cosine similarity** measures, which assess the coherence and topic consistency of conversations, were not among the most highly correlated features. While maintaining topic consistency may be useful, other factors such as participation levels, sentiment balance, and personality-driven interactions appeared to have a more substantial influence on group performance.

As illustrated in *Figure 2*, the coefficient of variation for each feature provides insights into the differences in team communication behaviors. Features with higher variability, such as Agreeableness kurtosis and Sentiment skewness, offer the model greater potential to distinguish between teams, as they indicate varying communication styles. In contrast, features with lower variability, such as average word count and sentiment range, suggest more consistent patterns across teams, potentially contributing less to differentiation.

Fig. 2. Feature Variability (Coefficient of Variation)



4.2 Model Evaluation and Performance Analysis

For model training, different feature sets were evaluated to identify the most effective predictors while minimizing overfitting. Using the proposed selection pipeline, two feature subsets were constructed. The first subset was derived from the intersection of the 40 features most strongly correlated with the target variable

and the 25 features with the highest coefficient of variation, resulting in 14 features, which were further reduced to 7 using PCA (see *Table 1* and *Table 2*). The second subset was formed by intersecting the top 25 most correlated features with the top 20 features exhibiting the highest coefficient of variation, yielding 7 features that were subsequently reduced to 5 via PCA. An overview of the Top 50 Correlations and Correlation Coefficients can be found in *Appendix Table A2* and *A3*.

Table 1. Final Feature Selection of First Subset

First Subset Features	Correlation	Correlation Coefficients
'Word_Count_kurtosis'	-0.142724	1.218458
'Sentiment Stars_kurtosis'	0.246652	1.025257
'Conscientiousness_skew'	0.239254	2.322555
'Berger Score_median'	-0.320585	1.949992
'IsWeekend_mean'	0.439734	1.078543
'ContainsInappropriateWords_mean'	0.373052	1.456162
'Agreeableness_kurtosis'	-0.329789	5.480681
'ContainsInappropriateWords_sum'	0.182239	2.513619
'DTA_trabajo_mean'	0.139187	2.135713
'Semantic Similarity_kurtosis'	0.147556	2.124410
'Extraversion_skew'	-0.201199	1.477168
'IsWeekend_sum'	0.331614	1.345185
'Agreeableness_skew'	0.212230	5.393610
'DTA_acuerdo_mean'	-0.163364	1.420363

Table 2. Final Feature Selection of Second Subset

Second Subset Features	Correlation	Correlation Coefficients
'Conscientiousness_skew',	0.239254	2.322555
'Berger Score_median'	-0.320585	1.949992
'ContainsInappropriateWords_mean'	0.373052	1.456162
'Agreeableness_kurtosis'	-0.329789	5.480681
'Extraversion_skew',	-0.201199	1.477168
'IsWeekend_sum',	0.331614	1.345185
'Agreeableness_skew'	0.212230	5.393610

To ensure model stability and prevent overfitting, various feature combinations were tested under the constraint that the number of features should not exceed 7, corresponding to approximately 25% of the number of available data points. The final configurations were selected based on their superior performance among models with 7 features and those with 6 or fewer.

As outlined in the methods section, model training and evaluation were conducted using a grid search in combination with five-fold cross-validation. The exact

parameter grids and training code are provided in the supplementary materials. The best-performing models were Gradient Boosting for the 7-feature configuration and LightGBM for the 5-feature configuration.

Across all five cross-validation folds, these models achieved MSEs of 0.438 and 0.463, translating to absolute deviations of approximately 0.66 to 0.7 grade points. When applied to the full dataset, they yielded R^2 scores of 0.343 and 0.393, indicating that the models captured approximately 34–39% of the variation in the target variable.

The predictive performance of the selected features was evaluated through the implementation of regression models, which were trained using a grid search and assessed using a five-fold cross-validation approach.

4.3 Key Takeaways and Practical Insights

This study provides insights into the relationship between communication patterns and student performance in team-based learning environments.

Participation levels and balanced contributions emerged as important factors for academic success. Teams that exchanged a higher volume of messages, as reflected in mean and median word count, tended to perform better. Additionally, greater variability in message length (e.g., interquartile range and standard deviation) was associated with higher grades, suggesting that diverse participation rather than dominance by a few individuals fosters collaboration.

The role of disagreement and engagement was also notable. Higher occurrences of disagreement ("desacuerdo" dialogue sections) appeared to encourage critical discussions and problem-solving. Interestingly, the presence of inappropriate language correlated positively with final grades, potentially indicating intense engagement and active discussions rather than purely disruptive behavior.

Personality traits further influenced team dynamics. Greater dispersion in agreeableness was linked to better performance, suggesting that a combination of cooperative and critical viewpoints enhances group interactions. In contrast, elevated levels of neuroticism, particularly its IQR and standard deviation, were negatively associated with performance, implying that emotional instability may hinder collaboration. Meanwhile, extraversion, especially its median and range, positively impacted group discussions and outcomes.

Emotional expression and sentiment balance also played a role in group performance. Teams exhibiting a broad spectrum of emotional expression - both positive and negative - tended to achieve higher grades, likely due to more dynamic and engaging discussions. However, extreme sentiment fluctuations could signal stress or conflict, potentially disrupting group cohesion and efficiency.

In addition, variability in communication patterns contributed to the model's ability to differentiate between teams. Features such as agreeableness kurtosis and sentiment skewness captured diverse communication styles that influenced performance predictions. On the other hand, features with lower variability, such as

average word count and sentiment range, reflected more consistent communication patterns across teams, making them less relevant for distinguishing group dynamics. Despite these findings, the predictive power of the model was limited. While it explained approximately 35% of the variation in final grades, its generalizability declined when applied to a different academic cohort. This suggests that external influences, including individual study habits, prior knowledge, and instructor effects, significantly impact student performance. These results highlight the complexity of academic success and the necessity of incorporating additional contextual factors beyond communication-based models.

4.4 Practical Implications

- Educators and team facilitators should encourage balanced participation and constructive debates to enhance group discussions.
- Personality composition within teams may impact collaboration, and awareness of emotional and social dynamics could help optimize group work strategies.
- While automated linguistic analysis can provide insights into teamwork effectiveness, its predictive power remains constrained by external influences, emphasizing the need for complementary assessment methods.

5 Discussion

The results of this study provide valuable insights into the relationship between digital communication patterns and academic performance in student group projects. While the results highlight key correlations between linguistic, structural, and sentiment-based features and team success, they also underscore various challenges and limitations associated with predictive modeling in small datasets. This discussion critically evaluates these findings, exploring both their implications and the methodological constraints that may influence their interpretation.

5.1 Strengths and Implications

As outlined in the Key Takeaways (see previous section), this study demonstrates that communication features can predict academic performance, though with limitations. The best models explained 34–39% of grade variance, confirming that participation levels, sentiment balance, and personality traits are meaningful predictors, even if external factors also influence outcomes.

The strong correlations between selected features and final grades highlight their relevance in understanding digital collaboration. Personality traits, particularly agreeableness and extraversion, shaped team dynamics, suggesting that structured participation strategies and personality-aware team formation may enhance learning outcomes. Additionally, automated sentiment monitoring could help educators identify disengagement or conflict early, improving group performance.

5.2 Challenges

Despite the promising findings of this study, several methodological challenges require cautious interpretation of the results. One of the primary limitations is the limited predictive power of the models when applied to unseen data. While the models performed reasonably well on the training data, achieving R^2 values between 0.3426 and 0.3926, their generalizability remained limited. This suggests that cohort-specific patterns of behavior, group dynamics, or external academic influences - not captured in the dataset - may play a critical role in student performance. Similar findings have been reported in academic prediction studies, where contextual variables significantly affect model accuracy (James, 2021, p.59-75).

Another notable contradiction emerged in the relationship between inappropriate language use and academic performance. Contrary to expectations, groups that used inappropriate or toxic language more frequently achieved higher final grades. This finding challenges conventional assumptions that respectful discourse fosters collaboration and raises questions about the role of informal or emotionally charged interactions in teamwork. Prior research suggests that the intensity of engagement, rather than the decorum of language, may be a stronger predictor of performance in collaborative environments (Figueira, 2015). However, the causal mechanism remains unclear and warrants further qualitative analysis. Schneider et al. (2024) provide a potential explanation for this phenomenon (see section 2.5). Their study suggests that emotional expressiveness, including strong emotional reactions, can foster engagement and improve collaboration in virtual teams. This could explain why the presence of inappropriate language in our dataset correlates positively with academic performance, as it may signal intense involvement in discussions rather than disruptive behavior.

Similarly, while constructive disagreement ("desacuerdo") was positively correlated with performance, it remains uncertain to what extent disagreement consistently improves collaboration. While structured debates are likely to improve problem-solving, excessive conflict or unstructured disputes could hinder teamwork. Because the study does not distinguish between productive and disruptive disagreements, the observed correlation should be interpreted with caution. Prior research on digital collaboration emphasizes that disagreement can be beneficial when managed constructively, but unregulated conflict can negatively affect group cohesion (Pulgar et al., 2019).

Another challenge is feature selection and model robustness. Given the high-dimensional nature of the dataset relative to the limited number of observations, there is a risk that certain features will be overfitted to the training data. The trade-off between preserving predictive signals and avoiding overfitting was particularly pronounced when selecting highly variable and highly correlated features. While PCA was used to mitigate this problem, future studies should explore additional dimensionality reduction techniques to improve model stability and

generalizability. Previous research on feature selection in small datasets has highlighted similar concerns, emphasizing that dimensionality reduction is essential to prevent overfitting (Kuhn & Johnson, 2013).

Finally, the role of topic coherence in communication remains unclear. While structured discussions are often associated with effective collaboration (Figueira, 2015), our results show that semantic similarity kurtosis, a measure of discussion coherence, was not among the strongest predictors of academic success. This suggests that flexibility in conversation flow, which allows teams to explore different perspectives rather than rigidly adhering to a single trajectory, may be as important as maintaining coherence. Alternatively, the weak correlation may reflect the limitations of semantic similarity metrics in capturing the nuances of informal digital communication. Prior research suggests that rigid discussion structures may limit creativity and engagement, highlighting the need to balance coherence with conversational flexibility (van Mierlo et al., 2016).

These challenges underscore the complexity of predicting academic performance based solely on communication data and highlight the need for future research to refine feature selection, improve generalizability, and integrate external academic factors for a more holistic understanding of team-based learning.

5.3 Practical Applications and Future Considerations

Despite these challenges, the study's findings provide actionable insights for educators seeking to enhance digital group work practices. The observed correlation between participation balance and academic success suggests that interventions encouraging equitable communication within teams may improve student outcomes. Educators could integrate communication-based engagement metrics into learning management systems, providing real-time feedback on team interactions.

Additionally, automated sentiment analysis tools could be leveraged to identify early indicators of communication breakdowns or emotional distress within teams. By flagging instances of highly negative sentiment or increasing emotional volatility, instructors could offer timely support to groups experiencing difficulties, fostering a more productive and supportive learning environment.

In conclusion, while the study presents compelling evidence for the relationship between digital communication and academic performance, it also underscores the complexities of predictive modeling in educational research. Addressing issues related to model generalizability, feature selection, and evaluation methodologies will be crucial for refining future predictive frameworks. The integration of additional contextual variables, such as task complexity and external academic pressures, could further enhance the robustness of predictive models and provide deeper insights into the mechanisms underlying effective digital collaboration.

6 Limitations and Future Work

6.1 Limitations

This study had several limitations, primarily related to the size of the data set and the evaluation of the model. With only 28 student groups, the models were susceptible to overfitting, which limited their generalizability. While fivefold cross-validation was used to increase robustness, the small test sets in each fold resulted in unstable performance estimates, making the results sensitive to outliers. The lack of an independent test set further limited the ability to validate the models on unseen data.

Another challenge was the reliability of NLP-derived features. A significant number of linguistic and sentiment-based variables were generated using pre-trained models that have not been extensively validated in educational contexts. This raises concerns about the propagation of uncertainty, where errors in feature extraction may have influenced predictions. In addition, the present study focused only on digital communication, excluding external factors such as individual study habits, teacher feedback, and prior knowledge, which are critical to academic performance.

6.2 Future Work

In the future, researchers should seek to expand the dataset by including multiple academic years and institutions. This could improve the generalizability of the results. To improve the reliability of performance estimates, researchers should implement improved validation techniques. Some possible techniques could include providing a special test set or using bootstrapping. In addition, benchmarking NLP-derived features against established datasets and incorporating manual annotation could also improve feature reliability.

In addition, the integration of multimodal data, including self-reported engagement, peer ratings, and instructor feedback, may have the potential to provide a more comprehensive framework for predicting performance. Further exploration through longitudinal studies could elucidate the evolution of communication patterns over time and determine whether early interactions serve as predictors of long-term success. In addition, improving the interpretability of the model through the implementation of explicable AI techniques would make the results more actionable for educators and students.

7 Conclusion

This study used NLP and machine learning to examine the relationship between digital communication patterns and academic performance in student group projects. Key communication characteristics - including participation levels, sentiment balance, and personality traits - emerged as significant predictors of success. Active engagement and emotional consistency contributed positively,

while excessive sentiment fluctuations and unstructured disagreements were associated with lower performance.

By translating language into numerical representations, this study allowed for scalability and repeatability, albeit at the expense of some nuance compared to manual analysis. Interestingly, simpler features such as participation level and sentiment balance were the strongest predictors, suggesting that basic communication patterns provide greater explanatory power than complex linguistic models.

Despite limitations in predictive power, the findings offer practical implications for optimizing team dynamics in digital learning. Encouraging structured participation, personality-aware team formation, and sentiment monitoring could improve collaboration and learning outcomes. Future research should focus on expanding datasets, refining models, and incorporating additional academic factors to improve predictive accuracy.

By combining computational analysis with educational research, this study advances the understanding of digital collaboration and provides insights for students, educators, and institutions seeking to improve team-based learning experiences.

References

1. Amelia, L. T. D., & Balqis, N. R. (2023). Changes in Communication Patterns in the Digital Age. *ARRUS Journal of Social Sciences and Humanities*, *3*(4), 544–556. https://doi.org/10.35877/soshum1992
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.
3. Boghrati, R., Berger, J., & Packard, G. (2023). Style, content, and the success of ideas. *Journal of Consumer Psychology*, *33*(4), 688–700. https://doi.org/10.1002/jcpy.1346
4. Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014
5. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024
6. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv.Org*. https://www.proquest.com/docview/2074118430?sourcetype=Working%20Papers
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/arXiv.1810.04805
8. Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x
9. Eming, J., & Philipowski, K. (2022, Mai 11). *Wie Corona die akademische Lehre dauerhaft verändert*. Forschung & Lehre. https://www.forschung-und-lehre.de/lehre/wie-corona-die-akademische-lehre-dauerhaft-veraendert-4678
10. Figueira, A. (2015). Predicting Results from Interaction Patterns During Online Group Work. In G. Conole, T. Klobučar, C. Rensing, J. Konert, & E. Lavoué (Hrsg.), *Design for Teaching and Learning in a Networked World* (S. 414–419). Springer International Publishing. https://doi.org/10.1007/978-3-319-24258-3_33
11. Frank, H., Lueger ,Manfred, & and Korunka, C. (2007). The significance of personality in business start-up intentions, start-up realization and business success. *Entrepreneurship & Regional Development*, *19*(3), 227–251. https://doi.org/10.1080/08985620701218387

12. Gloor, P. A. (2017). *Swarm Leadership and the Collective Mind: Using Collaborative Innovation Networks to Build a Better Business* (1. Aufl.). Emerald Publishing Limited. https://www.researchgate.net/publication/345743512_Swarm_Leadership_and_the_Collective_Mind_Using_Collaborative_Innovation_Networks_to_Build_a_Better_Business
13. Gloor, P. A. (2023). *Happimetrics: Leveraging AI to Untangle the Surprising Link Between Ethics, Happiness and Business Success (New Horizons in Management)*. Edward Elgar Publishing Ltd.
14. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*(null), 1157–1182.
15. Hadyaoui, A., & Cheniti-Belcadhi, L. (2023). Ontology-based group assessment analytics framework for performances prediction in project-based collaborative learning. *Smart Learning Environments*, *10*(1), 43. https://doi.org/10.1186/s40561-023-00262-w
16. Han, J., Kamber, M., & Pei, J. (2012). 3—Data Preprocessing. In J. Han, M. Kamber, & J. Pei (Hrsg.), *Data Mining (Third Edition)* (S. 83–124). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00003-4
17. Hrastinski, S. (2008). Asynchronous and Synchronous E-Learning. *EDUCAUSE Review*, *4*. https://er.educause.edu/articles/2008/11/asynchronous-and-synchronous-elearning
18. Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1. https://doi.org/10.1609/icwsm.v8i1.14550
19. Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.824819
20. James, G. (2021). *An introduction to statistical learning* (2., Second edition Aufl.). Springer.
21. Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2018). *Automatic Language Identification in Texts: A Survey* (No. arXiv:1804.08186). arXiv. https://doi.org/10.48550/arXiv.1804.08186
22. John, O., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, *2*, 102–138. https://www.semanticscholar.org/paper/The-Big-Five-Trait-taxonomy%3A-History%2C-measurement%2C-John-Srivastava/a354854c71d60a4490c42ae47464fbb9807d02bf
23. Kelleher, J. D., Brian, M. N., & Aoife, D. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press. https://www.researchgate.net/publication/281089751_Fundamentals_of_

Machine_Learning_for_Predictive_Data_Analytics_Algorithms_Worked_Examples_and_Case_Studies
24. Kim, S., Bak, J., & Oh, A. (2012). Do You Feel What I Feel? Social Aspects of Emotions in Twitter Conversations. *Proceedings of the International AAAI Conference on Web and Social Media*, *6*(1), Article 1. https://doi.org/10.1609/icwsm.v6i1.14310
25. Kohavi, R. (1995). A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Articial Intelligence*. https://www.researchgate.net/publication/2352264
26. Kuhn, M., & Johnson, K. (2013). Data Pre-processing. In M. Kuhn & K. Johnson (Hrsg.), *Applied Predictive Modeling* (S. 27–59). Springer. https://doi.org/10.1007/978-1-4614-6849-3_3
27. Laurillard, D. (2012). *Teaching as a Design Science: Building Pedagogical Patterns for Learning and Technology*. Routledge. https://doi.org/10.4324/9780203125083
28. Lim, S. L., Bentley, P. J., Peterson, R. S., Hu, X., & McLaren, J. P. (2023). Kill chaos with kindness: Agreeableness improves team performance under uncertainty. *Collective Intelligence*, *2*(1). https://doi.org/10.1177/26339137231158584
29. Owens, K. S., Kirwan, J. R., Lounsbury, J. W., Levy, J. J., & Gibson, L. W. (2013). Personality correlates of self-employed small business owners' success. *WORK*, *45*(1), 73–85. https://doi.org/10.3233/WOR-12153
30. Pulgar, J., Candia, C., & Leonardi, P. (2019). Undergrad classroom cooperation and academic performance: Beneficial for real-world-like problems but detrimental for algebra-based problems. *Physical Review Physics Education Research*, *16*(1), 010137. https://doi.org/10.1103/PhysRevPhysEducRes.16.010137
31. Ramaswami, M., & Bhaskaran, R. (2009). *A Study on Feature Selection Techniques in Educational Data Mining* (No. arXiv:0912.3924). arXiv. https://doi.org/10.48550/arXiv.0912.3924
32. Roessler, J., & Gloor, P. A. (2020). Measuring happiness increases happiness. *Journal of Computational Social Science*, *4*(1), 123–146. https://doi.org/10.1007/s42001-020-00069-6
33. Schneider, N., Vazquez, I., & Gloor, P. A. (2024). *No pain no gain—Giving real-time emotional feedback in a virtual mirror improves collaboration in virtual teamwork*. Business, Economics and Management. https://doi.org/10.20944/preprints202406.0147.v1
34. Smith, K. A., Blease, C., Faurholt-Jepsen, M., Firth, J., Van Daele, T., Moreno, C., Carlbring, P., Ebner-Priemer, U. W., Koutsouleris, N., Riper, H., Mouchabac, S., Torous, J., & Cipriani, A. (2023). Digital mental health: Challenges and next steps. *BMJ Mental Health*, *26*(1), e300670. https://doi.org/10.1136/bmjment-2023-300670
35. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–374.

36. Tenzer, H., Pudelko, M., & Harzing, A.-W. (2014). The impact of language barriers on trust formation in multinational teams. *Journal of International Business Studies*, *45*(5), 508–535. https://doi.org/10.1057/jibs.2013.64
37. Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, *63*(6), 384–399. https://doi.org/10.1037/h0022100
38. van Mierlo, T., Hyatt, D., & Ching, A. T. (2016). Employing the Gini coefficient to measure participation inequality in treatment-focused Digital Health Social Networks. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, *5*(1), 32. https://doi.org/10.1007/s13721-016-0140-7
39. Zhang, X., Gloor, P. A., & Grippa, F. (2013). Measuring creative performance of teams through dynamic semantic social network analysis. *International Journal of Organisational Design and Engineering*, *3*(2), 165. https://doi.org/10.1504/IJODE.2013.057014
40. Zhu, C. (2012). Student satisfaction, performance, and knowledge construction in online collaborative learning. *Educational Technology & Society*, *15*, 127–136.

Appendix

Figure A1: Data Aggregation Process (own representation)

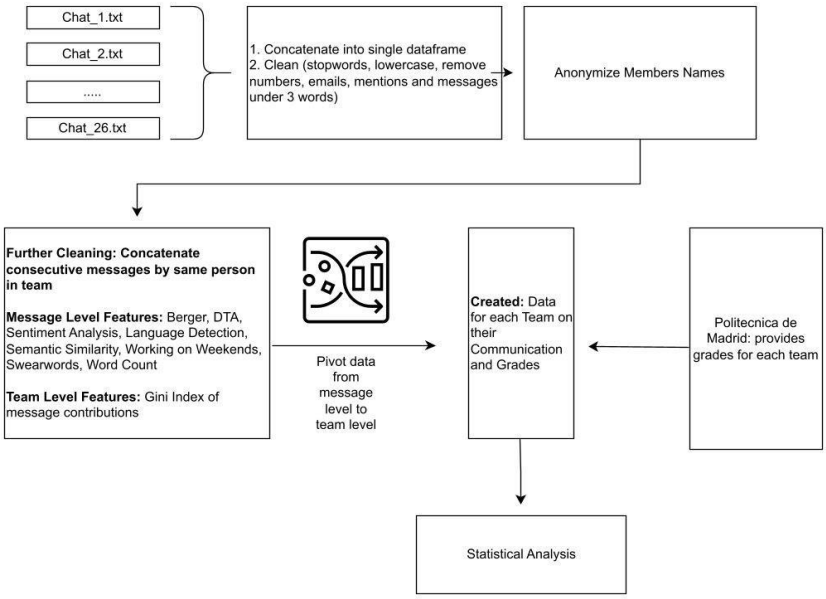


Table A1: Detailed Overview of Feature Engineering

Feature Name	Definition & Concept	Application & Implementation	Reason for Selection
Language Detection	Identifies the primary language of each message. Language uniformity within a group can impact clarity and collaboration efficiency (Jauhiainen et al., 2019).	A pre-trained language identification model was applied to classify each message's language, ensuring consistency across the dataset.	Groups that predominantly use a single language may have more effective communication, which can contribute to better teamwork and performance (Jauhiainen et al., 2019).
Sentiment Analysis	Measures the emotional tone of messages on a scale from -1 (negative) to 1 (positive). Emotional expression in conversations can influence engagement and group cohesion (Hutto et al., 2015).	Vader and TextBlob sentiment analysis models were used to assign sentiment scores to each message, allowing for sentiment trends to be tracked across group discussions.	Prior studies suggest that higher positive sentiment within teams correlates with better cooperation and performance outcomes, whereas excessive negativity may indicate conflicts (Hutto et al., 2015).
Berger Score Calculation	Analyzes linguistic style based on the use of pronouns and verb tense, where personal and present-tense language is linked to more engaging and effective interactions (Berger et al., 2020).	The proportion of personal pronouns and the ratio of present vs. past tense verbs were computed to assess engagement levels in group conversations.	Research indicates that individuals who use present-tense verbs and personal pronouns tend to be more actively engaged, which may positively influence group dynamics and academic outcomes (Berger et al., 2020).

Feature Name	Definition & Concept	Application & Implementation	Reason for Selection
Cosine Similarity	Measures the coherence of discussions by assessing the semantic similarity between consecutive messages. High similarity suggests structured and focused conversations, while low similarity may indicate fragmented discussions (Blei et al., 2003).	Sentence embeddings were computed, and cosine similarity was used to quantify the degree of semantic relatedness between consecutive messages in a conversation.	Structured and coherent discussions have been shown to facilitate more effective collaboration and problem-solving, which can enhance group performance (Blei et al., 2003).
Dialogue Act Tagging (DTA)	Categorizes messages into different communicative functions, such as statements, questions, or directives. Well-structured conversations improve coordination and team effectiveness (Stolcke et al., 2000).	A pre-trained model was used to classify messages based on their communicative function, helping to identify patterns in group interactions.	Teams that engage in more directive and informative exchanges tend to have clearer task coordination, which is associated with improved teamwork efficiency (Stolcke et al., 2000).
Sentiment Trajectory Analysis	Tracks the evolution of sentiment across messages, measuring how emotional tone fluctuates over time. Emotional variation can reflect stress levels, collaboration quality, or team dynamics (Zhang et al., 2013).	A time-series analysis of sentiment scores was conducted, allowing for an assessment of how sentiment changed over the course of group discussions.	Research suggests that sudden shifts in sentiment, such as a sharp increase in negativity, may indicate conflict or frustration, whereas stable sentiment trajectories are linked to steady cooperation (Zhang et al., 2013).

Feature Name	Definition & Concept	Application & Implementation	Reason for Selection
FFI Trait Extraction	Extracts personality traits from text based on linguistic markers. Different personality traits influence communication styles and teamwork efficiency (John & Srivastava, 1999).	A lexicon-based trait mapping method was used to infer personality traits of participants from their messages.	Prior studies suggest that individuals with traits such as openness and agreeableness contribute positively to group discussions, fostering a more collaborative environment (John & Srivastava, 1999).
Inappropriate Word Detection	Identifies messages containing offensive or inappropriate language, which can negatively affect collaboration and group cohesion (Davidson et al., 2017).	A predefined lexicon and classification model were used to detect and quantify the presence of inappropriate language in group chats.	Teams that exhibit high levels of toxic language are more likely to experience conflicts and decreased cooperation, ultimately impacting group performance (Davidson et al., 2017).
Gini Coefficient Calculation	Measures inequality in participation by analyzing the distribution of word counts among participants. A higher Gini coefficient indicates an imbalance in contribution levels (van Mierlo et al., 2016)	The Gini coefficient was computed based on the number of words contributed by each participant, assessing the level of participation equity within the group.	Prior research suggests that groups with a more balanced distribution of contributions tend to exhibit stronger teamwork and fairer workload distribution, which can enhance collective performance (van Mierlo et al., 2016).

Feature Name	Definition & Concept	Application & Implementation	Reason for Selection
Focus on Content over Traits	Estimates the Big Five personality traits (Openness, Conscientiousness, Extra-version, Agreeableness, Neuro-ticism) based on linguistic cues in WhatsApp messages. The FFI has established itself as a standard in measuring person-ality traits among literate, educated populations (John & Srivastava, 1999).	A lexicon-based personality mapping approach was applied to infer agree-ableness scores for each user, using message-level data. This was the only personality-related feature included.	Given that course participants under evaluation would have interacted with each other also outside of their digital communication, it is critical to attempt to control for this in understanding group success. Given the unavailability of the opportunity to interview the students themselves to gain insight into their personality, the subsequent most viable approach is to extract personality traits from their digital communication. This is of particular importance given the findings of studies on business leaders, which have demonstrated a strong correlation between leader personality and business success (Frank et al., 200; Owens et al., 2013).

Table A2: Top 50 Correlations (selected)

Feature	Value
IsWeekend_mean	0.439734
Word_Count_median	0.410472
DTA_desacuerdo_mean	0.395725
ContainsInappropriateWords_mean	0.373052
Word_Count_IQR	0.373022
Word_Count_mean	0.368248
DTA_fuera de tema_mean	-0.368209
Berger Score_std	0.366040
Agreeableness_IQR	0.350722
Neuroticism_median	-0.335909
IsWeekend_sum	0.331614
Agreeableness_kurtosis	-0.329789
Extraversion_median	0.323927
Berger Score_median	-0.320585
Word_Count_std	0.293977
Extraversion_mean	0.275350
Agreeableness_std	0.271523
Sentiment Stars_kurtosis	0.246652
Neuroticism_min	-0.245302
Berger Score_skew	0.240120
Conscientiousness_skew	0.239254
Sentiment Stars_max	0.227899
Berger Score_max	0.216327
Agreeableness_skew	0.212230
Extraversion_skew	-0.201199
Neuroticism_IQR	0.196190
Extraversion_IQR	-0.186168
Berger Score_mean	0.185156
ContainsInappropriateWords_sum	0.182239
Semantic Similarity_skew	0.175542
Berger Score_range	0.174674
Neuroticism_mean	-0.173649
DTA_acuerdo_mean	-0.163364
Semantic Similarity_kurtosis	0.147556
Sentiment Stars_range	0.144186
Neuroticism_std	0.143603
Openness_std	0.143124
Word_Count_kurtosis	-0.142724
DTA_trabajo_mean	0.139187
Word_Count_range	0.137435
Word_Count_max	0.137127

Semantic Similarity_range	0.137106
Semantic Similarity_max	0.134191
Openness_min	-0.133912
Agreeableness_min	-0.133816
Agreeableness_range	0.133627
Berger Score_IQR	-0.123731
Extraversion_kurtosis	0.110303
Conscientiousness_median	-0.103853
Openness_range	0.102974

Table A3: Top 50 Correlation Coefficients (selected)

Feature	**Value**
Agreeableness_kurtosis	5.480681
Agreeableness_skew	5.393610
Sentiment Stars_skew	4.695806
Openness_kurtosis	3.180121
Conscientiousness_kurtosis	2.890440
Extraversion_kurtosis	2.697708
ContainsInappropriateWords_sum	2.513619
Conscientiousness_skew	2.322555
DTA_trabajo_mean	2.135713
Semantic Similarity_kurtosis	2.124410
Semantic Similarity_min	2.105792
DTA_trabajo_sum	2.072751
Neuroticism_kurtosis	2.002006
Berger Score_median	1.949992
Openness_skew	1.529349
Extraversion_skew	1.477168
ContainsInappropriateWords_mean	1.456162
DTA_acuerdo_mean	1.420363
IsWeekend_sum	1.345185
Berger Score_kurtosis	1.308057
Neuroticism_skew	1.274055
Word_Count_kurtosis	1.218458
DTA_acuerdo_sum	1.187608
IsWeekend_mean	1.078543
Sentiment Stars_kurtosis	1.025257
DTA_desacuerdo_sum	0.932295
Berger Score_skew	0.929667
Message_Count	0.884416
DTA_fuera de tema_sum	0.864288
Berger Score_min	0.838323
Semantic Similarity_skew	0.835478

Word_Count_range	0.591873
Sentiment Stars_min	0.568176
Neuroticism_range	0.565138
Berger Score_range	0.563733
Word_Count_max	0.558080
Word_Count_skew	0.551982
Extraversion_range	0.541206
Berger Score_max	0.519135
Berger Score_mean	0.444202
Gini_Coefficient_Team_max	0.420180
Gini_Coefficient_Team_min	0.420180
Gini_Coefficient_Team_mean	0.420180
Gini_Coefficient_Team_median	0.420180
Word_Count_IQR	0.413447
Word_Count_std	0.410478
Berger Score_IQR	0.382839
DTA_fuera de tema_mean	0.382747
Semantic Similarity_median	0.368957
Semantic Similarity_range	0.328113

Appendix 5: Model Parameters

```
'RandomForest': {
    'verbose': [0],
    'n_estimators': [50, 100, 200, 500],
    'max_depth': [3, 5, 7, 10],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 5],
    'max_features': ['sqrt', 'log2'],
    'bootstrap': [True, False]
},
'GradientBoosting': {
    'verbose': [0],
    'n_estimators': [50, 100, 200, 500],
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
    'max_depth': [3, 5, 7, 10],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 5],
    'subsample': [0.7, 0.8, 0.9, 1.0],
    'max_features': ['sqrt', 'log2']
},
'XGBoost': {
    'n_estimators': [50, 100, 200, 500],
    'learning_rate': [0.01, 0.05, 0.1, 0.2],
```

```
    'max_depth': [3, 5, 7, 10],
    'min_child_weight': [1, 3, 5],
    'subsample': [0.7, 0.8, 0.9, 1.0],
    'colsample_bytree': [0.7, 0.8, 1.0],
    'reg_alpha': [0, 0.01, 0.1, 1],
    'reg_lambda': [0, 0.1, 0.5, 1],
    'gamma': [0, 0.1, 0.5, 1]
  },
  'Lasso': {
    'alpha': [0.001, 0.01, 0.1, 1, 10],
    'max_iter': [1000, 5000, 10000],
    'tol': [1e-3, 1e-4, 1e-5]
  },
  'Ridge': {
    'alpha': [0.001, 0.01, 0.1, 1, 10, 100],
    'max_iter': [1000, 5000, 10000],
    'tol': [1e-3, 1e-4, 1e-5]
  },
  'ElasticNet': {
    'alpha': [0.001, 0.01, 0.1, 1, 10],
    'l1_ratio': [0.1, 0.5, 0.9],
    'max_iter': [1000, 5000, 10000],
    'tol': [1e-3, 1e-4, 1e-5]
  },
  'LightGBM': {
    'verbose': [-1],
    'n_estimators': [50, 100, 200, 500],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [3, 5, 7, 10, -1],
    'num_leaves': [31, 50, 100],
    'min_child_samples': [5, 10, 20],
    'subsample': [0.7, 0.8, 0.9, 1.0],
    'colsample_bytree': [0.7, 0.8, 1.0],
    'reg_alpha': [0, 0.01, 0.1, 1],
    'reg_lambda': [0, 0.1, 0.5, 1]
  }
```

Appendix 6: Google Collab Link

- https://drive.google.com/drive/folders/1UXs3111wJwTbU_FSYgzjKEFnzGS5-7h0?usp=sharing

(If you encounter any difficulties opening the link, please contact one of the email addresses listed at the beginning).

Appendix 7: GitHub Repository

- https://github.com/AAV77/COIN---Content-Personality-and-Business-Success.git

(If you encounter any difficulties opening the link, please contact one of the email addresses listed at the beginning).