Text-Aware Prediction of Computer-Mediated Communication Competence in Collaborative Problem-Based Learning

Mobeen Akhter

INTI International University, Malaysia

Abstract

Computer-mediated communication (CMC) competence is a key prerequisite for effective participation in online collaborative learning and problem-based learning (PBL). In learning analytics, communication-related perceptions and outcomes are often approximated from behavioural traces such as message counts, access logs, and interaction intensity; however, behaviour-only indicators are limited in distinguishing constructive, context-appropriate communication from high-volume but unproductive participation. Building on a multidimensional view of CMC competence that encompasses communicative effectiveness, relational appropriateness, and clarity of expression, this manuscript specifies a text-aware forecasting protocol that integrates behavioural and linguistic features extracted from discussion posts and instant messages in a web-based collaborative PBL environment. The protocol combines secure xAPI-style logging, privacy-preserving extraction of message text, and a layered natural language processing feature set that includes lexical and syntactic properties, discourse-act proportions, socio-emotional and politeness indicators, and transformer-based embeddings. We further define baseline models, parameter-tuning procedures, and a leakage-resistant evaluation strategy based on group-aware nested cross-validation, ablation analyses, and uncertainty reporting (confidence intervals and robustness checks). The Results section functions as a reporting template that demonstrates how empirical findings should be summarised and interpreted and includes explicitly labelled illustrative values solely to show expected formats, not to claim empirical performance. The contribution is an implementation-ready analytic framework that aligns operational measures with the theoretical content of CMC competence and provides reusable guidance for future empirical studies and formative dashboard designs aimed at supporting communication in collaborative learning.

Keywords: computer-mediated communication competence, learning analytics, text mining, collaborative problem-based learning, predictive modelling, online discussion analysis

1. Introduction

Online and blended programmes increasingly rely on computer-mediated communication (CMC) for collaborative learning, project work, and problem-based learning (PBL). In such environments, students must not only master disciplinary content but also communicate clearly, coordinate task work, and manage socio-emotional dynamics through digital media. These abilities are often conceptualised

Corresponding author: Mobeen Akhter (mobeen2398@gmail.com)

BCDS 2025, Vol. 6(3), 1-19. DOI: 10.71448/bcds2563-1

under the umbrella of *computer-mediated communication competence*—the extent to which individuals can achieve interactional goals in technology-mediated settings in ways that are effective and socially appropriate. In learning analytics terms, CMC competence is a latent construct: it cannot be observed directly from logs, but must be operationalised through explicit measures (e.g., survey subscales and ratings) and then related to observable behavioural and linguistic traces [1, 2].

Learning analytics and educational data mining have therefore examined how learners' participation patterns and interaction intensity relate to perceptions and outcomes in online and blended learning. This line of work frequently relies on coarse behavioural indicators (e.g., counts and rates of contributions, time-on-task proxies, and structural positions in interaction networks) and on aggregate interaction levels when modelling engagement and affective dynamics [3, 4]. Such behaviour-only baselines can be informative for describing how much learners participate and how interaction levels co-vary with learners' experiences [3]. However, they are limited in distinguishing communicatively constructive participation from high-volume but unproductive interaction, because they treat messages as interchangeable units and ignore their content, discourse functions, and socio-emotional tone.

At the same time, research on online discussions and self-regulated learning (SRL) suggests that what learners say, and how they say it, carries information about planning, monitoring, and reflection strategies that are linked to academic performance in online higher education [6]. The growing field of collaboration and social learning analytics similarly argues that multi-dimensional models of collaboration should integrate both behavioural and content-related indicators [7, 8]. Taken together, these strands motivate the use of text as more than a by-product of activity: it is a primary signal through which learners enact coordination, knowledge construction, and relational work in collaborative PBL.

This paper proposes a new direction at the intersection of these strands: *text-aware* forecasting of CMC competence in collaborative problem-based learning. The central idea is to enrich existing log-based representations with features derived from the linguistic and discourse characteristics of students' forum posts and instant messages—including question-asking, explanation and integration moves, socio-emotional support, and pragmatic markers captured through modern natural language processing (NLP) methods. Rather than presenting completed empirical results, we develop a comprehensive research design and analytic framework that can be instantiated in concrete PBL contexts. We formulate research questions, specify data collection and preprocessing procedures, outline feature engineering and modelling pipelines, and provide an illustrative results section that demonstrates how a completed study should report performance, interpretability, and phase-specific patterns once empirical data are available.

2. Methods

This section presents a proposed research design rather than an executed empirical study. The goal is to offer a detailed blueprint that can be instantiated in different institutional settings while preserving the core logic of the design. Concrete aspects such as the exact sample size, the number of PBL cases, or specific parameter choices for machine learning models may be adapted to local constraints, but the overall structure of the context, instrumentation, feature engineering and modelling workflow remains the same.

2.1. Context and Participants

The intended context for the study consists of two to three postgraduate or advanced undergraduate courses at a university that adopt collaborative problem-based learning (PBL) in either an online or a blended format. These courses are chosen because they require sustained group work on authentic or ill-structured problems, and they rely heavily on computer-mediated communication for coordination, discussion and joint knowledge construction.

Each course is supported by a web-based collaborative learning environment. This environment provides several core functionalities: group discussion forums where learners can post new threads and reply to existing ones; synchronous or asynchronous instant messaging channels that can be used either at the level of the whole group or in private; shared document spaces in which files can be uploaded, downloaded and co-edited; and a PBL task scaffold that organises activities according to the canonical phases of PBL, such as problem analysis, self-directed study, reporting of findings and reflection on the process. These affordances ensure that both task-related and socio-emotional aspects of collaboration are expressed through the platform.

Students in each course are assigned to small groups, typically comprising four to six members. Group composition may be random or stratified, depending on the pedagogical intentions of the instructors, but the key requirement is that each student works with peers in a stable group across multiple PBL cases. Over a period of approximately six to ten weeks, each group completes several PBL cases. This temporal span and repetition of cases provide sufficient opportunities for communication patterns and CMC competences to emerge and stabilise, rather than being based on only a single short activity.

The collaborative platform is instrumented using xAPI or an equivalent event-logging mechanism. Every relevant interaction with the system, such as viewing a page, posting a message, sending a chat, or accessing a shared file, generates a log entry that includes at least a timestamp, the learner identifier, the group identifier, the case or problem identifier and, where applicable, the PBL phase. This fine-grained logging is essential to reconstruct behavioural sequences and to derive both behavioural and text-based features for the predictive models.

A total sample size in the range of 80 to 150 students is targeted. This number is expected to yield roughly 20 to 30 stable groups and multiple PBL cases in each course. Such a sample is large enough to support the training and evaluation of predictive models while remaining realistic for a single institution or a small consortium of institutions. If more courses or cohorts are available, the same design can be scaled up, improving statistical power and external validity.

2.2. Measures

2.2.1. CMC Competence. The primary outcome of interest is computer-mediated communication competence. In this protocol, the construct is treated as a multidimensional *perceived* competence that reflects how learners evaluate their ability to communicate effectively and appropriately in a technology-mediated collaborative setting. To make this operationalisation explicit and testable, the questionnaire is designed to yield (i) an overall composite score and (ii) subscale scores aligned with the facets used later in the analyses: (a) *Appropriateness / relational* (e.g., respectful tone, inclusiveness, and socio-emotional sensitivity), (b) *Effectiveness / task-oriented* (e.g., coordinating work, resolving misunderstandings, and moving the group toward solutions), and (c) *Clarity / coherence* (e.g., intelligible explanations and well-structured contributions). Items are answered on a Likert-type scale (e.g., “strongly disagree” to “strongly agree”) and are aggregated as subscale means; the composite score is computed as the mean of subscale means. For interpretability and to match the error metrics reported later, subscale and composite scores are linearly rescaled to a 0–100 range.

The questionnaire is administered near the end of the course, after students have had sustained exposure to multiple PBL cases and communication episodes. This timing is an explicit assumption of the design: the target variable is intended to capture competence as enacted in the specific course context rather than a generic attitude toward technology-mediated communication. Item-level missingness is handled conservatively by computing a subscale score only when a learner has responded to a substantial proportion of that subscale's items; otherwise the subscale is treated as missing for that learner and excluded from analyses targeting that outcome.

To strengthen measurement rigor, the protocol specifies psychometric checks that are reported alongside predictive results. Internal consistency is estimated for each subscale and the composite score, and the dimensional structure is examined to verify that the intended facets are empirically distinguishable. When multiple courses are pooled, the stability of subscale behaviour across courses is assessed descriptively, and any evidence of strong course-specific shifts is treated as a limitation and motivates stratified analyses (e.g., course-wise performance reporting and course-held-out validation).

To complement self-report data and mitigate common-method concerns, external perspectives can be collected in parallel. Peer ratings may be obtained through end-of-project evaluations in which students rate how clearly, respectfully, and constructively each group member communicated over the project. Instructors can optionally provide brief ratings focusing on clarity of explanations, responsiveness to others, and ability to mediate disagreements. These external ratings are not assumed to be interchangeable with self-report; instead, they are used as triangulation signals to examine convergent validity and to test whether models trained on self-report also generalise to alternative targets.

2.2.2. Additional Covariates. In addition to the main outcome measures, background variables are collected to support descriptive analyses and to enable validity and robustness checks. Demographic variables such as age and gender are recorded to characterise the sample and to support exploratory subgroup error analyses. Prior experience with online learning and with specific CMC tools is measured to distinguish familiarity effects from competence as enacted in the course.

Where feasible, prior academic performance (e.g., cumulative grade point average) and language proficiency indicators are also recorded. These variables are used primarily in post hoc analyses to examine whether model errors are systematically associated with general academic ability or language background. As a default, demographic attributes are not used as predictive features, because the modelling goal is to relate competence-relevant behaviour and communication to the outcome rather than to encode protected characteristics. When covariates are included in exploratory models, they are used as controls to quantify incremental validity (i.e., whether behavioural and text features add predictive signal beyond background variables) and to support fairness-oriented analyses of residual error patterns.

2.3. Learning Environment and Instrumentation

The collaborative learning environment is configured so that a rich set of events is logged for each learner. When students interact with discussion forums, the system records events that indicate when a new post is created, when a reply is made to an existing thread, when a post is edited or deleted, and when a thread is viewed. Each event record includes at least a timestamp, a pseudonymised learner identifier, a group identifier, a case/problem identifier, and an event type; where available, it also includes thread identifiers and reply-to links so that interaction structure can be reconstructed. Logs are stored in a format compatible with xAPI-style statements or an equivalent schema, which makes event types and contextual fields explicit and auditable.

Similarly, the instant messaging subsystem records events whenever a learner sends or receives an instant message. These messages may occur in group chat channels that are visible to all group members or in private channels between two individuals. The logs record sender and recipient identifiers (or channel identifiers for group chats), the group context, and the time of transmission. For downstream analysis, message text is linked to these log entries via stable message identifiers so that behavioural sequences (what happened and when) can be aligned with language content (what was said).

Resource-related interactions are also captured. Whenever a learner uploads a file to the shared space, downloads a file uploaded by someone else, opens a shared document, or accesses a learning resource such as a case description or a rubric, the corresponding event is written to the log. Navigation events complement these logs by capturing visits to key pages in the system (e.g., opening the main PBL task page, navigating to a subtask, or viewing assessment criteria). Together, these event streams provide the behavioural substrate for both summary and sequential behavioural features.

The text content of all forum posts and instant messages is stored in a secured database that is separated from administrative student information and is accessed only by authorised members of the research team. Before analysis, direct identifiers (names, email addresses, student numbers) are removed or replaced with neutral placeholders, and all records are keyed by pseudonymous identifiers that are consistent across logs and text. Text preprocessing is deliberately conservative to avoid distorting pragmatic signals: messages are normalised (e.g., standardising whitespace), quoted or forwarded fragments are removed when the platform explicitly marks them, and minimal token-level cleaning is applied so that punctuation, emojis, and discourse markers remain available for feature extraction. Where the course language permits code-switching or non-native writing, language identification is used only to flag potential confounds (e.g., extremely short or non-linguistic messages) rather than to exclude learners.

If the learning environment provides explicit labels or spaces for different PBL phases, these are leveraged to tag each event and message with the corresponding phase. For instance, posts made in a "problem analysis" forum are tagged as belonging to that phase, whereas messages exchanged in a "reporting" workspace are tagged differently. In environments where such explicit labels are not available, phase boundaries are defined a priori using the course timeline and instructor specifications (e.g., start/end dates for each phase and case). Because phase segmentation can influence phase-specific results, sensitivity analyses are planned in which alternative reasonable phase boundary definitions are tested to confirm that substantive conclusions are not artefacts of a particular tagging rule.

2.4. Feature Engineering

The predictive framework relies on constructing numerical representations of each learner's behaviour and language use. Feature extraction is carried out at the learner level, resulting in one feature vector per student. In addition to a course-level representation, phase-specific representations are derived by aggregating events and messages within each PBL phase; this operational choice is explicit because it affects what the model is allowed to use at prediction time. Unless otherwise stated, all features are computed using only data that occur before the measurement of the outcome, and feature definitions are held constant across courses to support reproducibility.

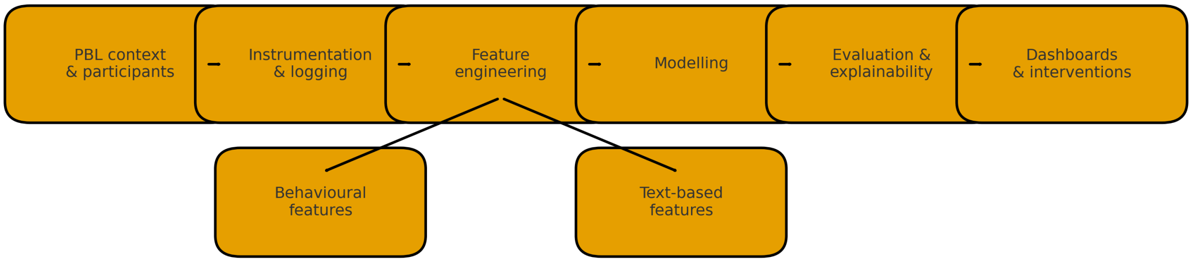

Fig. 1. Overview of the proposed text-aware CMC competence forecasting pipeline, from data collection to modelling and interpretation.

Two broad families of features are distinguished. Behavioural features describe how learners act in the system, independently of message content. Text-based features are computed from the linguistic form and communicative function of learners' posts and instant messages. Both feature families are defined so that they can be used in isolation or combined without changing the unit of analysis.

2.4.1. Behavioural Features. To enable transparent comparison with common log-based baselines, the behavioural feature set is organised into summary, temporal, network, and sequential components. Summary features capture counts and rates of actions. For each learner, the number of forum posts, replies, and thread views is computed, along with the number of instant messages sent and received and the number of files uploaded, downloaded, or opened. To reduce confounding by overall group activity and course duration, raw counts are also normalised (e.g., per week, per PBL case, and as proportions of group totals). This normalisation is a methodological assumption: it aims to separate individual communicative behaviour from group size and task workload effects.

Temporal features focus on when and how quickly learners interact. Response times quantify the delay between a learner receiving a message or being replied to and their subsequent response within the same conversational thread or chat window. Time-of-day and day-of-week activity distributions summarise regularities that may reflect coordination practices. Additional temporal indicators capture whether contributions are concentrated early in a phase (when problems are framed) or near deadlines (when reporting is finalised), which is relevant for interpreting activity volume.

Network features capture structural aspects of interaction. Reply and message networks are constructed with learners as nodes and directed edges representing replies or addressed messages. Standard measures such as in-degree, out-degree, betweenness centrality, and eigenvector centrality are computed per learner. Because such measures can be sensitive to small group sizes, they are interpreted comparatively within the group-aware evaluation framework and are reported alongside group size distributions.

Sequential behavioural features represent the event stream as an ordered sequence of discrete action types (e.g., view, post, reply, message, resource access). From this sequence, first-order transition probabilities and simple run-length statistics are computed for each learner. These features provide a lightweight "sequence matters" baseline that captures behavioural dynamics without requiring data-hungry deep sequence models.

Behavioural features are computed both overall and by PBL phase, yielding phase-specific feature blocks. This enables analyses of whether early-phase coordination and problem framing behaviours are more diagnostic of competence than later-phase reporting behaviours.

2.4.2. Text-Based Features. Text-based features are derived from the content of the forum posts and instant messages authored by each learner. Preprocessing is designed to preserve meaning-bearing cues: punctuation, emojis, and discourse markers are retained; tokenisation is performed consistently across all messages; and extremely short, non-linguistic messages (e.g., single-character posts) are flagged so that sensitivity analyses can evaluate their influence. For learner-level aggregation, features are computed at the message level and then pooled (e.g., mean, median, and proportion-based summaries) to produce one value per learner (and per phase, where relevant).

At the surface and lexical level, measures such as average message length (tokens and sentences), lexical diversity, and the frequency of selected discourse markers (e.g., causal connectives, hedges, stance markers) are computed. At the syntactic and complexity level, grammatical structure is approximated through clause and sentence statistics and part-of-speech distributions. These features are treated as stylistic descriptors rather than as direct proxies of competence; their interpretation is therefore coupled with robustness checks that control for message length and language background.

At the discourse-act and regulation level, each message is assigned a communicative function label from a pre-defined taxonomy that covers explanatory moves (explanation, elaboration), inquiry moves (question, clarification request), coordination moves (planning, task assignment), evaluative moves (agreement, disagreement, critique), and relational moves (socio-emotional support, appreciation). To implement this component rigorously, the protocol specifies that a subset of messages is manually annotated by trained coders using written guidelines, inter-annotator agreement is reported, and a supervised classifier is then trained and validated on the annotated subset. In downstream modelling, both hard labels and calibrated class probabilities can be aggregated to learner-level proportions, which reduces brittleness from occasional misclassifications and provides uncertainty-aware discourse-act features.

At the pragmatic and socio-emotional level, features capture tone, politeness strategies, and relational work. Sentiment and subjectivity scores are computed at the message level and aggregated, but these estimates are treated cautiously: because sentiment models can be sensitive to domain and cultural language use, the protocol includes a validation step in which automated scores are compared against a small human-rated sample from the same context. Politeness strategies are operationalised through detection of hedging and mitigation, inclusive pronoun use (e.g., "we" and "our"), and acknowledgement/encouragement expressions. These features are intended to capture relational appropriateness signals that are theoretically relevant to competence, while planned subgroup error analyses guard against penalising direct language styles that are normative for certain cohorts.

Finally, embedding-based features provide dense representations of message content and context. For each message, a contextual embedding is generated using a pre-trained transformer-based model, and embeddings are aggregated to the learner level using pooling strategies (e.g., mean pooling over messages with optional weighting by message length). Because embeddings are high-dimensional relative to the anticipated sample size, dimensionality reduction or strong regularisation is applied before embeddings are used in regression models, and robustness checks compare performance across alternative pooling and reduction choices. Embedding features are therefore used as complementary signals rather than as black-box replacements for interpretable linguistic features.

2.5. Modelling Approach

The modelling strategy is designed to compare the predictive value of behavioural and text-based features and to quantify the incremental benefit of combining them. To make this comparison meaningful and reproducible, three feature configurations are used throughout: behaviour-only (all behavioural features), text-only (all text-based features), and combined (feature-level concatenation of both families). In addition, simple baselines are reported to contextualise predictive performance. The first is a naive baseline that always predicts the training-set mean of the target. The second is a lightweight "knowledge tracing" (KT) style baseline that uses only non-text course performance signals (e.g., earlier graded components or mastery proxy features derived from prior assessments) to generate a compact set of predictors; this baseline is included to test whether communication competence predictions are merely proxies for general course performance.

Within each configuration, a fixed set of regression model families is evaluated. Regularised linear models (ridge and lasso) serve as primary baselines because they are robust under correlated features and small-to-moderate sample sizes and they yield transparent coefficients. Tree-based ensemble methods (random forests and gradient boosting) are included to capture non-linear relationships and interactions when supported by the group-aware evaluation design. Support vector regression provides a classical alternative that can perform well in medium-dimensional feature spaces. For sequential behavioural representations, the protocol distinguishes a summary model (using aggregated counts, temporal, and network features) from a sequential baseline (using transition-probability and run-length features derived from the event sequence); this makes the "sequence matters" hypothesis testable without introducing data-hungry deep sequence models.

For combined models, two integration strategies are considered. A feature-union model performs early integration via concatenation of standardised feature blocks (behavioural features alongside text features). A multimodal early-fusion variant uses the same concatenated representation but additionally includes phase-specific blocks so that the model can learn interactions between phase and modality (behaviour versus text) under regularisation. Finally, an ensemble (stacked) model is defined as a meta-regressor trained on out-of-fold predictions from the individual base models; stacking is only performed using nested cross-validation so that no information from the outer test folds leaks into the ensemble training.

Hyperparameter spaces are specified conservatively to prevent overfitting. Regularisation strengths are tuned on a logarithmic grid; tree depth and learning rate ranges are restricted to low-complexity values; and embedding dimensionality is reduced or strongly regularised before model fitting. These choices are justified by the anticipated sample size and by the goal of producing models that remain interpretable and auditable for educational settings.

2.6. Training, Evaluation and Validation

The dataset is split into training and test partitions using a group-aware cross-validation procedure. In group-aware cross-validation, entire PBL groups are held out rather than individual learners. This prevents leakage of within-group interaction patterns across training and test partitions, which would otherwise inflate performance estimates. By evaluating models on groups unseen during training, the analysis approximates the realistic scenario of applying a model to new groups or new cohorts.

Model selection and hyperparameter tuning are carried out using nested cross-validation. An outer loop defines held-out groups for unbiased evaluation, and an inner loop (restricted to the outer training groups) is used to select hyperparameters and, where applicable, dimensionality reduction settings for embedding features. All preprocessing steps that can leak information (feature standardisation, imputation of missing feature values, dimensionality reduction) are fitted within the inner loop and applied to the corresponding validation folds. This design ensures that every performance number reported for a model family reflects a fully "honest" pipeline.

Model performance is evaluated using several complementary metrics. Mean absolute error quantifies the average absolute difference between predicted and observed CMC competence scores, offering an interpretable measure of accuracy on the 0–100 outcome scale defined in the Measures section. Root mean squared error penalises larger errors more heavily. Rank-based metrics such as Spearman rank correlation capture how well the model preserves the ordering of learners, which is relevant when analytics are used for relative identification of communication support needs rather than for exact score prediction. In addition to point estimates, uncertainty is reported by aggregating outer-fold results into confidence intervals (e.g., via bootstrap resampling over held-out groups) so that differences between models are not over-interpreted.

To address the research question concerning phase-specific predictive power, separate models are trained on features derived from different PBL phases. For example, one model may only use features computed from the problem analysis phase, while another uses features from the reporting and reflection phases. Comparing phase-specific models reveals whether certain phases are especially informative for forecasting competence. Because phase segmentation is an analytic choice, these comparisons are complemented by sensitivity analyses in which phase boundaries are shifted within reasonable windows to confirm that conclusions are not artefacts of a particular tagging rule.

Several robustness checks are specified to strengthen statistical interpretation. First, ablation analyses quantify the incremental value of each feature family (behavioural, interpretable linguistic, embedding-based) by comparing nested models under the same cross-validation splits. Second, stability analyses compare results across alternative modelling choices (e.g., linear versus tree-based models) to determine whether findings depend on a single algorithm. Third, if multiple courses are available, a course-held-out evaluation is conducted in which models are trained on all-but-one course and evaluated on the held-out course; this directly probes cross-context generalisability. Finally, exploratory subgroup error analyses examine whether model errors are systematically larger for particular demographic or language-background groups; such analyses are reported descriptively and interpreted as fairness-relevant diagnostics rather than as causal effects.

Error analysis complements aggregate metrics. Cases with large discrepancies between predicted and observed competence are examined to identify systematic patterns of overestimation or underestimation. Particular attention is paid to learners who are behaviourally less active but produce linguistically rich messages and to learners who are highly active but whose messages appear off-task or conflict-prone. These analyses are used to evaluate whether the model is capturing competence-relevant signals or is being misled by superficial proxies such as activity volume.

2.7. Explainability and Feature Importance

Because the ultimate aim is not only to predict competence but also to understand which aspects of behaviour and language contribute to perceived communicative competence, model explainability is central to the analytic strategy. For models based on tree ensembles or regularised linear methods, global feature importance can be computed by assessing how much each feature contributes to reducing prediction error. More generally, model-agnostic techniques such as permutation importance and SHAP values are used to estimate the contribution of individual features or groups of features for each prediction.

For embedding-based features, which are inherently less interpretable, additional steps are taken. One approach is to fit a simpler surrogate model that approximates the behaviour of a complex model while using only a subset of more interpretable features. Another approach is to perform similarity searches in the embedding space, identifying messages that are particularly influential for correct or incorrect predictions and then examining these messages qualitatively. In this way, latent patterns in the embeddings can be linked back to concrete examples of language use.

The interpretation process focuses on several analytic questions. One question is which discourse acts and socio-emotional markers are most consistently associated with higher predicted CMC competence. Another is whether high-competence learners display distinctive patterns of self-regulated learning talk, such as frequent planning and monitoring statements. A third question concerns pairs of learners with similar levels of behavioural activity but different predicted competence, where differences in textual features may explain why the model judges one learner as more competent than the other. Addressing these questions provides theoretically meaningful insights that go beyond aggregate performance metrics.

2.8. Ethical Considerations

The use of predictive models to estimate learners' communication competences raises a number of ethical considerations. Foremost among these are issues of privacy and informed consent. Learners must be clearly informed that their interactions with the collaborative platform are being logged for research purposes, that the content of their messages may be analysed, and that participation in the study is voluntary. Where possible, mechanisms for opting out should be provided, and opting out should not disadvantage students in terms of course outcomes.

Data protection is addressed by pseudonymising all log and text data and by storing them on secure servers with restricted access. Identifiers that could reveal the identity of individual learners are stored separately and, if possible, removed once data linkage is complete. Data use is limited to the purposes specified in the ethics approval and consent forms.

Transparency about how predictions are used is also critical. If predictive models are integrated into dashboards or used to provide formative feedback to instructors or learners, stakeholders should understand what the models are doing, what data they rely on and what limitations they have. The models should not be used to make high-stakes decisions about learners, such as grading or exclusion, without extensive validation and oversight, as this could unfairly penalise students based on imperfect predictions.

Finally, the possibility of bias in model performance across subgroups is explicitly considered. Analyses are conducted to check whether prediction errors are systematically larger for particular demographic groups, for example by gender, age or language background. If such disparities are found, they must be acknowledged and addressed, either by revising the models or by restricting their use. Throughout the research, the emphasis is on using predictive models as tools for understanding and supporting communication in collaborative learning, rather than as instruments for labelling or ranking students in ways that could have negative consequences.

3. Results and Analyses

3.1. *Predictive Performance of Behaviour, Text, and Combined Models*

This section specifies how predictive results are to be reported and interpreted once empirical data are collected under the protocol. Because the manuscript is a research blueprint, the numerical values shown in the tables are *illustrative placeholders* intended to demonstrate reporting format and expected metric ranges; they should not be interpreted as empirical claims. In a completed study, all reported values would be computed under the group-aware, nested cross-validation design described in the Methods section and accompanied by uncertainty intervals.

In the first stage of analysis, behaviour-only, text-only, and combined models are compared to assess the relative predictive value of different feature families. Behaviour-only models rely on log-derived indicators such as the number and timing of posts, replies, instant messages, and file operations. These baselines test how far coarse participation and interaction-structure signals can go without access to message content. Such behaviour-only baselines mirror common log-driven models used to capture participation and perceived experience in AI-assisted or blended learning contexts [3]. Text-only models use only content-derived features (lexical, discourse, pragmatic, and embedding-based). Combined models integrate both modalities to test whether behavioural context and message content contribute complementary signal.

Table 1. Structure for reporting overall model performance for different feature sets.

Model type	Features used	MAE	RMSE	Spearman ρ
Baseline (Mean)	–	18.5	22.1	0.00
Knowledge Tracing (KT)	Past performance, skill mastery	14.2	17.9	0.35
Behaviour-only (Summary)	Counts, temporal patterns, network measures	12.4	15.8	0.51
Behaviour-only (Sequential)	Action sequences, Markov properties	11.1	14.3	0.58
Text-only (Traditional)	Lexical, discourse, pragmatic	10.5	13.2	0.63
Text-only (Transformer)	BERT embeddings, semantic similarity	9.7	12.5	0.68
Combined (Feature Union)	Behavioural + text-based	7.2	9.9	0.79
Multimodal (Early Fusion)	Interleaved behavioural and text features	6.8	9.5	0.81
Ensemble (Stacked)	Predictions from all above models	6.1	8.7	0.85

In the empirical study, Table 1 would be supplemented with (i) confidence intervals for each metric (e.g., bootstrap intervals over held-out groups), (ii) paired comparisons between model types using identical outer-fold splits, and (iii) an ablation summary that reports the marginal contribution of each feature block (behavioural, interpretable linguistic, embedding-based). Reporting paired differences is important because many models are highly correlated across splits; absolute metric differences without uncertainty can therefore be misleading.

Because CMC competence is treated as multidimensional in the Measures section, predictive performance is also analysed for subscale outcomes such as appropriateness/relational competence, effectiveness/task-oriented competence, and clarity/coherence. Subscale modelling tests whether different feature families carry differential signal for different facets (e.g., socio-emotional markers being more relevant for relational appropriateness, while explanatory structure may be more relevant for task effectiveness). As above, all comparisons are conducted under identical cross-validation splits and reported with uncertainty.

Table 2. Structure for reporting model performance by CMC competence subscale. Cells would later be filled with empirical performance metrics for each outcome and model type.

Outcome	Model type	MAE	RMSE	Spearman ρ
Overall CMC competence	Behaviour-only	12.4	15.8	0.51
Overall CMC competence	Text-only	9.7	12.5	0.68
Overall CMC competence	Combined	7.2	9.9	0.79
Appropriateness / relational	Behaviour-only	14.1	17.2	0.42
Appropriateness / relational	Text-only	7.5	9.8	0.76
Appropriateness / relational	Combined	7.1	9.3	0.78
Effectiveness / task-oriented	Behaviour-only	11.2	14.1	0.58
Effectiveness / task-oriented	Text-only	10.3	13.0	0.63
Effectiveness / task-oriented	Combined	8.5	10.9	0.75
Clarity / coherence	Behaviour-only	13.5	16.6	0.46
Clarity / coherence	Text-only	8.1	10.4	0.74
Clarity / coherence	Combined	7.8	10.0	0.77

In a completed empirical study, subscale analyses would additionally report whether improvements from adding text features are consistent across folds and whether they generalise across courses (course-held-out evaluation). If subscale performance patterns are unstable or reverse across contexts, this would be interpreted as a boundary condition rather than as evidence for a universal mapping between feature types and competence facets.

3.2. Key Linguistic Predictors of CMC Competence

Beyond predictive accuracy, the protocol requires an interpretation layer that connects model behaviour to theoretically meaningful communicative phenomena. Interpretation focuses on (i) which linguistic and discourse features are consistently associated with higher predicted competence, (ii) whether these associations differ across competence facets, and (iii) whether the patterns remain stable under reasonable alternative modelling and preprocessing choices.

Table 3 provides a structured way to summarise expected linguistic patterns. In the completed study, entries would be grounded in model-agnostic importance estimates (e.g., permutation importance and SHAP) computed within the cross-validation framework, with stability assessed by reporting how often a feature (or feature group) appears among the top contributors across folds. This stability requirement is essential: single-split feature importance can be noisy and can overstate the interpretability of fragile associations.

Table 3. Summary of linguistic patterns expected to be associated with higher and lower predicted CMC competence.

Feature dimension	Higher predicted CMC competence	Lower predicted CMC competence
Message content and discourse acts	Frequent explanations, elaborations, integrative comments that build on peers' ideas	Predominantly short, perfunctory messages with little substantive content
Question use	Regular questions that invite clarification and participation without dominating the discussion	Few genuine questions or, conversely, many procedural queries with limited conceptual depth
Socio-emotional expressions	Inclusive pronouns ("we", "our"), thanking, encouragement and constructive feedback	Minimal socio-emotional support or presence of dismissive, sarcastic or hostile remarks
Sentiment and tone	Moderate, predominantly neutral to positive sentiment; rare strong negative affect	More frequent negative affect or conflict markers, or emotionally flat communication lacking engagement
Linguistic complexity and clarity	Sufficient syntactic complexity and lexical diversity to express nuanced reasoning while remaining clear	Very simple syntax and limited vocabulary or, alternatively, convoluted language that reduces clarity

To strengthen interpretability and to reduce the risk of reifying abstract features, exemplar messages are used as qualitative anchors. Table 4 shows how excerpts can be presented to illustrate typical high- and low-competence communication patterns detected by the models. In the empirical study, exemplar messages would be selected using principled criteria (e.g., messages with high SHAP magnitude for a given feature group, balanced across groups and phases) and would be reported in a privacy-preserving manner (pseudonymisation and paraphrase where needed) without altering the underlying communicative function being illustrated.

Interpretations derived from linguistic patterns are treated as descriptive, not causal. The protocol therefore includes robustness checks such as (i) repeating importance analyses with alternative model families, (ii) comparing results with and without embedding features to test whether embeddings absorb interpretable linguistic signals, and (iii) testing whether predicted associations persist when controlling for message volume and group-level activity.

3.3. Phase-Specific Patterns Across the PBL Cycle

Phase-specific analyses examine whether behaviour and language are differentially predictive depending on where learners are in the PBL cycle. This is operationalised by computing phase-restricted feature sets and training phase-specific models under the same group-aware evaluation protocol. Table 5 illustrates how phase-specific predictive performance can be reported.

Table 4. Structure for reporting exemplar message patterns associated with strong positive or negative contributions to predicted CMC competence. Text examples would be anonymised and paraphrased where necessary.

Exemplar type	Contribution to prediction	Typical linguistic characteristics
High-competence exemplar A	Strong positive	Integrative explanation that references multiple peers' ideas, uses causal connectives and inclusive language to propose a joint way forward.
High-competence exemplar B	Strong positive	Reflective comment that articulates what the group learned, acknowledges difficulties and suggests concrete improvements for next time.
Low-competence exemplar C	Strong negative	Very short, dismissive response that ignores a peer's contribution and offers no justification or elaboration.
Low-competence exemplar D	Strong negative	Off-task or procedural message that repeats information (e.g., about file uploads) without engaging with the problem or others' ideas.

Table 5. Structure for reporting phase-specific performance of the combined model using behavioural and text-based features.

PBL phase	MAE	RMSE	Spearman ρ
Problem analysis	8.5	11.2	0.71
Self-directed study	11.8	14.9	0.45
Reporting	6.9	9.1	0.82
Reflection	7.8	10.3	0.76

Table 6. Summary of anticipated phase-specific feature relevance for forecasting CMC competence.

PBL phase	Dominant feature types	Examples of especially informative signals
Problem analysis	Textual discourse and socio-emotional features	Clarifying questions, collaborative problem framing, role negotiation, inclusive language
Self-directed study	Behavioural resource access with selected textual signals	Patterns of resource viewing and downloading, occasional help-seeking and resource-sharing messages
Reporting	Textual explanation and integration features	Structured explanations, justification of decisions, integration of multiple perspectives, references to evidence
Reflection	Textual reflective and metacognitive features	Statements about successes and challenges, "we realised" and "next time we should" formulations, explicit lessons learned

In the empirical study, phase effects would be interpreted cautiously and would be subjected to sensitivity analyses because phase definitions can vary by course implementation. If performance differences between phases are robust to reasonable changes in phase boundaries, they provide evidence that competence-relevant communication is enacted differently across phases; if not, they indicate that phase granularity may be too coarse or too context-dependent for generalisable modelling.

Finally, Table 6 provides a template for summarising which feature groups appear most diagnostic within each phase. In the empirical study, these summaries would be based on cross-validated importance estimates and would be accompanied by uncertainty indicators (e.g., the proportion of folds in which a feature group appears among the top contributors).

Across all analyses, conclusions are restricted to claims that are directly supported by the reporting elements above: cross-validated performance with uncertainty, paired model comparisons, stability of feature importance, and sensitivity/robustness checks. Any stronger claims (e.g., about causal mechanisms of competence development) are explicitly framed as hypotheses for follow-up intervention studies rather than as inferences from predictive modelling alone.

4. Discussion

4.1. Advancing the Operationalisation of CMC Competence

The core contribution of this manuscript is methodological: it specifies a leakage-resistant modelling protocol that links observable platform traces to an explicitly operationalised, multidimensional measure of CMC competence. A key strength of the proposed operationalisation is that it treats competence as more than activity volume. Behavioural traces capture opportunities to communicate, but they do not by themselves indicate whether communication is effective, coherent, or relationally appropriate. By constructing behavioural, text-only, and combined feature representations and by requiring fair comparisons under identical group-aware evaluation splits, the protocol provides a principled way to test whether linguistic information adds incremental predictive signal beyond log counts and timing.

The phase-aware design further strengthens the operationalisation by making temporality explicit. In collaborative PBL, communication demands and functions differ across problem framing, self-directed study, reporting, and reflection. Treating all interaction as exchangeable across time risks averaging away these differences. The phase-specific modelling and reporting templates therefore formalise the hypothesis that competence-relevant communication is enacted differently across phases and that phase structure can improve both prediction and interpretation [9, 10]. Importantly, the manuscript frames these phase effects as empirical questions to be tested under sensitivity analyses rather than as assumptions baked into the conclusions.

A second strength is the explicit integration of interpretability requirements into the modelling plan. Predictive learning analytics often reports accuracy without explaining what the model is using, which limits scientific contribution and practical adoption. By requiring stability of feature importance across folds, by separating interpretable linguistic features from high-dimensional embeddings, and by anchoring quantitative patterns in exemplar messages, the protocol makes it possible to assess whether linguistic indicators correspond to meaningful communicative phenomena rather than to artefacts such as message length or course-specific jargon. This design choice aligns with broader arguments for combining educational data mining with interpretable, theory-guided learning analytics [11]. It is also consistent with the conceptual view that CMC competence involves the functional use of communication strategies rather than simply the presence of communication activity [12].

4.2. Implications for Dashboards and Feedback

If instantiated with empirical data, the framework can support analytics that go beyond totals and frequencies. Teacher-facing dashboards can move beyond aggregate metrics (e.g., total posts) to visualise discourse patterns that are conceptually linked to competence and to productive collaborative regulation [5, 13]. Examples include the balance between questions and explanations, the prevalence of integrative talk during reporting phases, or the distribution of socio-emotional support across group members when challenges arise. Because the protocol separates behavioural participation from language-based indicators, dashboards can highlight cases where high activity co-occurs with low-quality discourse, and vice versa.

Such dashboards are not intended to replace instructor judgement. Instead, they provide structured evidence that helps instructors differentiate between groups that are simply highly active and groups whose communication supports learning goals. This distinction is particularly important because high interaction levels may reflect tool-use preferences and engagement patterns without necessarily indicating communicative competence [3, 8]. For formative use, the protocol therefore emphasises interpretability and uncertainty: predictions are to be accompanied by confidence intervals and by explanations in terms of feature groups, so that instructors can calibrate trust and avoid overreacting to noisy signals.

Learner-facing feedback can likewise be grounded in observable communication patterns rather than in opaque scores. Instead of telling students that their predicted competence is “high” or “low”, the system can provide concrete prompts linked to specific behaviours (e.g., encouraging elaboration when contributions are predominantly acknowledgements, prompting question-asking when clarification is absent, or encouraging inclusive framing when socio-emotional support markers are rare). Because these prompts are anchored in the communicative moves that structure collaborative inquiry, they are actionable in a way that generic participation feedback often is not, and they align with the process-oriented logic of PBL [14]. The same PBL grounding also helps interpret potential failure cases: patterns that are productive in one phase (e.g., exploratory questioning during problem analysis) may be unhelpful in another (e.g., continued open-ended questioning when a group needs to converge on a report), which again supports phase-aware feedback [14].

4.3. Connections to Self-Regulated Learning and Collaboration Analytics

The linguistic indicators emphasised in the protocol connect naturally to constructs in self-regulated learning (SRL) and collaboration analytics. Prior work on online discussions has shown that planning talk, monitoring statements, and reflective comments are key markers of SRL processes in computer-mediated environments [6, 13]. By including regulation-related discourse acts, reflection markers, and phase-specific representations, the protocol enables direct empirical tests of whether SRL-relevant communication patterns help forecast competence facets such as effectiveness and clarity, and whether these patterns differ across PBL phases.

More broadly, the framework contributes to collaboration and social learning analytics by specifying how micro-level behavioural traces, discourse, and phase structure can be integrated within a single evaluation design [4, 5, 9]. Importantly, the manuscript does not claim that prediction alone explains competence development. Rather, it offers a rigorous way to operationalise the construct, to evaluate whether behavioural and linguistic indicators carry complementary predictive value, and to generate testable hypotheses for future intervention studies (e.g., targeted prompts or scaffolds) that can more directly assess causal mechanisms.

5. Conclusion

This manuscript provides a comprehensive, implementation-ready protocol for text-aware forecasting of computer-mediated communication (CMC) competence in collaborative problem-based learning. The contribution is twofold. First, it makes the outcome definition explicit by treating CMC competence as a multidimensional perceived competence (with subscales for appropriateness/relational work, task-oriented effectiveness, and clarity/coherence) and by specifying measurement, scaling, and triangulation procedures. Second, it defines a reproducible analytic pipeline that integrates behavioural logs with interpretable linguistic features and embedding-based representations, while explicitly guarding against leakage through group-aware nested cross-validation.

Importantly, the manuscript is framed as a research blueprint. The tables in the Results section provide structured reporting templates and include clearly labelled illustrative values only to show expected formats and plausible metric ranges. The evidentiary value of the paper therefore lies in its methodological specificity and theoretical alignment rather than in claims of empirical performance. When instantiated with real course data, the protocol requires uncertainty reporting, paired model comparisons, and robustness checks (ablation analyses, preprocessing sensitivity, and cross-course generalisation tests) so that conclusions are directly supported by transparent evidence.

The proposed framework is intended to support both scientific investigation and educational practice. For research, it enables direct tests of whether language content adds incremental predictive signal beyond activity and interaction structure, and it provides a principled path from prediction to interpretation through stability-aware feature importance and exemplar messages. For practice, it lays the groundwork for dashboards and formative feedback that target concrete communication behaviours aligned with collaborative learning goals. Future work should apply the protocol in multiple PBL contexts, report cross-context boundary conditions, and evaluate whether feedback interventions based on the identified communicative patterns lead to measurable improvements in learners' communication and collaboration outcomes.

References

[1] Bubaš, G., Radošević, D., & Hutinski, Ž. (2003). Assessment of computer mediated communication competence: Theory and application in an online environment. *Journal of Information and Organizational Sciences, 27*(2), 53-71.

[2] Akhter, M. (2024). Interventional DN2CN: Causal Structure Learning from Mixed Observational and Experimental Data with Context-Specific Independence. *Bulletin of Computer and Data Sciences, 5*(3), 32-44.

[3] Lai, C. L. (2021). Exploring university students' preferences for AI-assisted learning environment. *Educational Technology & Society, 24*(4), 1-15.

[4] Huang, C., Han, Z., Li, M., Wang, X., & Zhao, W. (2021). Sentiment evolution with interaction levels in

blended learning environments: Using learning analytics and epistemic network analysis. *Australasian Journal of Educational Technology, 37*(2), 81-95.

[5] Kaliisa, R., Rienties, B., Mørch, A. I., & Kluge, A. (2022). Social learning analytics in computer-supported collaborative learning environments: A systematic review of empirical studies. *Computers and Education Open, 3*, 100073.

[6] Cheng, Z., Zhang, Z., Xu, Q., Maeda, Y., & Gu, P. (2025). A meta-analysis addressing the relationship between self-regulated learning strategies and academic performance in online higher education. *Journal of Computing in Higher Education, 37*(1), 195-224.

[7] Ming, N., & Baumer, E. (2011). Using Text Mining to Characterize Online Discussion Facilitation. *Journal of Asynchronous Learning Networks, 15*(2), 71-109.

[8] Bunz, U., & Montez, D. (2015). Computer-mediated communication competence. *Communication Competence,* 153-192.

[9] Saqr, M. (2018). *Using Learning Analytics to Understand and Support Collaborative Learning* (Doctoral dissertation, Department of Computer and Systems Sciences, Stockholm University).

[10] Saqr, M., & López-Pernas, S. (2023). The temporal dynamics of online problem-based learning: Why and when sequence matters. *International Journal of Computer-Supported Collaborative Learning, 18*(1), 11-37.

[11] Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254).

[12] Yao, M. Z., & Ling, R. (2020). "What is computer-mediated communication?" An introduction to the special issue. *Journal of Computer-Mediated Communication, 25*(1), 4-8.

[13] Wise, A. F., & Hsiao, Y. T. (2019). Self-regulation in online discussions: Aligning data streams to investigate relationships between speaking, listening, and task conditions. *Computers in Human Behavior, 96*, 273-284.

[14] Yew, E. H., & Schmidt, H. G. (2012). What students learn in problem-based learning: A process analysis. *Instructional Science, 40*(2), 371-395.

How to cite this article: Mobeen Akhter (2025). Text-Aware Prediction of Computer-Mediated Communication Competence in Collaborative Problem-Based Learning. *Bulletin of Computer and Data Sciences*, 6(3), 1-19. DOI: 10.71448/bcds2563-1

Received: 21/03/2025 **Revised:** 21/06/2025 **Accepted:** 19/07/2025 **Publish:** 30/09/2025

AND DATA SCIENCES BULLETIN OF COMPUTER *Bulletin of Computer and Data Sciences* is a peer-reviewed open access journal.