Chapter 7

Study 3: Predicting rapport levels based on keystroke patterns

Study 3 aims to classify users, based on their reported level of rapport during a conversation: one group reported a very high level of rapport during a conversation, while the other group reported medium to low levels of rapport. Since rapport has been shown to be detectable in speech prosody, and prior studies have shown that typing patterns can parallel aspects of speech prosody, the aim of this study is to discern if feelings of rapport are also detectable in typing patterns. Moreover, this study investigates how keystrokes from different subsets of a conversation predict rapport levels. By better understanding this, future researchers will be able to pinpoint areas of a conversation in which levels of rapport could be more accurately detected. Since rapport detection in language production is dependent on when the language is produced within a dialogue, this is especially important (Tickle-Degnen and Rosenthal, 1990).

Study 3 aims to detect as many cases of the medium-to-low rapport group as possible, since in a situation where we are trying to improve an interaction, this would be the group most in need of detection and intervention.

However, rapport is a very complex social dynamic (e.g. Tickle-Degnen and Rosenthal, 1990). This adds to the importance of finding additional sources of information, e.g. keystroke timing, that are sensitive to rapport levels. Many definitions of rapport exist (see Section 7.1 and Table 7.1 for examples); one of the most apt comes from early work on rapport, though: "...an individual's experience of harmonious interaction with another person, often described as 'clicking' or 'having chemistry"' (Tickle-Degnen and Rosenthal, 1990, p. 286) Despite the complexity of rapport, typing patterns provide a promising way to understand and measure this underlying social dynamic, since typing is sensitive to a number of cognitive and social processes (Schilperoord, 2002). Study 3 collected survey data from each subject after a conversation, where each question in the survey asked the subject to rate the conversation and their partner on different dimensions of rapport. I then took the mean of all ratings, and used this as a measure of rapport during the conversation. I then extracted features from typing patterns during the conversation, and used machine learning to predict rapport ratings based on these typing features.

Since establishing high rapport can be an important element of a successful interaction or productive collaborations , whether doctor/patient, salesperson/customer, manager/employee, etc., it could be very helpful to be able to continuously measure and predict rapport as an interaction unfolds. This study pursues that line of inquiry by looking at not only how well a full typing session can predict rapport, but also how well different subsets can predict final rapport. If this is successful, it could allow rapport level to be measured before an interaction concludes, so that one party can make adjustments to their interaction style in order to improve rapport.

As an example, if a customer is interacting with a salesperson online, the salesperson would want to make sure that the customer is enjoying their interaction. Furthermore, the salesperson would want to know as early as possible whether the customer is enjoying themselves, so that the salesperson can make adjustments to their approach.

However, study 3 investigates whether each portion of a conversation is equally valuable in predicting overall rapport, in order to establish if the influence of rapport is stronger in temporal slices of a conversation, as well as in slices that are segmented by the role that the typist is playing in the overall experimental task. Finally, an additional subset randomly samples a proportion of keystrokes, to see if the aforementioned slices *per se* are responsible for changes in predicting rapport, or if the type of subsetting is important. (See Figure 7.1 for a visual representation of these subsets.)

As a reminder, Study 3 aims to answer these research questions:

RQ 3a) Can typing patterns over an entire conversation be used to predict low levels of rapport between partners in an interaction?

RQ 3b) Can a random subset of keystroke data predict conversational rapport as well as a complete set of keystrokes?

RQ 3c) Does a subset of keystrokes from the first half of a conversation predict low rapport as well as a subset of keystrokes from the second half of a conversation?

RQ 3d) Does a subset of keystrokes from when a subject is providing recommendations predict low rapport as well as a subset of keystrokes from when a subject is receiving recommendations?

The motivation for these questions is to find the most valuable subsets of an interaction for predicting rapport. Tickle-Degnen and Rosenthal (1990) hypothesized that different components of rapport are more important or less important at different stages of an interaction; by slicing a conversation temporally, we can show which of those components is most important for overall, final rapport. In addition, study 3 is motivated by the ultimate application of my research: recommender "systems." A system could be a chatbot providing automated recommendations to a customer or a doctor providing well-being recommendations to a patient. In both cases, though, the user of interest is *receiving* recommendations, and so it is important to see if a user in this role is expressing signs of rapport.

To answer these questions, Study 3 used a uniform set of features for each model, but different subsets of keystroke data went into calculating each feature. A machine learning model was then trained on each subset, and the models' predictive accuracy was compared. As discussed in Section 7.2, though, the most important class to predict was when a subject reporting lower levels of rapport, rather than accurately predicting when a subject reports high rapport. Therefore, the comparison metrics reflect how well the model detected this class, rather than, e.g., overall accuracy.[1] The

[1] It should be noted that for all of the models, the ZeroR, simply classifying every instance as the majority class, would be 136/192, or 71%. But by definition, this would misclassify every instance of the minority class; therefore

features utilized in this study include both keystroke timing signatures as well as stylometric features. However, in each experiment, feature importance is also assessed because the same feature derived from a different subset of data could make that feature more or less valuable for predicting overall rapport.

Overall, I found that keystrokes are useful for detecting low rapport, especially when keystrokes from an entire conversation, rather than a random subset, are utilized. In addition, I found that keystroke patterns from when a user is *receiving* recommendations are more useful at predicting low rapport than keystrokes patterns from when a user is *providing* recommendations.

7.1 Related Work: Rapport

Rapport is essential to establishing a good relationship or good interaction, but is notoriously difficult to define and measure.

7.1.1 Defining rapport

Tickle-Degnen and Rosenthal (1990), a seminal study on the conceptualization of rapport, defines it as "...an individual's experience of harmonious interaction with another person, often described as 'clicking' or 'having chemistry."' While this definition is comprehensive, and has been adopted by many subsequent researchers, the definition still resorts to phrases such as "harmonious interaction" and "having chemistry," which are themselves difficult to define.

Lubold and Pon-Barry (2014) defines rapport in terms of a feeling of closeness, and many rapport-related questionnaires ask about a feeling of connectedness. Along those lines of connectedness, other studies on rapport ask subjects to what degree the partner "paid attention" to the subject, or the conversation was "engrossing" and "worthwhile" (LaBahn, 1996). Regardless of definitions, it seems that researchers must repeatedly resort to abstract concepts in order to explain rapport.

it is not a germane comparison. All of the classifiers have an overall classification accuracy near 71%, but this is not reported because Study 3 is concerned with correctly detecting the minority class.

These definitions point to the possible utility of connecting rapport to keystroke analysis. Table 7.1, below, highlights how keystroke timing could add a concrete measurement to some of the more abstract definitions of rapport.

The motivation behind why keystroke timing analysis could be a more accurate method to measure rapport comes from Chung and Pennebaker (2014, p. 5):

> A consistent finding is that many of the word categories used to reliably classify psychological states can be considered to be a part of language style as opposed to language content. That is, **how** people say things is often more revealing that **what** they are saying. [Emphasis added]

In other words, this study (as well as previous studies in this thesis) look at not only lexical choices (*what* is said) but also timing characteristics of keystroke production (*how* it's said).

7.1.2 Measuring rapport

Early attempts to quantify rapport mostly focused on the relationship between psychotherapists and their patients, since rapport is critical to building a productive therapeutic relationship. Anderson and Anderson (1962) quantified rapport by looking at the proportion of matching definitions between a therapist and client, where higher rapport was correlated with a greater proportion of matching definitions, since both parties saw certain concepts in the same light.

Relying on word-matching, though, is a crude estimate; by using both linguistic choices and production patterns, I hope to be able to capture more robust similarities. Müller et al. (2018) ran a study with many similarities to my own study. They aimed to predict low rapport based on non-verbal cues. These included facial expressions and speech prosody. They found these non-verbal cues to be extremely helpful in establishing high rapport and in their own experimental predictions. For example, facial expression and body posture can signal "attention," which is an important element of rapport. Further, Müller et al. (2018) hypothesizes that these non-verbal cues are more often imitated by the other participant, and this reflection is important for establishing rapport.

Definition (source)	Findings from keystrokes and prosody
"…an individual's experience of harmonious interaction or 'clicking' …" (Tickle-Degnen and Rosenthal, 1990)	Pauses during typing, as well as increased mistakes, are associated with increased cognitive load or strain (Schilperoord, 2002). In a "harmonious" interaction, it seems that one would expect fewer prolonged pauses and fewer mistakes, because the subject is more comfortable with expressing their thoughts.
"…the perception that a relationship has the right 'chemistry' and is enjoyable." (LaBahn, 1996)	Multiple studies have found that typing patterns are sensitive to a typist's emotions (Epp et al., 2011; Lee et al., 2015a). Notions such as enjoyment, which are a component of rapport, seem to fit this category, and are therefore likely to be perceptible in typing patterns.
"engrossing…involving… worthwhile…" (Grahe and Bernieri, 2002)	Studies that looked at the connection between keystrokes and emotions also study the intensity of emotions, not just the positivity/negativity of the emotions themselves (Maalej and Kallel, 2020). It seems that a conversation that is engrossing or involving will evoke more intense and less apathetic contributions. This should be reflected in the energy with which keystrokes are produced, as realized in keystroke dwell time.

Table 7.1
Definitions of rapport and possible quantification by keystroke patterns

These findings are especially germane since my thesis approaches typing patterns as a manifestation of "silent prosody." However, whereas prosody in Müller et al. (2018) is visible/audible to other participants, my study aims to show that prosody manifested in typing patterns (that are likely not visible to a partner) is also valuable.

Herzig et al. (2016) performed a study that bears a number of parallels to my own study, where they aimed to predict customer satisfaction of an interaction based on previously compiled personality profiles, as well as *affective*, as opposed to purely lexical features, of prior conversations. Their goal was to be able to make predictions, and perhaps route the customer to the appropriate representative, before the current interaction even began. In my own study, this is similar to Experiment 3b, where I aim to predict rapport based only on the first half of a conversation. I also use features beyond purely lexically-based features, and look at production patterns, including semantically-contextualized production patterns.

In terms of the actual classifications of rapport, two primary methods exist: self-reports from participants, and external observers who assign a rapport rating to an interaction they observe. This study relies on participants' answers to a questionnaire after the experiment, which is a form of self-reporting. Future studies could also use external annotators, as both self-reports and external annotation have been shown to yield similar results, where a strong correlation exists between self-reporting ratings and an external annotator's ratings, even when the external judge only views a small slice of an interaction (Carney et al., 2007).

7.1.3 The complexity of rapport

Seo et al. (2018) highlights the complex interactions of individual verbal behaviors that contribute to a sense of high rapport. For example, asking an off-topic question, such as personal information, can increase rapport in certain settings, whereas in other settings or at the wrong time it can seem rude. As such, measuring straightforward semantic similarity between turns would not be a sufficient surrogate for rapport estimation. Similarly, some statements or questions are properly responded to with short responses, while others have more appropriate long responses. Therefore, in this case, measuring simple turn length, or turn length similarity, would also be insufficient.

The model devised by Tickle-Degnen and Rosenthal (1990) provides an attractive apparatus for experimentation, as the study breaks down rapport into three dimensions: *attentiveness*, *positivity*, and *coordination*. But the study also found that each dimension does not exude equal influence on rapport throughout the course of a conversation. For example, early in an interaction positivity and attentiveness are most important for establishing good rapport; in the later stages of an interaction coordination and attentiveness are more influential on good rapport. For this reason, my study breaks down conversations into a first half and a second half, in order to assess both the predictive capability of each half as well as which features are most important. Combining my findings, i.e. which temporal slice is most important, with Tickle-Degnen and Rosenthal (1990)'s most important components during that slice, will provide insight into which components of rapport are the most accurate predictors of overall, final rapport.

Finally, Raj Prabhu et al. (2020) provides a robust confirmation of my own research methods, specifically combining a subject's answers to multiple questions into a single metric, which they call *Conversation Quality*. Because the perception of quality is so multi-faceted, they affirm that it can only be derived from the answers to multiple questions. Similar to the ultimate goals of my thesis research, they "intend to quantify spontaneous interactions by accounting for several socio-dimensional aspects of individual experiences" (p. 196). By quantifying an experience, researchers can compare different experiences and also quantify how much a certain change improved an experience.

7.1.4 Rapport and spoken language production

As mentioned in the introduction, Study 3 utilizes both keystroke timing features and stylographic features. These stylographic features, however, go beyond metrics such as the average length of an utterance, and instead are focused on elements such as the ratio of utterance length between one subject and their partner. This is related to the notion of "coordination," which Lubold and Pon-

Barry (2014) found to correlate strongly to rapport. This ratio, rather than outright measurement, highlights the differences with studying underlying motivations in interaction rather than in isolation. In other words, understanding a typist in dialogue requires more than understanding the typist's own metrics, but also understanding the relationship of those metrics to their partner's metrics.

Michalsky and Schoormann (2017) shows that when a subject finds their partner to be more likable and socially attractive, the subject puts more cognitive effort into matching their partner's style, where style-matching is another form of coordination. In this study, greater cognitive effort is manifested by typing rate and edit patterns. As such, it will be interesting to see if these variables are predictive of rapport levels.

Finally, it seems that rapport is ideal to study through keystroke analysis, given the sensitivity of this type of analysis to changes in cognitive load (Brizan et al., 2015, *inter alia*). Barnett et al. (2018) and Barnett et al. (2020) found that when an examiner intentionally established either high or low rapport with a subject, even though the experimenters did have meaningful interactions with the subjects, the level of rapport affected performance on cognitive tasks such as the Stroop test and word association tests. In these investigations, it was found that high rapport improves results on cognitive assessments.

Findings such as those by Barnett and colleagues seem to underscore the importance of studies in my thesis. If keystroke analysis can provide accurate predictions of perceived rapport, and rapport helps improve cognitive functioning, then increasing rapport will create a more productive interaction, e.g. more accurate movie recommendations because the partners in an interaction are able to better articulate their thoughts.

7.2 Methodology

7.2.1 Dataset partitioning

In order to test the predictive accuracy of different subsets of keystroke data, the entire dataset was partitioned in different ways, as illustrated in Figure 7.1. The random subset used 50% of the keystrokes of each subject in each conversation, so as to include approximately the same number of keystrokes as the other partitioning methods. In addition, if one subject had 100 messages and the other subject had 50 messages, then taking an equal proportion of each would keep the comparative ratio between the two subjects the same.

As can be seen in Figure 7.1, the role partitions included keystroke data from both halves of the conversation; vice versa, the half partitions included data from each role.

7.2.2 Determining number of classes

This study used each subject's post-conversation questionnaire responses to construct a 6-dimensional vector of conversational impressions for each subject. These 6 questions are reproduced in full in Section 4.4.5. They asked participants about enjoyability, conversational smoothness, partner connections, etc. In other words, each question attempted to get at a different dimension of overall rapport (LaBahn, 1996; Lubold and Pon-Barry, 2014; Tickle-Degnen and Rosenthal, 1990).

Before proceeding with any prediction preprocessing, the subjects' survey answers had to be divided into the appropriate number of classes, so that classes were as discrete as possible without being too small. As can be seen in Figure 7.2, survey responses were heavily positively skewed (a Likert scale from 1-7 was used for each question, with 1 being low enjoyability, low connection, etc. and 7 being high ratings of those qualities). Looking at the average response value to all questions, 40 subjects had a 7.0 average, i.e. every question received a 7 rating. Only 15 subjects had a mean answer of less than 3.5.

In order to determine the optimal number of clusters, I used `NbClust`, an R package that tests 30 different distance algorithms to measure the appropriate number of clusters (Charrad et al., 2014). Each algorithm uses different instantiations of cluster analysis such as within cluster sums of squares, average silhouette and gap statistics. The majority of distance metrics recommended 2 clusters, while a handful recommended 3 clusters. This study will use 2 clusters, although studies can also investigate classification for 3 clusters. Figure 7.3 uses PCA to visualize 2 versus 3 clusters. As can be seen in Figure 7.3b, there is not clear partitioning between the highest 2 clusters, which

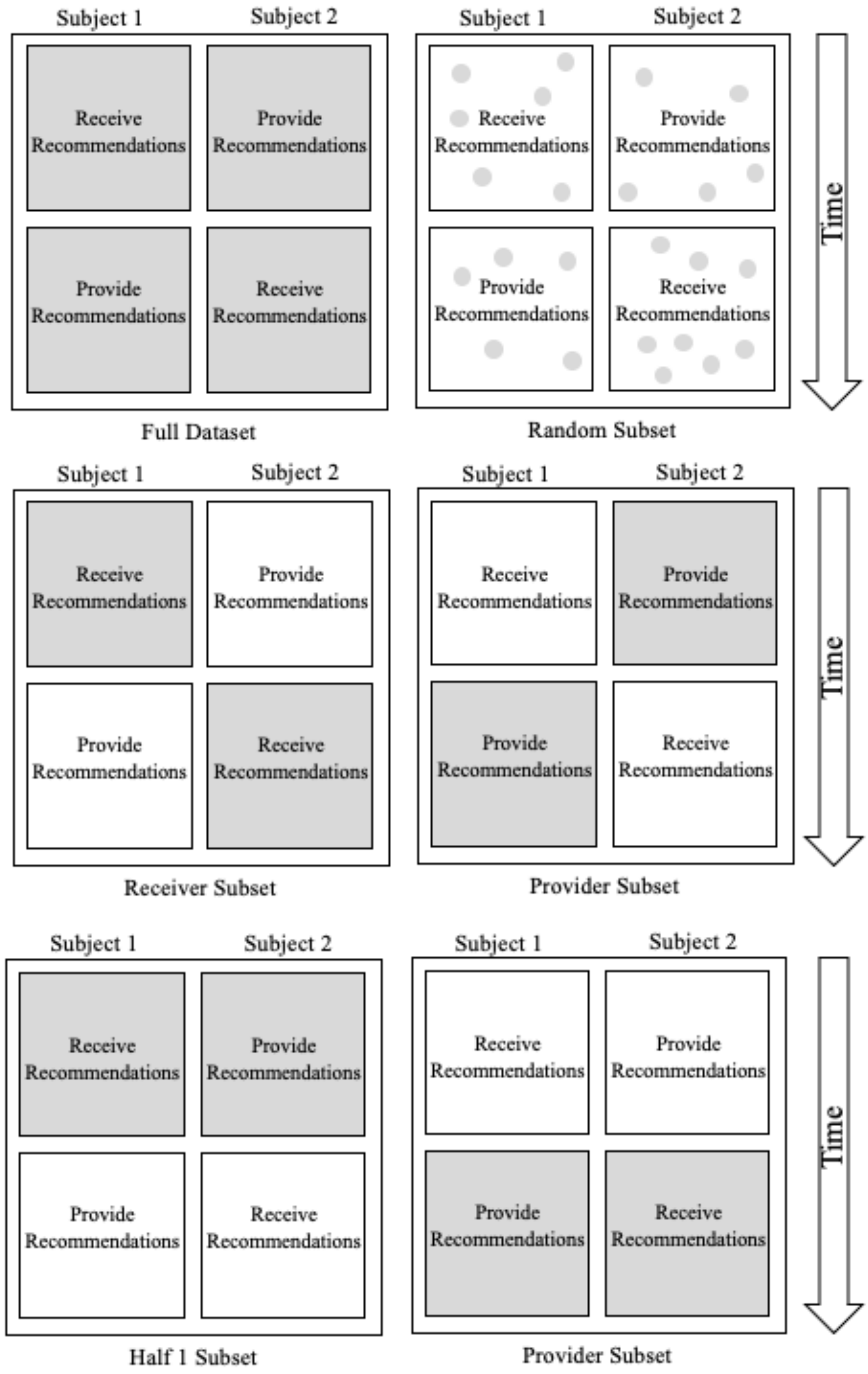


Figure 7.1
Each subset was created by dividing the full dataset in different ways, either by receiver/provider roles, first half/second half, or random sampling. The grayed out areas indicate the selected data.

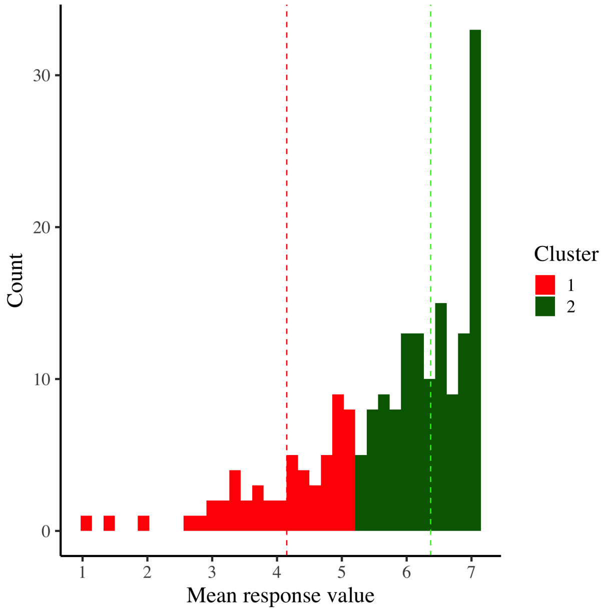


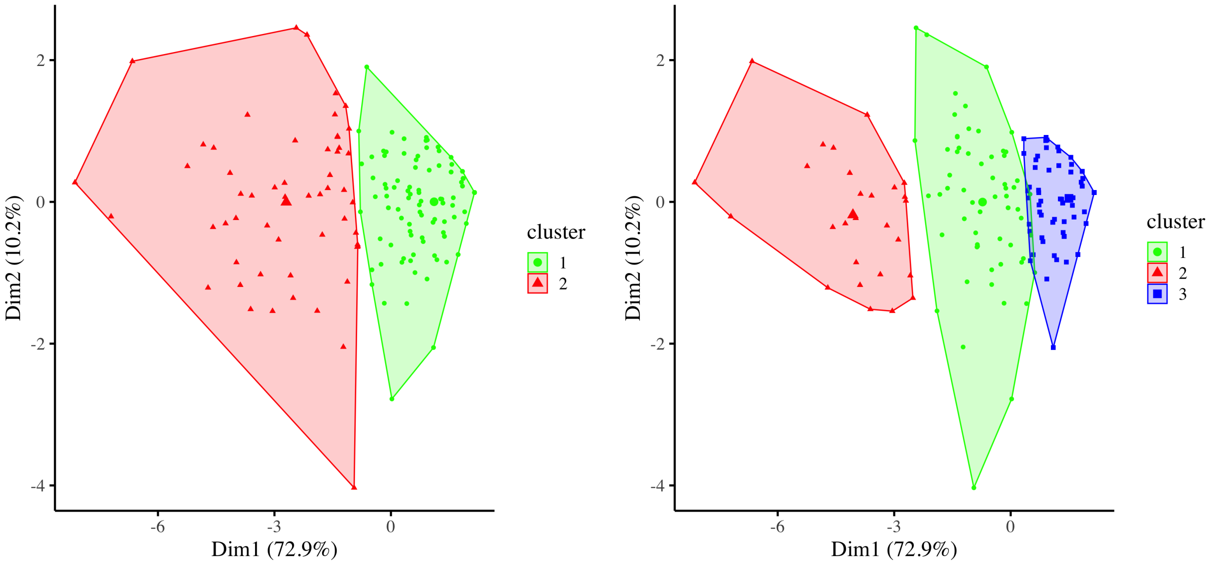
Figure 7.2
A histogram of the mean response value from each subject. The bottom lines represent the mean response value for each cluster. As can be seen, the overall response value was heavily positively skewed.

would hint that these clusters are not necessarily well-separated. Once it was determined that 2 clusters was optimal, k-means clustering was used to classify each subject into a cluster.

Nonetheless, class size was still skewed. Cluster 1, made up of subjects who had lower rapport ratings, included only 56 subjects, with a mean question response of 4.15 (out of 7.0). Cluster 2 included 136 subjects with a mean response rating of 6.38. Because of this imbalance, synthetic minority over-sampling (SMOTE) was used to balance training data (Chawla et al., 2002).

7.2.3 Metrics selection

As mentioned previously, the minority class, consisting of subjects who reported lower rapport levels, was the class of interest. For this reason, overall accuracy was not used, since the majority class consists of those subjects who feel higher rapport; in a real-life setting these partners would not necessarily require extra attention. In order to use more meaningful metrics, this study focuses on the following 4 metrics:



(a) Creating 2 clusters, or classes, from the 6-dimensional answer vector for each subject. PCA was used to visualize the clusters.

(b) Creating 3 clusters, or classes, from the 6-dimensional answer vector for each subject. PCA was used to visualize the clusters.

Figure 7.3

The difference in cluster partitioning between two clusters and three clusters. In partitioning the data into3 clusters, more overlap exists between the top two clusters.

- **Area under the ROC curve (AUC)** - Studies such as Wardhani et al. (2019), among others, suggest that AUC is more robust than F1-score when evaluating imbalanced data. Further, AUC uses prediction probabilities, whereas F1 looks at correct/incorrect predictions. Because rapport is not a discrete binary, probabilities also seem more appropriate for the problem at hand.

- **Matthews correlation coefficient (MCC)** - The MCC is a convenient metric in that it is a single numeric representation of all cells in a confusion matrix (Chicco and Jurman, 2020). However, its performance can sometimes degrade on imbalanced data (Zhu, 2020). As a result, it is still reported in the results as all elements of the confusion matrix are important for this task; however, AUC will still take precedence.

- **Positive predictive value (PPV)** - The PPV measures proportion of positive cases (those who *actually* report low rapport) against the number of predicted class members. Unlike a straightforward true-positive or true-negative rate, the PPV also takes into account the *prevalence* of a class, which is the proportion of the class of interest within the entire dataset (Altman and Bland, 1994). For this reason, it is germane to this study, where low rapport prevalence could vary in different situations; therefore, it is reported but also secondary to AUC.

- **F1 score** - Although F1 scores may not be the most appropriate for imbalanced data, it does balances precision and recall, and takes the harmonic mean of both, thus giving them equal weight. While empirical evidence exists to show that F1 core might not be ideal for my scenario, the scores are still reported here for the sake of completeness, and because they are widely used and understood in the machine learning community.

7.2.4 Feature selection

The full list of features, with brief explanation, is enumerated out in Table 7.2. Most features are repeated from previous studies. However, since Study 3 uses data from both sides of an entire conversation, it also includes dyadic features such as ratios between subject language production and conversational partner language production.

7.2.5 Classifier choice

All of the models in this study were built using the `tidymodels` ecosystem in R (Kuhn and Silge, 2022). This allowed for all model types to be built from the same syntax and code blocks, which aids reproducibility within this study as well as future reproducibility.

A tuning subset of data, a holdout set consisting of 5% of the total data, was run using H2O's automatic machine learning (AutoML) (LeDell and Poirier, 2020). The algorithm tested 83 different

Feature	Comments
Inter-keystroke interval (IKI)	Similar to Epp et al. (2011), times were log-transformed and pruned to the top 95%
IKI of content words	Content words (Pennebaker, 2011) have intrinsic meaning and include nouns, verbs, etc.
IKI of function words	Function words include determiners, pronouns, etc.
IKI at word beginning	This measures the hesitation before selecting a word, rather than mid-word when the lexical item has already been retrieved (Logan and Crump, 2011).
IKI in mid-word	This measures motor execution, rather than lexical retrieval (Logan and Crump, 2011).
IKI at phrasal boundaries	As described in Galbraith and Baaijen (2019), these pauses reflect different cognitive processes that focus on phrase planning rather than word retrieval.
IKI before sending message	This reflects hesitation before transmitted a message.
Dwell time	Dwell time is strongly connected to emotion (Lee et al., 2014).
Dwell time of content and function words	If dwell time is more connected to emotions, and content words have more intrinsic meaning, then the semantic words types should be measured separately.
Edit count	The frequency of edits should reflect uncertainty in the language being produced (Olive et al., 2009).
IKI of lexical density	Lexical density measures the ratio of different types of words, such as nouns to total words (e.g. Khawaja et al., 2014). Study 3 measures the IKI timing ratio of content or function words to all words.
Turn count ratio	The ratio of subject turn count to partner's turn count, in order to test the importance of coordination between partners (Gravano and Hirschberg, 2009).
Turn-type ratio	Similar to the above, this tests if partners overlap the other with the same frequency, as opposed to how often they let the other partner complete a turn.
Word count ratio	This looks at the ratio of how many words each partner is producing. If rapport is high, word count should be approximately equal (Erkens et al., 2005).

Table 7.2

A list of features used in Study 3

models. The top scoring models, as measured by the range of AUCs, were random forests, boosted trees, and neural networks.

In future studies, more sophisticated modeling techniques should also be considered, including deeper neural networks, LSTMs with recurrence, and Transformer models. These more sophisticated neural networks are especially useful for modeling keystroke patterns on mobile devices (Chang et al., 2021b; Stragapede et al., 2022). However, Study 3 only used a relatively simple multilayer perceptron because it allowed for straightforward comparisons between models and was thus useful for an initial investigation.

For each subset, the tuning set was then used to calculate optimal hyperparameters for each model. In other words, models were retrained for each subset of data. The rationale for this is that the objective of Study 3 was not to measure how well subsets perform on a model tuned from a full dataset, but rather how well each subset performed as a complete dataset, i.e. training and testing.

Because my dataset was small and imbalanced, using traditional partitioning of the data would result in very small sample sizes. For example, if I had used a 75%/25% training/testing split, then the minority class in the testing set would only have 14 instances, which seems insufficient for obtaining reliable prediction results.

In order to circumvent this size issue, I used cross-validation on the entire dataset (minus the tuning set). Specifically, I use Monte Carlo cross-validation, also known as *repeated random sub-sampling validation* (Burman, 1989; Shao, 1993). Unlike *k*-fold cross validation, in this setup the folds are not mutually exclusive, and are randomly resampled with replacement. As a result, the test set would not need to be broken up into *k* groups; rather, overlap is allowed between folds. The disadvantage to this approach, though, is that it leaves open the possibility that a datapoint would never be selected. That being said, I ran 25 repetitions for each model, which should minimize the risk of datapoints being left out.

Model	Dataset	Correct Predictions	Mean Certainty
Neural Net	Receiver	36	0.59
Neural Net	Half 2	32	0.53
XG Boost	Random	31	0.52
Neural Net	Half 1	30	0.52
Neural Net	Full	28	0.52
Neural Net	Random	30	0.51
XG Boost	Half 1	27	0.49
Random Forest	Half 1	27	0.49
Neural Net	Provider	24	0.48
Random Forest	Receiver	26	0.48
Random Forest	Random	26	0.47
XG Boost	Receiver	27	0.47
Random Forest	Half 2	19	0.44
Random Forest	Full	18	0.44
XG Boost	Half 2	22	0.44
Random Forest	Provider	20	0.44
XG Boost	Full	20	0.43
XG Boost	Provider	22	0.42

Table 7.3

This table is made up of each model/dataset combination. Results are arranged by the mean certainty for each combination in predicting the minority class. As can be seen, the neural network trained on the receiver subset had both the highest number of correct minority class predictions, as well as the highest average confidence in predicting the minority class.

7.3 Results

As mentioned earlier, all models were tested on all relevant subsets. However, for those experiments reported here, only the neural network model results are reported. This decision was made because the neural network reported the strongest AUC results on the full dataset. This partially represents the model's high confidence in its predictions. This can be visualized in Figure 7.4 below, where the neural network's predictions have a higher density in the high probability end of the x-axis. A density plot was appropriate because the number of instances (of the minority class) was constant, which makes the densities appropriate for comparison. In addition, Table 7.3 is arranged by the mean confidence in predictions of the minority class. Most of the top spots are occupied by the neural network.

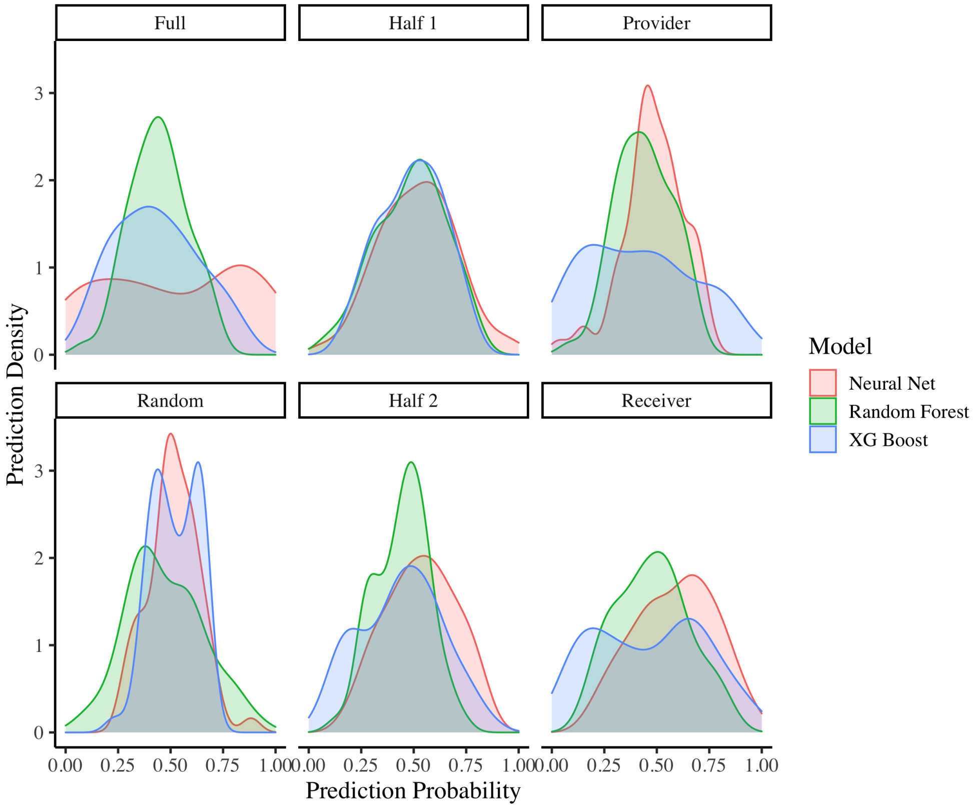


Figure 7.4
A density plot of mean predictive confidence in predicting the minority class. In many instances, the neural network has the greatest density of high-confidence predictions. This is not exclusive to the neural network, though, and some classifiers are more confident on certain subsets.

Dataset	Mean AUC (SD)	Mean MCC (SD)	Mean PPV (SD)	Mean F1 (SD)
Full dataset	0.71 (0.08)	0.27 (0.14)	0.47 (0.11)	0.48 (0.10)
Random subset	0.60 (0.11)	0.07 (0.19)	0.34 (0.11)	0.39 (0.13)
p-value	$< .0001$	$< .0001$	$< .00001$	$< .001$
Effect size (d)	1.18	1.15	1.23	0.81
df	44	44	48	46

Table 7.4
For each dataset, the AUC, MCC, PPV, and F1 score of the neural network are reported. In addition, for the comparison of the two datasets, the p-value, effect size, and degrees of freedom are reported. In this comparison, the differences in each metric score are significant.

For each experiment, the AUC, MCC, PPV and F1 score of the neural network model trained on that subset are reported, and the distributions are visualized. In addition, for each metric a two-sample *t*-test was run on each pair, and those results also reported.

Finally, the most important features for each classifier are visualized, although these are from the boosted tree models. The reasoning behind this was two-fold: (1) single feature importance is difficult to assess in neural networks, and (2) the boosted tree still had very good performance on many datasets.

7.3.1 Experiment 3a: Full dataset vs random subset

The first experiment tested whether rapport ratings could be equally well predicted from the entire dataset as compared to a random subset. This test was performed first so that any significant differences detected in Experiments 3b and 3c could be more appropriately attributed to that specific subset, rather than subsetting in general. The random subset used 50% of each subject's total keystroke output.

As can be seen in Figure 7.5 and Table 7.4, in all cases each metric was significantly better for the full dataset. All effect sizes are considered large (Cohen, 1988).

As can be seen in Table 7.4, the variance is lower for all full datasets across metrics. A possible explanation for these results, then, is that adding more keystrokes reduced uncertainty and led to more confidence in model predictions.

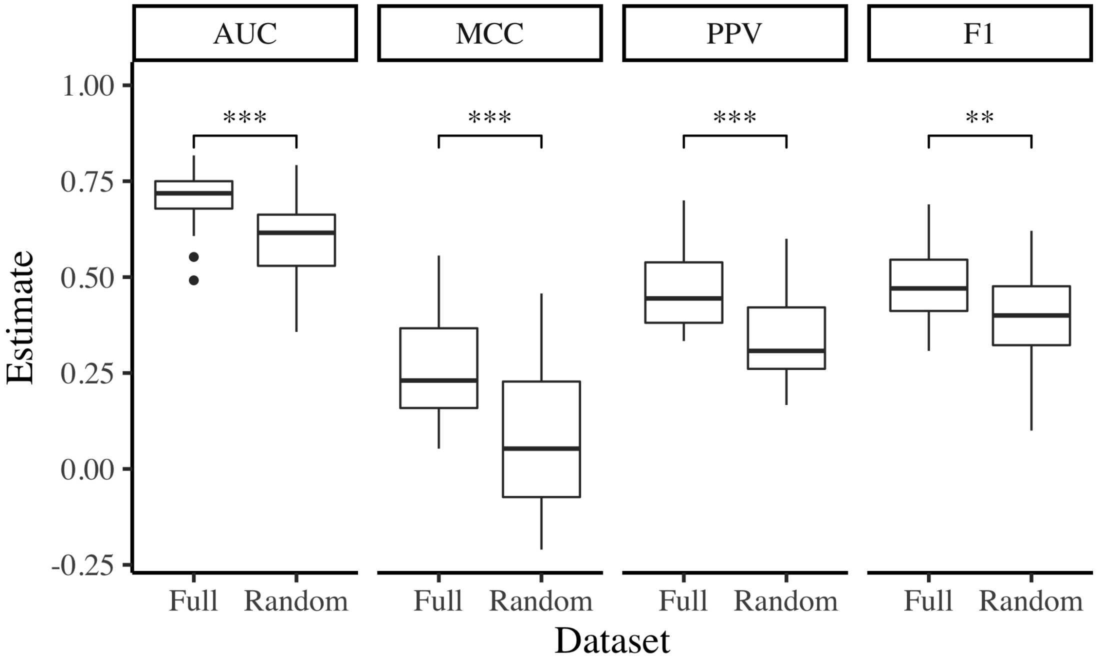
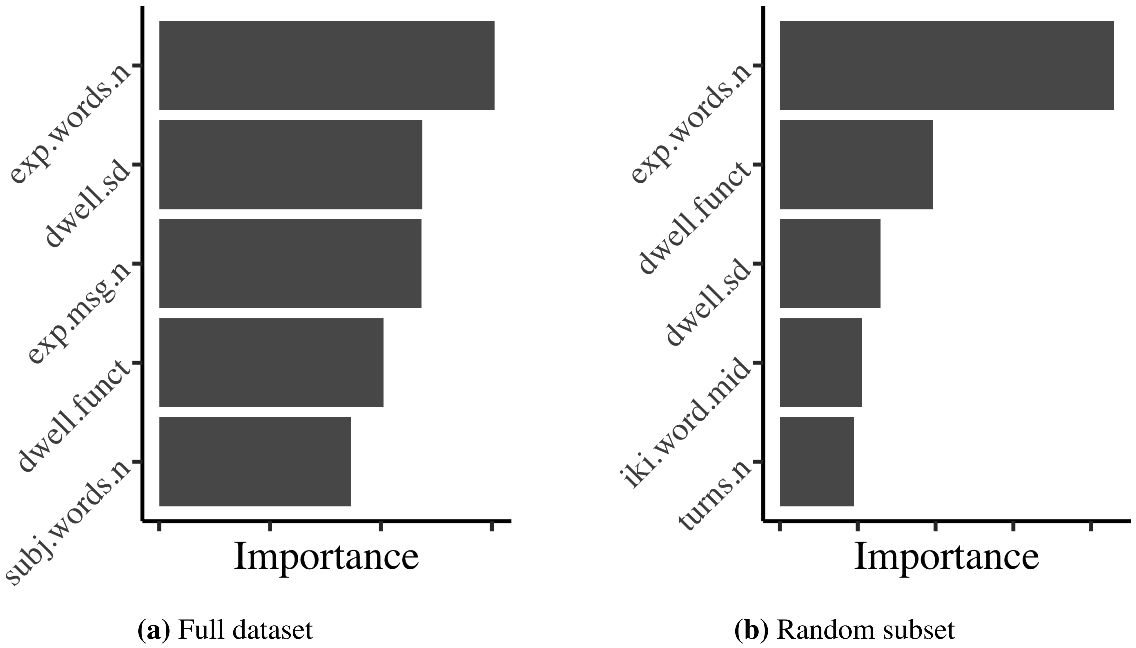


Figure 7.5

The distributions of metric scores for the full dataset versus a random subset. All metrics were significantly higher for the full dataset as compared to the randomly selected subset. It is important to bear in mind that different metrics are calculated on different scales, and so comparing, e.g. AUC to MCC is not meaningful. The only meaningful comparisons are within each metric.

Because this study is inferential, I also wanted to extract *which* features were the most important. The `vip` package in R calculates importance using the Shapley values of each feature, where Shhapley values are similar to coefficient values in linear regression (Shapley, 1953). Crucially, though, because of the opaque nature of neural networks, these charts were made using the optimal boosted tree models. Although the results of the boosted trees were not consistently as accurate as the neural networks, they still scored highly, and so the important features of the boosted trees should not be radically different than the important features in the neural network models.

The five most important features in the boosted trees for each dataset are visualized in Figure 7.6. In both cases, the absolute number of words in the experiment were the most important predictor. However, key dwell times and typing rate (IKI) were also important.



(a) Full dataset

(b) Random subset

Figure 7.6
These figures illustrate variable importance for the full dataset and random subset. Feature importance is based on the boosted tree models rather than the neural network models, due to the opaque nature of neural networks.

Dataset	Mean AUC (SD)	Mean MCC (SD)	Mean PPV (SD)	Mean F1 (SD)
Half 1 subset	0.66 (0.09)	0.21 (0.15)	0.42 (0.11)	0.45 (0.11)
Half 2 subset	0.61 (0.01)	0.12 (0.17)	0.35 (0.09)	0.41 (0.12)
p-value	.09	.055	.01	.23
Effect size (d)	0.48	0.56	0.74	0.34
df	47	47	46	48

Table 7.5
For each dataset, the AUC, MCC, PPV, and F1 score are reported for the neural network models. In addition, for the comparison of the two datasets, the p-value, effect size, and degrees of freedom are reported.

7.3.2 Experiment 3b: First half subset vs second half subset

The next experiment subsetted keystroke data by whether the keystroke was typed in the first 8 minutes of the experiment or the last 8 minutes. The tests in this experiment were aimed at answering whether data from the initial half or latter half of the conversation was a better predictor of whether rapport was classified as high or medium-to-low.

The results of subsetting by conversation half were not as clear as Exp 3a. As can be seen in Table 7.6, most of the differences in metric scores were marginally significant, and all less than 0.10. However, only one metric, PPV, was significant below the 0.05 alpha level. In addition, the effect size of the AUC and F1 comparisons is considered small while the effect size for the MCC and PPV is considered moderate.

Finally, the five most important features for each dataset are visualized in Figure 7.8. In the first half, both keystroke and stylometric features were important, where stylometric features are measures of writing style such as word count. However, in the second half, only stylometric features were important. Although it fell outside the scope of study 3, future work will delve into why typing patterns were less important in the second half, as it may point to the notion that certain features are reflective of rapport only when those features occur within certain temporal slices of an interaction.

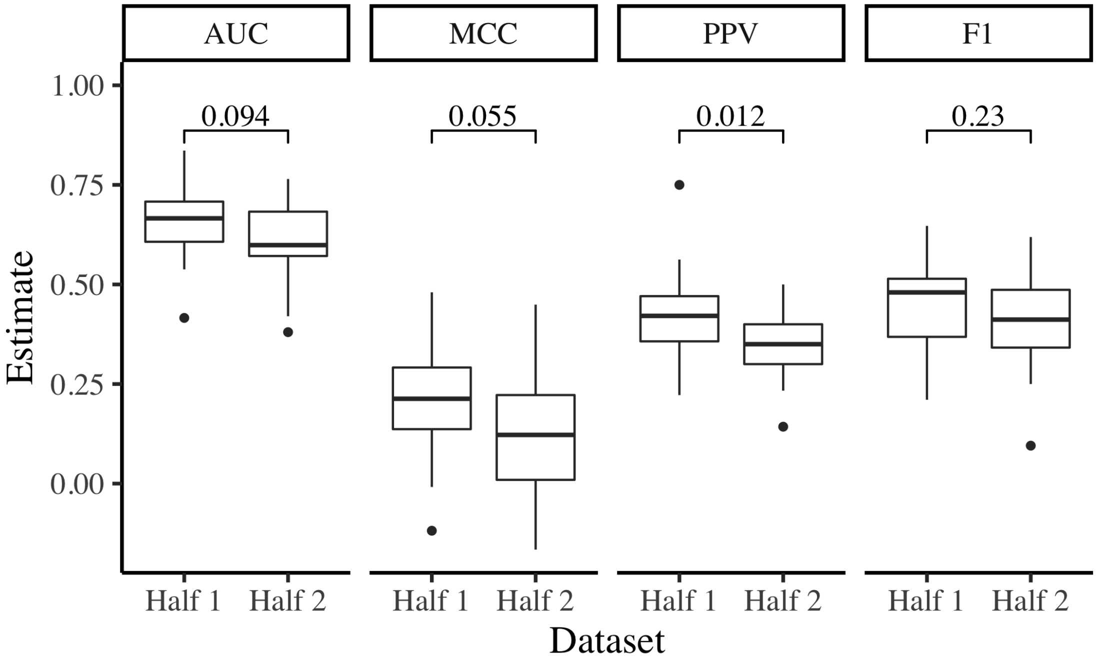
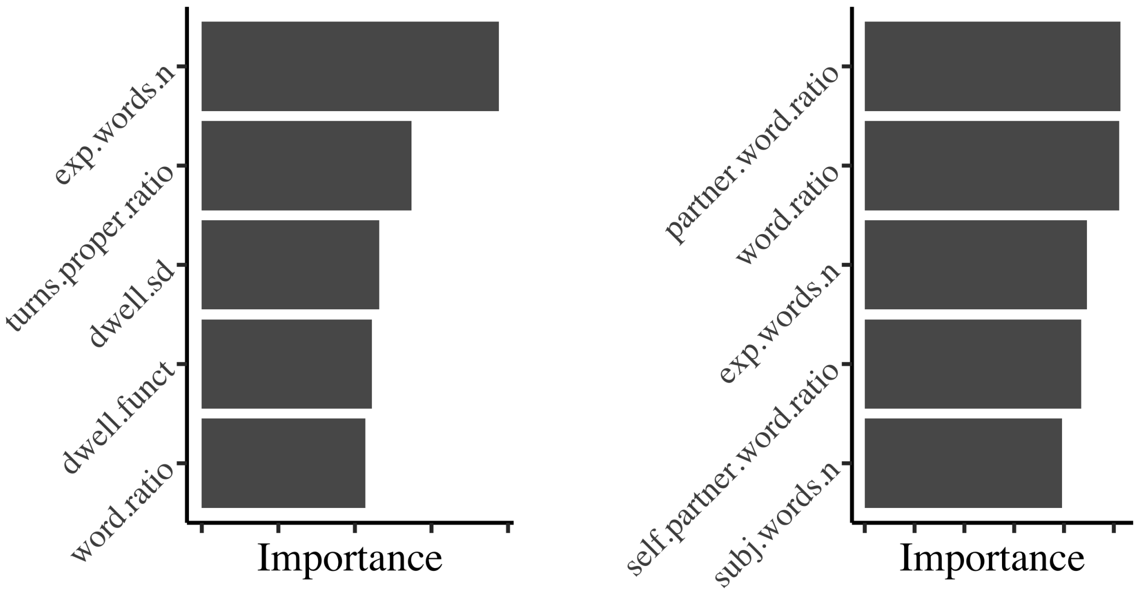


Figure 7.7

The distributions of metric scores for the first half subset versus the second half subset. The metrics are better for the first half subset as compared the second half, but only marginally so. It is important to bear in mind that different metrics are calculated on different scales, and so comparing, e.g. AUC to MCC is not meaningful. The only meaningful comparisons are within each metric.



(a) Half 1 subset

(b) Half 2 subset

Figure 7.8
These figures illustrate variable importance for the first half subset and second half subset. Feature importance is based on the boosted tree models rather than the neural network models, due to the opaque nature of neural networks.

7.3.3 Experiment 3c: Recommendation provider vs recommendation receiver

The final experiment investigated how well typing data from subjects in different "roles" predicted rapport ratings. By roles, I mean when a subject was *providing* movie recommendations compared to when a subject was *receiving* recommendations. This experiment tests whether the task at hand influences typing patterns, specifically how well typing patterns when occupying different roles predict feelings concerning rapport.

The results seem to paint a clear picture that typing patterns when a subject is receiving recommendations is a better predictor of rapport than typing patterns when a subject is providing recommendations. Moreover, the effect size of all of the differences reported above are considered large.

Dataset	Mean AUC (SD)	Mean MCC (SD)	Mean PPV (SD)	Mean F1 (SD)
Provider subset	0.59 (0.07)	0.10 (0.12)	0.35 (0.10)	0.39 (0.11)
Receiver subset	0.73 (0.10)	0.32 (0.16)	0.47 (0.10)	0.53 (0.11)
p-value	<.000001	<.00001	<.0001	<.0001
Effect size (d)	-1.65	-1.53	-1.21	-1.36
df	42	45	48	48

Table 7.6
For each dataset, the AUC, MCC, PPV, and F1 score are reported. In addition, for the comparison of the two datasets, the p-value, effect size, and degrees of freedom are reported.

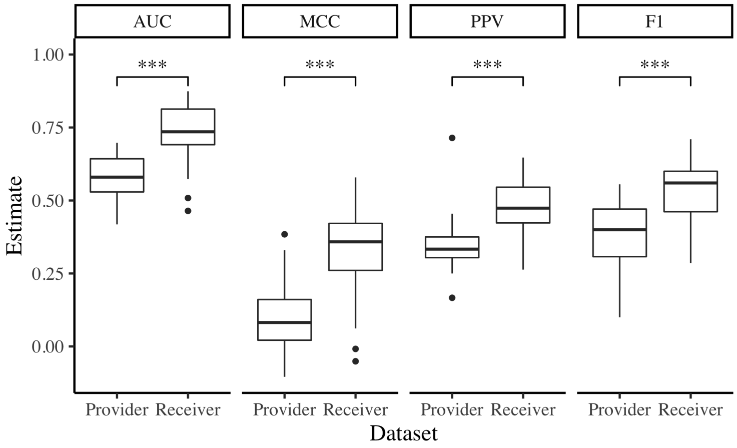
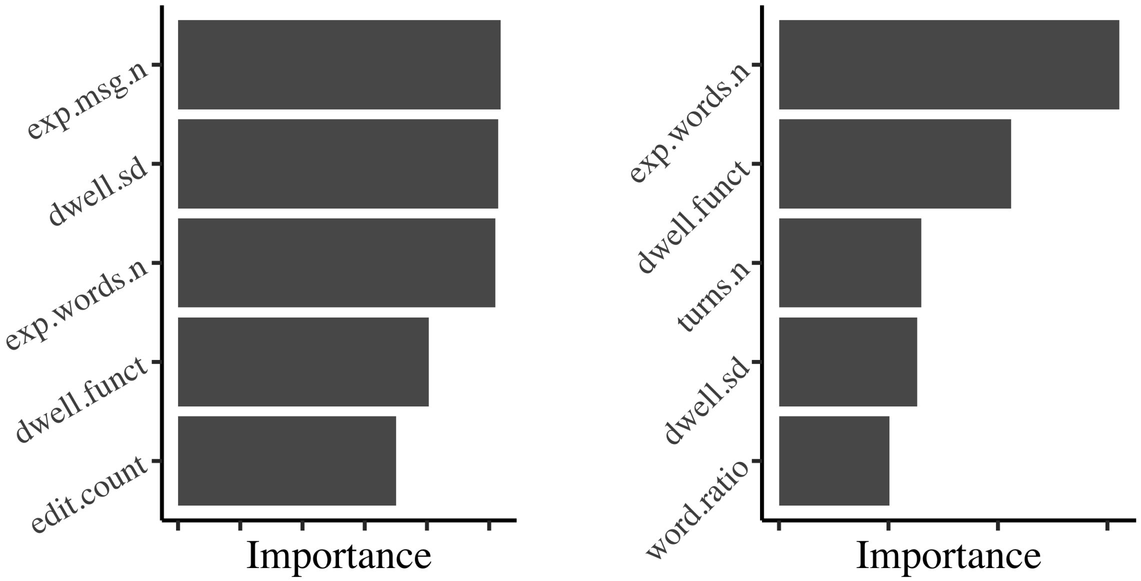


Figure 7.9
The distributions of metric scores for the Provider subset and Receiver subset for the neural network model. The metric scores from the receiver subset were significantly higher than the scores derived from data from the provider subset. It is important to bear in mind that different metrics are calculated on different scales, and so comparing, e.g. AUC to MCC is not meaningful. The only meaningful comparisons are within each metric.

Finally, the five most important features for each dataset are visualized in Figure 7.10. In both cases, the absolute number of words or messages in the experiment were the most important predictor. However, key dwell times were also important in both cases.



(a) Provider subset

(b) Receiver subset

Figure 7.10
These figures illustrate variable importance for the Provider subset and Receiver subset. Feature importance is based on the boosted tree models rather than the neural network models, due to the opaque nature of neural networks.

7.4 Discussion

The comparisons performed in Experiments 3a, 3b and 3c provide an intriguing picture of how accurately certain subsets of a conversation predict overall feelings of rapport during that conversation. Moreover, the different types of important predictors help to fill in this picture.

Looking at the overall findings, it is interesting how well the neural network performed on many of the datasets. One of the strengths of a neural network is its ability to perform feature engineering. However, this strength is sometimes limited only to deeper networks (Seide et al., 2011). Since this study used a multi-layer perceptron, it is possible that this deep feature engineering was a key to the model's success. This seems worth mentioning because it points to notion that perhaps the features constructed for this study (as well as the entire thesis) were too fine-grained or misguided, which is why the network's own constructed features were superior. Although this is a possible advantage of neural networks, one disadvantage is the "black box" nature of neural network predictions. To be more precise, it is very difficult to understand why the model made the predictions it did (Linzen et al., 2018). Nonetheless, methods do exist for understanding predictions, and these should be explored in the future to better understand exactly which dimensions of language production best predict underlying mindsets.

For the boosted tree model, which was fairly strong in its predictive accuracy, it is possible to extract variable importance, as seen in Figures 7.6, 7.8, and 7.10. Of note are two features types: subject/partner word ratios and key dwell times.

Ratios of words, messages, etc. between a subject and partner are interesting because they point to the importance of coordination between partners (Scissors et al., 2009), which is an *extrinsic* marker. On the other hand, a partner cannot see a subject's individual keystrokes, and so these features are considered intrinsic and not communicated.

If more coordination is a sign of higher rapport between partners, then it stands to reason that ratios between a subject's language production and their partner's production should be closer to equal, or 1:1. Indeed, a quick analysis of word ratios and message ratios both found that for cluster 2, the group with higher rapport ratings, these ratios were much closer to 1.0. For word-level ratios, this difference was found to be significant, while it was not statistically significant for message ratios.

A possible reason that word ratios were more important than message ratios was the nature of the experimental task: subjects were encouraged to engage in a "text message"-like conversation; similar to Ling and Baron (2007), this could result in strong reciprocity between a message and a response, regardless of the content of those messages. However, in high rapport conversation, the content or at least length of the messages would be more equal, which would result in the statistically significant differences in the ratio of word counts but not message counts.

The other interesting feature that plays an important predictive role in every dataset, and is the most important keystroke-based feature, is dwell time, or how long a key is held down for. As observed in studies such as Lee et al. (2014) and Lee et al. (2015b), dwell time is more strongly associated with emotional responses. It would then stand to reason that rapport would be more closely connected to dwell time than, e.g., typing speed. Moreover, the most important sub-feature of dwell times is often not the overall timing, but rather the timing associated with a semantic subset of words, such as function words or content words (Chung and Pennebaker, 2014; Pennebaker, 2011). Since the ultimate goal of my thesis research is to use keystroke timing to infer the way a human feels when talking to another human or computer agent, it is essential that a keystroke-based system also knows where to look for keystroke information.

The fact that these two types of features were both important also highlights an important theoretical dispute in conversational coordination: Is the phenomenon of coordination *mediated* by the environment or *unmediated* and the result of the user's own mental state? Another way to view this is through the lens of audience design, and whether coordination is intended for the audience, or is a natural response.

Like many previous studies, Branigan et al. (e.g. 2011); Garrod and Pickering (e.g. 2009), the fact that both intrinsic typing patterns and extrinsic word transmission are important points to a synthesis of these two theories. This has wide-ranging implications for the future design of HCI systems such as chatbots. On the one hand, a chatbot will need to use a measure of reflection in order to encourage a feeling of connection with a human user. On the other hand, a system should also monitor the cognitive effort, e.g. keystroke patterns, behind the coordinated word production of the user, to better understand the user's mindset and whether a feeling of connectedness exists.

Experiment 3a sought to answer **RQ 3a** and **RQ 3b**. **RQ 3a** was concerned with classifying rapport level over an entire conversation, and it appears that the distinction between high and lower rapport can be predicted. We can see this in Table 7.4, where the full dataset had a mean AUC of .71, which is above chance, and considered "acceptable discrimination" (Hosmer Jr et al., 2013, p.

177). On the other had, regarding **RQ 3b**, it appears that a random subset cannot predict rapport as well as a full dataset, with the AUC being considered "poor discrimination."

However, it is possible that this was due to the method of random sampling. For example, when subsetting a contiguous chunk such as the first half, in a 5-letter word, there would be 3 mid-word keystrokes, 1 word-initial keystroke, and 1 word-final keystroke. In the random sampling, hypothetically every keystroke could have been a mid-word keystroke. This speaks to the importance, in future work, of using a repeated random subsampling, so as to balance out these types of disparities, or only sampling full words and sentences.

The fact that the random subset was significantly worse was surprising because studies such as Pecune et al. (2018) or Olsen and Finkelstein (2017) used external annotators to make rapport judgments from very brief slices of a conversation, and found that these judgments were highly accurate. However, the fact that the judgments in other studies were accurate while the random subset in this study was significantly inferior, possibly points to the need to use a contiguous subset, rather than a random subset. As Zhao et al. (2014) points out, rapport is a dyadic phenomenon, co-constructed over time by both members of the dyad. It is possible, then, that tracking this continuous development is also important for rapport judgment.

That being said, the medium and methodology of my thesis compared to prior studies is vastly different. Moreover, there exists serious methodological issues with my random subset; specifically, the random subset was just *a* random subset. Because randomization was used, the random subset should have been resampled a number of times so that the full dataset could be compared to the randomization process in general. I will discuss this further in the Future Work section.

Experiment 3b provided answers regarding **RQ 3c**, the comparison of data from the first half of the conversation to the second half, which is also intriguing. Although not all of the differences in metrics were significant, they were all at least approaching significance ($p < .10$). In answer to **RQ 3c**, then, it appears that keystroke data from the first half of an experiments does predict final rapport as well as keystroke data from the latter half of a conversation.

However, the comparison of PPV, which specifically measured how well a classifier predicts low rapport (the “positive” case), was significant ($p = .01$). This finding aligns well with previous research on trust development in HCI. Tolmeijer et al. (2021) performed a longitudinal study of rapport and found that while first impressions are strongly influential on overall impressions, sometimes a negative first impression can be slightly improved with subsequent interaction. It seems that the results of Experiment 3b corroborate a very similar conclusion: data from the first half (first impressions) is significantly more accurate at predicting a poor final relationship. However, because not all metrics were significant, it could be interpreted as the second half data (subsequent dialogue) slightly improving a final impression.

In designing an intelligent computer agent, evidence from Experiment 3b would imply that the initial parts of a dialogue should be more heavily weighted and closely monitored, but that subsequent dialogue should not be disregarded.

Finally, Experiment 3c answers **RQ 3d** by shedding light on the importance of different tasks or roles in relationship building. The findings show that typing data from when a subject who is receiving recommendations is significantly better at predicting rapport than typing data from when a subject is providing recommendations. In fact,the AUC of 0.73 is higher than the AUC of the full dataset.

The fact that the receiver subset is more accurate than the provider subset seems to intuitively make sense, as a recipient would judge the quality of recommendations they are receiving, and use these judgments to form an impression of their partner (the provider). On the other hand, a provider is primarily only outputting recommendations, and so the provider has very little feedback information on which to judge their partner and form an impression.

It is possible that these findings follow from the theory put forth by Clark and Schaefer (1989) that the establishment of common ground requires both a *presentation* phase and *acceptance* phase. Gergle (2017) enumerates the varied ways that acceptance can be signaled, beyond verbal feedback. As such, perhaps typing patterns are also an internal signal of acceptance, and are therefore more strongly connected to a sense of rapport in the conversation.

For the larger aim of my thesis research, it is also beneficial that recipient information is more informative regarding rapport. In recommender systems or customer service chatbots, the user is the recipient of the information. For example, in a digital mental health app, a designer would want a computer agent to provide counseling to a user of the app. Since recipient typing data seems more informative about rapport, a system could better rely on the typing patterns being produced by the user (receiver) in order to assess how the recommendations are being received, and the level of rapport between the user and computer agent.

7.5 Future work

In future iterations of these experiments, a few changes should be made to both data collection methods and data processing methods.

The first change that should be made in future work is a new experimental setup that generates a greater proportion of low rapport levels. As mentioned repeatedly throughout this study, a limitation on classifier performance was the imbalanced dataset (although methods such as SMOTE and metrics such as PPV were used to partially circumvent this). Nonetheless, having more examples of low rapport interactions will improve classifier performance. This improvement is critical if future dialogue systems aim to detect low rapport in its many guises.

Regarding overall prediction methodology, this study used classification, where subjects were divided into high rapport and low rapport classes. However, rapport is not binary; rather, it exists on a continuous scale. For this reason, a regression analysis would also be very appropriate. Moreover, when I was analyzing the 6-dimensional vector for each subject, based on their answers to the 6 survey questions, I also ran a Kaiser analysis to determine the "true" number of factors/dimensions in the 6-dimensional vectors (Auerswald and Moshagen, 2019). Similar to Liebman and Gergle (2016b), I found that because the answers to each question were highly correlated, there is really only a single factor in the answers. The upshot of this is that it would be appropriate to reduce the 6 answers to a single mean, and use that number as the response variable in a regression analysis.

This will be done in a future analysis; for my dissertation I chose to begin the process using less granular clustering.

Regarding specific analyses, it would be helpful to also test two changes. The first change regards the random subset. In these experiments, I only used a single random subset, due to methodological constraints. As a result, my conclusions about the random subset can really only be extended to that specific random subset, rather than random subsetting in general. In future iterations, I would resample a random subset, and test all of these samples against the full dataset.

To extend random subsetting, it would also be helpful to determine what proportion of the full dataset can be randomly selected so as to achieve comparable results. Study 3 used 50% of the data, but it is possible that subsetting only 25%, for example, would achieve similar results while lowering the amount of data that needs to be extracted. Collecting less data would also help to improve the anonymity of the keystroke data, providing greater privacy while still achieving comparable performance (see Manandhar et al. (2019) for an example of a continuous but anonymous authentication system using keystrokes).

Regarding input features, this study only looked at overall ratios, such as word ratios between what a subject produced and their partner produced. However, studies such as Lubold and Pon-Barry (2014) also looked at turn-by-turn ratios, to measure coordination between interlocutors, where coordination is usually a sign of connectedness. This is important if the goal of a future system is to continuously monitor rapport, since a full dataset would not be available in the middle of a conversation.

Moreover, as mentioned multiple times, rapport is a dyadic feature that emerges from an interaction. Therefore, it would be helpful, especially for human-to-human dialogue management, to also take into account the mindset of the partner, based both on final questionnaire data and the partner's typing patterns.