Chapter 6

Study 2: The relationship between keystrokes and sentiment

This study examines how sentiment, sentiment change, and opinions in a conversational dyadic relationship affect the way a user types. This is important for a number of reasons: Most importantly, not all underlying user sentiment is apparent from word choices, and often sentiment based on word choices can only be measured after a complete message or conversation is complete. On the other hand, if sentiment also affects keystroke patterns, then this allows for real-time or continuous sentiment measurement, where user sentiment is being measured *as* a conversation progresses. Since studies such as Lee et al. (2014) and Lee et al. (2015a) have shown that emotion affects the way a user types in isolation, this study's findings about sentiment in dialogue will be fruitful for future developments such as a more empathetic chatbot.

Moreover, a user's underlying opinions are almost never overtly obvious from word choice. Aside from a user uttering a message such as, "I am enjoying this conversation," these underlying opinions are almost never realized in text. However, if keystroke patterns reflect opinions, then it would be possible to use information from typing patterns to deduce underlying opinions.

This study also demonstrates the usefulness of incorporating keystroke data into sentiment analysis for dialogue. While using keystrokes to infer sentiment is not new (e.g. Epp et al., 2011;

Lee et al., 2015a; López-Carral et al., 2019; Vizer, 2009; Yang and Qin, 2021), it has not been extended to conversational data. This seems to be a natural extension of sentiment measurement using keystrokes: If typing patterns are viewed as implicit prosody, and it has long been known that prosody is used to a greater extent in dialogues versus isolated speech (Blaauw, 1994; Bruce and Touati, 1990; Hieronymus and Williams, 1991), then typing pattern differences should be more apparent in dialogue than monologue. As such, I will also demonstrate that adding keystroke data to a lexical text-based classifier can add further accuracy to sentiment prediction in dialogues. The reasoning behind this is that while both word choice and keystroke timing are sensitive to cognitive patterns, each of these might be sensitive to different cognitive (e.g. Logan and Crump, 2011). Therefore, by adding an additional source of cognitive information, a researcher can learn more about underlying sentiment.

Further, conversational settings introduce the notion of sentiment *change*, where we can measure how much the sentiment changed between turns, rather than simply looking at the sentiment of a turn on its own. This is important in an interaction setting, where it is important to understand if conversational partners are on the same emotional level, or different levels as seen in sentiment change between turns. Moreover, sentiment *change* is unique from sentiment *per se* in that a change may be less likely to be evident on a lexical level. For example, a user's sentiment might shift very negatively, although they restrain themselves and use similar or neutral lexical choices. This restraint might affect keystroke production without affecting word choice. This is partially supported by findings such as Lee et al. (2014, 2015a), which show that emotion affects typing patterns. However, the "gold standard" manually-annotated sentiment that I use as an outcome variable is based on lexical evidence, and would therefore not detect these latent signals. Future studies will need to create an outcome variable based on subjective self-reported sentiment change, and then test how well a lexical model and a keystroke model can predict this change, in order to understand the improvement from keystrokes.

Finally, in a social or conversational environment, a user will also develop opinions about their conversational partner, which can affect the sentiment with which the user produces language.

Because of this unique feature of conversational language, I will also show how opinions can interact with sentiment to affect typing patterns, so that in the future typing patterns can be used to infer opinions.

As outlined in Section 3, Study 2 sought to answer the following research questions:

RQ 2a) Does keystroke information provide additional information about message sentiment as well as sentiment change between turns, above standard lexically-determined sentiment values?

RQ 2b) Are typing patterns sensitive to a user's opinion of their partner, when considered independently from the sentiment of a user's utterances?

The study is made up of two experiments:

Experiment 2a aims to replicate previous studies that used keystrokes to predict sentiment classification: Very Negative vs. Very Positive and Extreme (positive or negative) vs. Neutral sentiment. Because exact sentiment prediction can be difficult to judge, these distinctions are also often used in training sentiment models, e.g. Epp et al. (2011). Experiments 2a also examines a unique aspect of conversation interaction: sentiment change. Rather than only considering sentiment in an isolated message, I examine how changes in sentiment from message-to-message also are reflected in typing patterns. I found that adding keystroke information to estimates made by an algorithmic lexical-based model does significantly improve the amount of deviance explained by the baseline lexical model.

Finally, one aspect unique to interactions is that a user forms an opinion of their partner. This opinion may not only affect how a user types in general, but also affect how they type messages of certain sentiments. As an example, perhaps a user already does not like their partner; as a result, no special effort is required to type a negative message, whereas typing a positive message requires great effort. Experiment 2b looks at the distinct influences of message sentiment and a user's opinion of their partner, to see if opinions also have an effect on keystroke timing. While I do find that certain overall opinions have an independent effect on typing patterns, it does not appear that

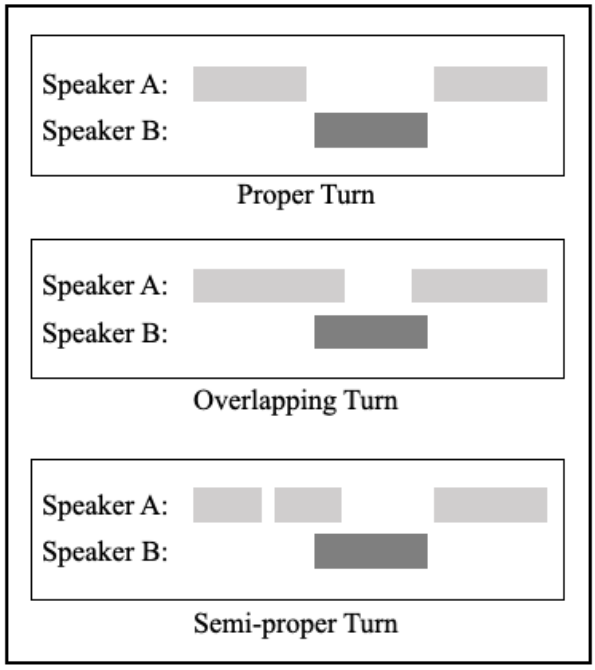


Figure 6.1
Different turn types within a dialogue. In a *proper turn*, which is the only turn type used in Study 2, a turn begins after the preceding turn ends, and then it ends before the onset of the following turn. In an *overlapping turn*, the turn begins or ends while the other interlocutor is typing. In a *semi-proper turn*, a turn begins after the first the first message of the preceding turn is sent, but before the conclusion of the entire turn.

this affect consistently exerts an independent influence. Nonetheless, a number of models show promising independent effects which should be studies further in future studies.

The two experiments use models that flip the dependent and independent variables. Experiment 2a predicts sentiment from a combination of keystroke predictors. Experiment 2b uses sentiment and opinion predictors to predict keystroke timing patterns. The reasoning behind these setups will be expanded upon in the discussion section. In brief, the two experiments are answering different questions: Exp 2a asks whether a set of keystroke features can make distinctions about sentiment; Exp 2b asks if sentiment and parter-opinion have independently robust effects on different keystroke metrics.

This study will focus exclusively on what I term “proper” turns (see Figure 6.1). This is similar to the original conception of turn-taking in the origins of conversational analysis (Sacks et al., 1974;

Current Turn Sentiment	Following Turn Sentiment		
	Negative	Neutral	Positive
Negative	8%	1%	3%
Neutral	2%	14%	4%
Positive	7%	5%	55%

Table 6.1
The sentiment of the current turn, broken down by the sentiment of the following turn. As can be seen, for each sentiment level, the highest proportion of following turns is the same sentiment level. Furthermore, the majority of adjacency pairs (55%) are made up of positive turns followed by positive turns.

Wilson et al., 1984). Overlapping turns, while ubiquitous in naturalistic conversation, also introduce a large amount of variation into language production, before, during, and after the interruption. For example, if a user starts typing while they know their partner is typing, these typing patterns are likely to be different as compared to when they know their partner isn't typing, because the user knows that their partner is also producing language. Similarly, if another message pops up in the middle of a user typing their message, this new message could likely distract the user. In fact, the way that a speaker responds to an interruption is not ubiquitous, but rather highly dependent on cultural norms and gender norms (Tannen, 1984). This additional variability will be expanded in the discussion section of this study (Section 6.4).

While understanding overlapping turns is essential for the future development of using keystroke patterns during online conversations, this is outside the scope of this study.

This distinction is also important because comparing full turns is more similar to comparing asynchronous dialogues such as back-and-forth tweets. A tweet is not made up of a single sentence, but rather multiple ideas, which make them more analogous to full turns. This highlights the importance of studying full turns in Study 2, rather than individual utterances, because full turns provide of view of an entire idea creation rather than a single component message in that idea. In fact, conversational analysis sometimes uses the turn, rather than utterance, as its base unit (Crookes, 1990).

6.1 Related Work: Sentiment and dialogue

As mentioned in Section 2.2.3, keystroke patterns have previously been utilized for sentiment detection primarily when typing monologues in isolation. In this study, though, I highlight two specific extensions of this work: sentiment analysis in dialogue and sentiment analysis using additive models.

6.1.1 Sentiment analysis in dialogue (versus monologue)

As devices such as Alexa and Siri evolve from simple question-answering voice assistants to pseudo-social companions (Pradhan et al., 2019), it is important to detect sentiment in human-computer interactions rather than only detecting sentiment in isolated language production. Bertero et al. (2016) trained a model to perform sentiment analysis alongside speech recognition in a real-time dialogue system. As they point out, when emotion detection can be done quickly, it allows for speech recognition accuracy to also be increased. In other words, rather than simply decoding a speech signal, this decoding can be aided when the emotional dimension of the speech is also provided. My study, in a similar vein, examines the relationship between sentiment and keystroke patterns, so that future researchers can create better computer-agents, such as chatbots, that generate a more relevant and emotionally-appropriate response to a user.

The question remains, though, as to how well research from isolated sentiment analysis can transfer to sentiment analysis in dialogue. As Zhang et al. (2019) points out, sentiment in dialogue is sensitive to both the speaker themselves as well as previous context. Gergle (2017) similarly points out that sentiment in dialogue is simultaneously sensitive to individual-, group-, and even network-level properties. In an isolated setting, however, previous context and other group-level properties does not exist or at least does not exert the same influence on the present language choices.

Ghosal et al. (2020) created an utterance-level model of sentiment analysis in conversations. However, the model was significantly improved by incorporating elements of commonsense knowl-

edge, in order to better understand relationships and sentiment shifts that were not apparent on a purely lexical level. This study has a similar aim, by demonstrating that keystroke knowledge can also complement lexical knowledge to improve sentiment understanding.

Welch et al. (2019) provides an interesting parallel to my own study. The researchers used longitudinal asynchronous dialogue data to predict not only the content of the next message, but also the timing of when the next message would be sent. As a caveat, because their data was asynchronous it was on a very different time scale than my keystroke data. Nonetheless, similar to my own study it shows the use of timing-related data for understanding a dialogue and the temporal or linguistic relationships between messages.

Finally, Ganesan et al. (2022) presents an interesting parallel to my own study of sentiment change. The researchers used large language models such as RoBERTa (Liu et al., 2019) to predict "moments of change" in sentiment, such as a sentiment switch or an escalation of the same sentiment between successive online posts. They compared these results to a transformer model that also incorporated "psychological features" of the user. However, the purely lexical model outperformed this augmented model. Given this finding, it will be informative to see if cognitive-based features are more informative in synchronous conversations as opposed to the asynchronous posts used in Ganesan et al. (2022).

As a last note, it is important to study sentiment in dialogue because dialogues can elicit different emotional reactions than monologues consisting of the same content. Stranc and Muldner (2019) studied student sentiment after watching a teacher deliver a lecture as a monologue versus delivering the same material in dialogue. They found that, while controlling for retention of material, dialogues provided more positive sentiment in student comments after watching the lecture. A study such as this is important because it highlights the need to study sentiment specifically in dialogue, rather than viewing dialogue similarly to monologue.

6.1.2 Sentiment analysis using generalized additive models (GAMs)

Additive models such as those employed in this study are well-suited to capture a nonlinear relationship between typing patterns and sentiment level. Moreover, an additive model is made up multiple smoothing functions, and so it can be used to model phenomena where a latent or unidentified factor is also influencing an outcome. As Hastie and Tibshirani (1987) points out in one of the first studies of GAMs, these models have "the advantage of being completely automatic, i.e. no 'detective work' is needed on the part of the statistician."

As an example of the flexibility of GAMs in sentiment analysis, Qi and Li (2014) used these models to predict sentiment from nouns and noun phrases, rather than the traditional approach of using verbs and adjectives to locate emotional word choices. Because nouns, in isolation, look largely the same in negative and positive language, it was important to use a GAM that could also detect latent factors.

In another interesting application of GAMs, Wang et al. (2022) used these models to determine consumer satisfaction, assuming it was related to, but not identical to, sentiment in reviews. The additive models seemed to be capable of teasing apart these variables.

In my own study I am using dialogue data, where the language a user produces and the way they produce it is inherently complex. Language production in dialogues can be influenced by many factors such as a user's state of mind, by a previous turn, or by the beginning of the conversation. Using GAMs will allow my models to account for multiple variables which may or may not be an explicit part of a model.

6.2 Methodology

Study 2 uses the same data that was collected for Study 1 (see Chapter 4 for details of the data collection procedure). However, rather than studying individual messages (as in Study 1), this study concatenates adjacent messages from the same user into a "turn", and then looks at adjacent pairs of turns, called an *adjacency pair* (Schegloff and Sacks, 1973). A "turn" is equivalent to a sent

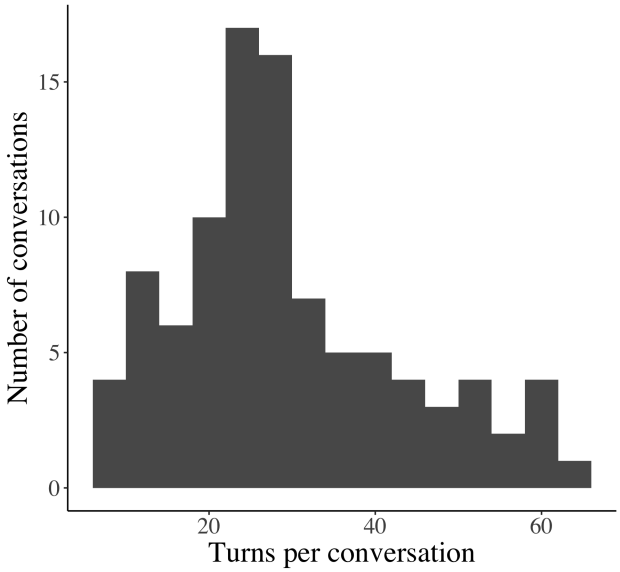


Figure 6.2
The distribution of the number of turns per conversation.

Turn Type	Occurrences	Length (characters)	Word Count
Proper	676	79	16
Overlapping	905	91	18
Semi-proper	1154	88	17

Table 6.2
Occurrence count and features of different turn types. These turn types are illustrated in Figure 6.1.

message. In other words, a turn is not delineated by punctuation, but rather only when the ENTER key was pressed to transmit a message.

The final dataset included 2,890 turns, with a mean word count of 30 words per turn. However, there exists significant variability in the number of turns per conversation: the shortest conversation comprised 8 turns while the longest conversation was 66 turns. The standard deviation for turns in a conversation was 14 turns. The variation in turns per conversation is illustrated in Figure 6.2.

Although there were 2,890 turns in the entire dataset, as Table 6.4 illustrates, each model in Study 2 used only between 262-450 proper turns. The size of this dataset is relatively small compared to datasets used for comparable tasks. For example, the IEMOCAP dataset, considered one of the gold standards in conversational sentiment analysis, was trained on 5,810 utterances (Busso et al., 2008).

As mentioned elsewhere in this thesis, keystroke data has high variability: this is not only because language data is inherently messy, but specifically because timing such as pauses in typing data can occur for many reasons, some of which are impossible to detect. As an example, a typist can pause for cognitive reasons, e.g. thinking of what to say next, physical reasons, e.g. they are tired, or completely unrelated reasons, e.g. distracted by a fly in the room (Dahlmann and Adolphs, 2007; Leijten and Van Waes, 2013). For this reason, the initial filter on all data in Study 2 was to only keep the first 95% quantile of keystroke data, to prevent especially long pauses from skewing the findings, and then center and scale the remaining data (Epp et al., 2011).

To illustrate why this was done, it is unlikely that a 20 second pause had a different underlying motivation than a 10 second pause, and so the distinction is not meaningful. In future work, it may be useful to take an approach similar to Baaijen et al. (2012), in which pauses are binned together, e.g. a bin for pauses less than 1 second, or a bin for pauses between 5 and 10 seconds. By taking this approach, no pauses are eliminated but some precision is lost.

6.2.1 Feature selection

The feature-set for Study 2 was slightly modified from Study 1. Many features were reused when creating the optimal model. However, since this study uses whole turns rather than individual utterances, this introduces certain features that would not have been available when looking at single messages, e.g. the pause been messages or multiple phrasal boundaries.

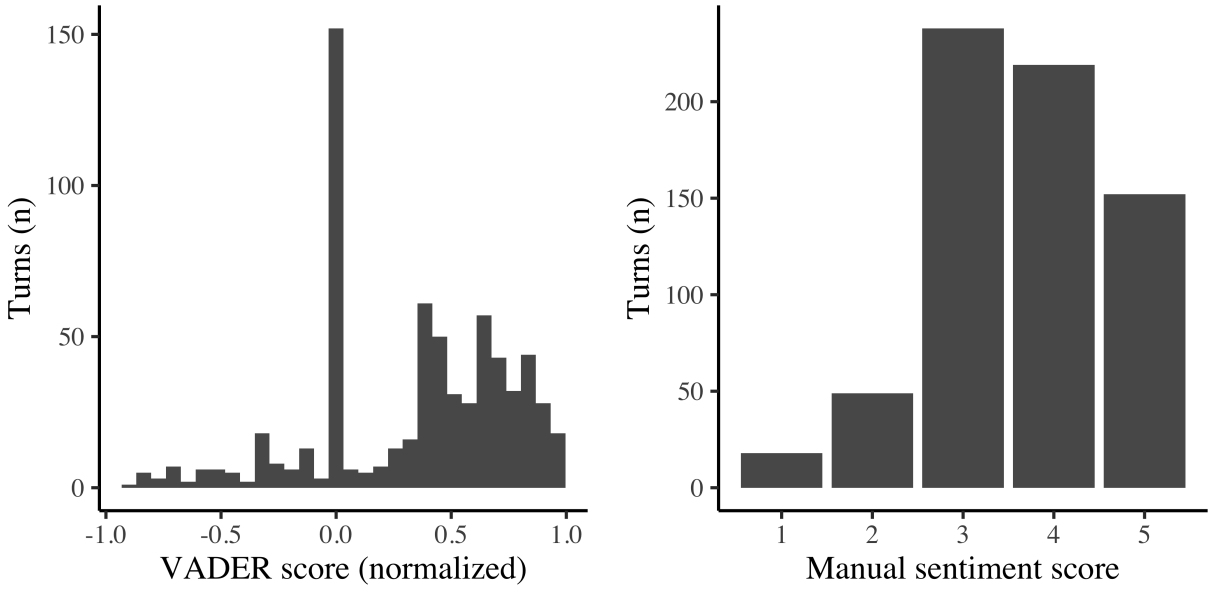
Keystroke timing features were selected primarily for their cognitive relevance, rather than for strictly practical reasons. For example in keystroke dynamics research it is common to measure the timing of every keystroke bigram, trigram, or *n*-gram. Because the current dataset is relatively small, though, this feature would have high variability.[1] The turn features that were tested are below.

- Pauses (inter-keystroke intervals or IKIs) before, within and after a word (Conijn et al., 2019)
- Pauses at phrase, sentence and message boundaries (Galbraith and Baaijen, 2019)
- Dwell time, or how long a key is depressed (Lee et al., 2014, 2015a)

[1] However, Study 3 will use *n*-grams.

- Edits or deletions during a message (Olive et al., 2009)

Galbraith and Baaijen (2019) noted that the importance of different types of pauses depends on where they occur in a text, i.e. in the middle of a sentence, at the end of a sentence, or at a phrase boundary (delimited by a comma or semicolon). In other words, not all pauses should simply be aggregated and averaged, but rather different pause locations imply different reasons for pausing, and these different reasons could be more or less influenced by sentiment change.



(a) **(b)**

Figure 6.3
Figure 6.3a illustrates the distribution of algorithmically-determined sentiment values, using VADER. Figure 6.3b illustrates the manually-assigned sentiment scores, using human annotators. As can be seen, in both instances most turns are neutral, and more turns are positive than negative.

The sentiment measures used for this study were collected in two ways: algorithmically and manually. Algorithmic sentiment analysis was done using the `VADER` Python package (Hutto and Gilbert, 2014).[2]

[2]I also ran a pilot study using the `sentimentr` R package for sentient analysis. However, too many turns were labeled neutral sentiment, but only because the algorithm could not make any decisions, whereas VADER was trained on social media and seems to better understand language that appears neutral on the surface but in informal settings is used to convey sentiment. Despite choosing one algorithm over the other, it should be noted that the results of both were

Manual sentiment analysis was performed by a research assistant and me. Annotation guidelines were drawn up that followed Mohammad (2016) and specified what signals to look for in a turn in order to assign a sentiment score. The inter-annotator agreement, measured by Cohen's κ, was 0.94, which is considered "near perfect agreement" (McHugh, 2012). While I only had one additional annotator available, it should be noted that manual text annotation is essential for sentiment analysis learning (Bobicev and Sokolova, 2017).

6.2.2 Generalized Additive Models (GAMs)

Finally, I chose to use additive models built in `mgcv` in order to capture nonlinearities in the data (Wood, 2022, 2006). In addition, each predictor also output an *effective* degree of freedom (edf), which represents how many knots or change-points actually exist for the predictor. As an example, an edf of 1 would mean that the predictor is essentially linear, whereas an edf of 3 would imply two points at which the slope of the line changed. By having a sense of how many change-points actually exist, it is possible to make further inferences about how many groups actually exist in the data, since a different group would necessitate a change-point.

In addition, GAMs are considered to be an especially interpretable method of machine learning (Chang et al., 2021a). As an example, Hegselmann et al. (2020) presented the visual output of GAMs to healthcare professionals who needed to assess clinical levels of risk. They found that the doctors were able to "mentally simulate" the output of the additive models, and felt comfortable making an informed decision.

All predictors of interest used the default thin plate regression spline, while a smoothing spline was also added to use the individual subject as a random effect. An additional parameter was added to increase penalization so that less informative smooths could be reduced to an edf of zero. Without this additional penalization, a non-informative spline could only be reduced to a linear function (Marra and Wood, 2011). This penalization is similar to LASSO regression for linear models; since

strongly correlated. A Pearson correlation coefficient was calculated at $r(2888) = .58, p < 0.0001$, and a Wilcoxon paired t-test found no significant difference ($p = 0.65$).

the dataset was small, preventing overfitting was important. Aside from these changes, no other hyperparameters were set. Because the dataset was so small, hyperparameter tuning would have most likely had a minimal effect. To assess model fit as well as model comparison, recent work has suggested using AIC comparison as well as ANOVAs (Wood et al., 2016). Both of these tests have been specifically adapted for generalized additive models in the aforementioned `mgcv` package.

6.2.3 Selecting an optimal additive model

To test the various hypotheses, an optimal keystroke model was selected. In order to select this model, the response variable was the full continuous sentiment values in VADER, from -1 to +1; for predictors, different subsets of keystroke patterns were tested in combinations. Ultimately, the most accurate model used four smoothed predictors: overall IKI mean, average dwell time within content words, average IKI before function words, and average IKI at all phrasal boundaries. The model also included a random effect spline for each individual subject. The implications of these predictors providing the most accurate fits will be included in the discussion section.

Predictor	Effective df	Referential df	F score	p-value
Inter-keystroke Interval (IKI)	4.0	9	1.454	0.007
Dwell within content words	2.7	9	1.35	0.003
IKI at beginning of function words	0.8	9	0.195	0.111
IKI at phrasal boundaries	≈ 0.0	9	0	0.892
By-subject	12.7	159	0.089	0.220

Table 6.3
An ANOVA highlighting the importance of each (smoothed) predictor in the optimal keystroke model. While not all predictors attained statistical significance, the combination of these four predictors explained 12.9% of the deviance of the sentiment values. This was the best fitting model of all of those tested. Effective df represents how many change-points were *actually* needed for the predictor's best fit, whereas referential df is the expected degrees of freedom.

This optimal model was compared to a baseline model that used all IKIs and all dwell times as predictors. Compared to this model, a one-way ANOVA comparing the baseline model to the optimal model provided a marginally significant improvement ($p = 0.07$) using a χ^2 test. The

optimal model also had an AIC of 482, whereas the baseline model had an AIC of 487.2, which provides evidence that the additional predictors did not only add complexity while providing a better fit.

Further, the optimal model also used full penalization of less informative splines. The usefulness of penalizing less informative splines is illustrated by fact that the AIC of the optimal model improved from 487 to 482, while the effective degrees of freedom only increased by 4.2. This increase in effective df should be compared to the difference in *referential* df between the models which was 18. The referential degrees of freedom represent the maximum possible increase in degrees of freedom necessary to accommodate the complex model, whereas the difference in effective degrees of freedom shows how many additional degrees were actually required for this extended GAM.

Figure 6.4 below shows the partial effects of each predictor. As can be seen, some predictors have a strong linear or at least monotonic effect, while others are nonlinear and not easily definable by a polynomial. One advantage of additive models is that they provide the ability to capture this.

A *dis*advantage of additive models, however, is that using splines rather than linear predictors makes the effect size and direction of effect difficult to interpret (Wood, 2013). Since non-linear effects, such as IKI time in Figure 6.4, can have varying slopes and directions, an interpretable β coefficient cannot be derived to make clear a predictor's precise effect. As described by Wood (2013), though, the p-values use a Wald t-test that uses a null hypothesis that $s(x) = 0$. A low p-value, then, indicates a low likelihood that the splines that make up the function are jointly zero.

6.3 Results

Experiment 2a examines the degree to which keystroke information can augment lexically-determined sentiment information. Experiment 2b then looks at the effects of users' opinions on different keystroke patterns, when user opinion is considered independently of sentiment information.

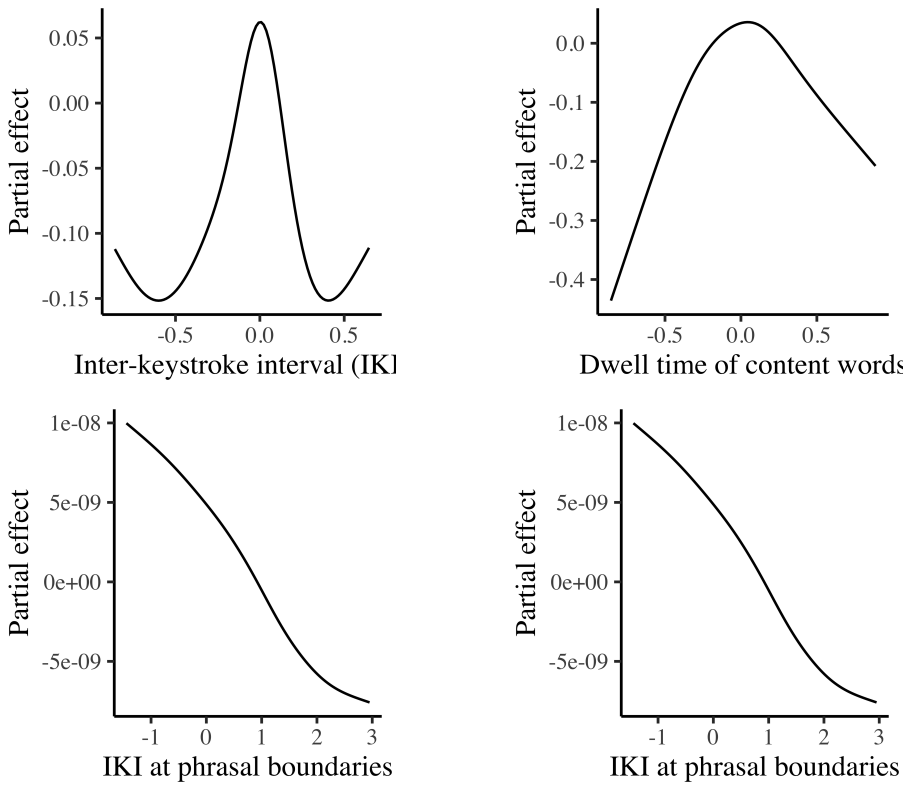


Figure 6.4
The nonlinear functions that define each predictor in the optimal keystroke model

6.3.1 Experiment 2a

The first experiment sought to determine the extent to which adding keystroke information to lexical sentiment information could improve the model's predictive power. The lexical baseline model used traditional algorithmic sentiment analysis based solely on printed text (VADER, Hutto and Gilbert, 2014). Table 6.4 shows the results of adding this information to baseline lexical models for four different sentiment analysis tasks.

Table 6.4 is broken down into four different prediction tasks. The percentage of deviance explained by the model is reported for each model, rather than reporting R^2 or adjusted R^2; the reasoning for this is that R^2-derived measures are only accurate when sums of squares are used, whereas the GAM fitting process makes this an inaccurate measure (Wood, 2006). The base model uses an off-the-shelf sentiment analysis library, VADER, which only considers (surface-level) lexical

Prediction Task	*n* (turns)	Base VADER Model Deviance Explained (edf, AIC)	Keystroke Model Deviance Explained (edf, AIC)	Combined Model Deviance Explained (edf, AIC)	ΔAIC	p-value (χ^2 test)
Exact Sentiment (1 – 5)	450	26.9% (25.2, 1184.6)	15.3% (32.4, 1264.9)	27% (24.7, 1184.5)	-0.1	$p = 0.28$
Positive (4,5) v. Negative (1,2)	303	21.6% (14.3, 230.9)	7.1% (8.5, 256.9)	33% (28.1, 229.1)	-1.8	$p = 0.04^{*}$
Extreme (1,5) v Neutral (3)	262	34.4% (26.1, 287.9)	25.5% (37.4, 342.5)	46.5% (44.6, 281.2)	-6.7	$p = 0.005^{**}$
Sentiment Change (-4–4)	386	12.3% (3.0, 1191.9)	2.8% (4.4, 1234.5)	15.4% (5.7, 1183.3)	-8.6	$p = 0.008^{**}$
Signif. codes: $***-p<0.001, **-p<0.01, *-p<0.05, \dagger-p<0.1$						

Table 6.4

The model results for four prediction tasks. The *n* represents the number of turns, not the participant count. The model results report the percentage of deviance explained by the model, with the effective degrees of freedom and the AIC in parentheses. The *p*-values were determined using an ANOVA comparison of the base model versus the combined model, with a χ^2 test. The influence of the individual predictors used for the combined models are unpacked in Table 6.5. For most models, the addition of keystroke information resulted in a significant improvement in model fit.

content, but takes into account intensifiers and dependencies. The keystroke model uses only the predictors mentioned in the Methodology section that make for an optimal model. Finally, the ΔAIC metric and *p*-value are derived from an ANOVA comparison between the baseline model and combined model, to assess the value of adding keystroke data to a lexical baseline.

The first task had a goal of predicting the exact sentiment of a turn, i.e. 1, 2, 3, 4 or 5. Although the keystroke predictors reduced AIC slightly, by 0.1, this was not statistically significant[3].

The second task used a binomial model to predict either positive sentiment (a value of 4 or 5) or a negative sentiment (a value of 1 or 2). Adding keystroke information resulted in a statistically significant improvement in AIC, by 1.8; deviance explained improved by 11.4% ($p < .05$) from 21.6% to 33%.

The third task also predicted a binomial outcome, i.e. whether the sentiment was extremely negative/positive (a value of 1 or 5) or whether the sentiment was neutral (a value of 3). An ANOVA

[3]This is ironic since the optimal model was fit using this prediction task. Nonetheless, it seems that this task might not be ideally suited to keystroke information: Perhaps the exact sentiment values are too fine-grained to be detected in keystrokes, or perhaps the lexical model is exceptionally good at predicting this.

also showed a statistically significant improvement in model fit, and improved AIC by 6.7; deviance explained improved by 12.1% ($p < .01$), from 34.4% to 46.5%.

The final task predicted the sentiment change between the preceding turn (from another participant) and the current turn, using a continuous measure rather than a categorical measure such as *same* or *different*. This prediction task is unique to dialogues, because language produced in isolation does not have a previous turn produced by a different conversant. Adding keystroke data to this lexical model also provided a statistically significant improvement, and improved AIC by 8.6; deviance explained improved by 3.1% ($p < .01$) from 12.3% to 15.4%.

I will delve further into the final task as well as the prior three in the following discussion, but it is perhaps telling that cognitive signals such as keystroke timing patterns are highly informative as to changing mindsets. On the other hand, a purely lexical analysis may be less effective at detecting changes.

Table 6.5 breaks down the Combined Models in Table 6.4 into individual predictors. Each column in Table 6.5 is a different prediction task, which corresponds to a row in Table 6.4. As can be seen, when combining lexical information (VADER) with keystroke information, the lexical information appears to be more influential than the keystroke information. This can be expected as the VADER model was built with a very large number of parameters and is likely more nuanced. As a note, though, p-values are difficult to interpret when using additive models (Wood, 2013), and so even though these have been specifically built for GAMs, these values should not always be taken at face value. Rather, the most reliable measurements are the ANOVA model comparisons in Table 6.4.

Nonetheless, the fact that keystroke information can still be valuable in addition to lexical information demonstrates that keystroke information is not merely redundant to lexical information, but can provide complimentary information.

The p-values in Table 6.5 are derived from a type III ANOVA of each model that used a χ^2 test. For predicting the exact sentiment score ($p = 0.08$) as well as predicting extreme versus neutral sentiment ($p = 0.01$), the by-subject variation provided significant additional information. In other

Predictor	Exact (F, edf))	Positive v Negative (χ^2, edf)	Extreme v Neutral (χ^2, edf)	Change (F, edf)
VADER	$p \approx 0.0$ **(12.3, 1.2)**	$p < 0.0001$ **(31.2, 2.1)**	$p \approx 0.0$ **(55.4, 3.4)**	$p \approx 0.0$ **(6.2, 0.9)**
IKI	$p = 0.85$ (0.0, 0.0)	$p = 0.15$ (7.4, 4.2)	$p = 0.20$ (16.0, 9)	$p = 0.9$ (0.0, 0.0)
Dwell Time (Content Words)	$p = 0.23$ (0.1, 0.3)	$p = 0.32$ (0.01, 0.0)	$p = 0.20$ (1.1, 0.6)	$p = 0.007$ **(0.6, 1.2)**
Pre-word Pause (Function Words)	$p = 0.89$ (0.0, 0.0)	$p = 0.16$ (9.9, 5.9)	$p = 0.50$ (0.0, 0.0)	$p = 0.24$ (0.1, 0.5)
Boundary pause	$p = 0.80$ (0.0, 0.0)	$p = 0.79$ (0.0, 0.0)	$p = 0.45$ (0.0, 0.0)	$p = 0.02$ **(0.4, 1.0)**
Subject RE	$p = 0.08$ **(0.2, 21.1)**	$p = 0.13$ (17.1, 1.5)	$p = 0.01$ **(41.0, 3.1)**	$p = 0.51$ (0.0, 0.0)

Table 6.5
The table above shows the influence of each individual predictor on the outcome of the combined models (lexical + keystrokes) in Table 6.4. F-scores of 0.0 indicate no variance, while edf's of 0.0 indicate linear predictors.

words, each subject was unique in how their production patterns differed when delineating sentiment in these two tests, as opposed to the delineators in the other two experiments.

The model predicting sentiment change was also notable in that two keystroke-based predictors provided significant additional information: the dwell time within content words ($p = 0.007$) and the pauses at phrasal boundaries ($p = 0.02$).

6.3.2 Experiment 2b

Experiment 2b looks at sentiment and keystroke timing from a different perspective, where sentiment and opinion scores predict keystroke patterns (the inverse of Experiment 2a). The goal of understanding how a participant's opinions about a conversation affect their keystroke timing patterns. Whereas in Experiment 2a, the sentiment value was the response variable and the keystroke

patterns are the predictors, in 2b the keystroke patterns are the response variable and the sentiment value plus participants' opinions are the predictors. This change was made so that the independent effects of many different user opinions could be tested on each keystroke feature, where keystroke metrics are held constant while opinion and sentiment change.

In Experiment 2b, a baseline model measured how well manually annotated sentiment scores could predict keystroke patterns. This baseline was then compared to an expanded model that added a participant's opinion rating as a predictor. By comparing the two models, I am able to better isolate the impact of a user's opinions on the way that they type.

Table 6.6 summarizes the results of Experiment 2b.

	Opinion Question						
Keystroke Feature	Watch with partner	Smooth convo	Enjoy convo	Watch recommendations	Rapport	Mean	Self-opinion
Pre-turn pause	$p = 0.38$	$p = 0.78$	$p = 0.12$	$p = 1.0$	$p = 0.85$	$p < 0.0001^{***}$	$p = 0.62$
IKI	$p = 0.20$	$p < 0.0001^{***}$	$p < 0.0001^{***}$	$p = 0.11$	$p < 0.0001^{***}$	$p < 0.0001^{***}$	$p = 0.16$
Dwell	$p = 0.09^{\dagger}$	$p < 0.0001^{***}$	$p < 0.0001^{***}$	$p = 0.18$	$p < 0.0001^{***}$	$p = 0.08^{\dagger}$	$p = 0.17$
Edit ct	$p = 0.01^{*}$	$p = 0.15$	$p = 0.38$	$p = 0.08^{\dagger}$	$p = 0.09^{\dagger}$	$p = 0.09^{\dagger}$	$p = 0.07^{\dagger}$
Pre-word pause	$p = 0.19$	$p = 0.10$	$p = 1.0$	$p < 0.0001^{***}$	$p = 0.32$	$p = 0.28$	$p = 0.29$
Boundary pause	$p = 0.22$	$p = 0.08^{\dagger}$	$p = 0.43$	$p < 0.0001^{***}$	$p = 1.0$	$p = 0.14$	$p = 0.13$
Before send pause	$p = 0.98$	$p = 1.0$	$p = 1.0$	$p < 0.0001^{***}$	$p = 0.11$	$p = 0.10$	$p < 0.0001^{***}$
Signif. codes: $***-p<0.001, **-p<0.01, *-p<0.05, \dagger-p<0.1$							

Table 6.6

The table above shows the effects of each opinion rating on the timing of the respective keystroke features. Each full opinion question is listed in Section 4.4.5. The *p*-values are derived from an ANOVA comparing a baseline model using just the manual sentiment rating to an expanded model that used the sentiment rating as well as the opinion ratings. The mean score is the average of the first five opinion questions. The final question was not averaged in because it partially depended on self-reflection, rather than the participant's opinion of their parter. As can be seen, a number of different partner opinions affected the inter-keystroke interval and dwell times. In addition, the final question asking what the partner thought of the participant (self-awareness) affected the interval before a participant sent a message.

The responses of interest were:

- Mean pre-turn pause - the mean interval between when the previous message was sent and the current message was begun
- Mean inter-keystroke interval - the mean interval between each keystroke, analogous to typing speed
- Mean dwell time - the mean time of how long a key is pressed, i.e. key-down to key-up
- Edit count - the total number of times a participants uses the BACKSPACE or DELETE keys
- Mean pre-word pause - the mean interval that occurs before a word is typed, i.e. between SPACE and the first letter of the word
- Mean boundary pause - the mean interval surrounding a phrasal boundary, delimited by a comma, period, question mark, etc.
- Mean pre-send pause - the mean interval between when the last character of a message is typed and the ENTER key is pressed to send the message

The full text of the opinion questions that make up the predictors of interest is outlined in Section 4.4.5. An example question was: *Hypothetically, how much do you think you'd enjoy watching a movie with your partner? [1=Not at all, 7=I would definitely enjoy it]*

Table 6.6 shows the significance of different opinions on typing patterns. The values were derived from an ANOVA comparing a baseline model to a test model that added opinion ratings as a predictor. Thus, the *p*-values reflect the influence of the additional opinion predictor in predicting keystroke patterns, or to what extent different opinions affect the way a participant types while controlling for the sentiment of a turn.

As a point of clarification, the column labeled "Mean" is the average of all proceeding opinion values. These questions asked a participant what they thought of the conversation or partner *per se*. On the other hand, the final question also required introspection and so I did not want to confound the other opinion values with this latter value.

Looking at each keystroke pattern (i.e. each row), the overall mean inter-keystroke interval and mean dwell time are strongly influenced by a number of different opinions. Moreover, the

two opinion values that do not have a statistically significant or marginally significant effect on keystroke patterns still have a p-value below 0.20. This result will need to be further investigated in future studies.

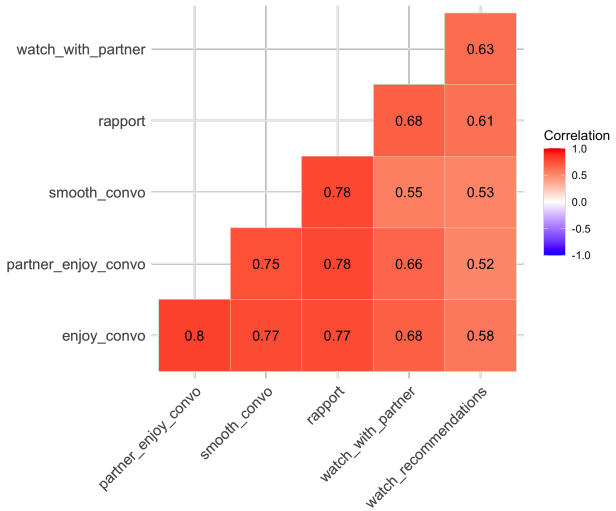


Figure 6.5
The correlation of scores between each opinion question. As can be seen, all of the correlations are very strong in a positive direction. In a post-hoc test, all correlations were statistically significant to an extremely low α.

Before looking at the influence of specific opinion scores, it is also important to note the non-independence of each survey question. As seen in Figure 6.5, all of the answers to the individual questions were strongly positively correlated to one another. The similarity can be seen in Figure 4.8, where the distribution of responses to most questions was very similar.

However, the influence of different individual opinions (i.e. each column in Table 6.6) shows interesting results. Specifically, a participant's opinion on whether they will watch a partner's recommendations is closely related to a number of keystroke patterns, although the direction of causality between opinions and keystrokes is unclear and will require follow-up experimentals. Moreover, the mean opinion score also significantly affects multiple keystroke patterns. This latter result might be the result of the high levels of correlation seen between individual opinion values, as seen in Figure 6.5. Because the mean score essentially accounts for many opinions, it may be

that this multi-faceted score is the most accurate predictor for physical change as manifested in keystroke timing.

A final result is the significant influence of self-awareness on the pause time before a message is sent. One way to look at this is that the less confident a participant is that their partner is enjoying the conversation, the more that participant might hesitate before sending a message.

6.4 Discussion

As a reminder, Study 2 sought to answer the following research questions:

RQ 2a) Does keystroke information provide additional information about message sentiment as well as sentiment change between turns, above standard lexically-determined sentiment values?

RQ 2b) Are typing patterns sensitive to a user's opinion of their partner, when considered independently from the sentiment of a user's messages?

As briefly noted in the introduction to Study 2, I limited my dataset only to *proper* turns, or turns without interruption. This was done in order to avoid introducing further variation into the timing surrounding turns, where a user seeing an "is typing" indicator might disrupt their flow of language production. This is further complicated by the notion that different cultures respond differently to conversational interruption. Tannen (1984) classifies these conversational styles into two groups: "high involvement" speakers do not mind interruptions or overlapping speech and will even intentionally use simultaneous speech to show agreement or enthusiasm; "high considerateness" speakers, on the other hand, are more concerned with being considerate of others and prefer not to impose on the conversation as a whole or on specific comments of another conversant.

These styles of interruption also have an intriguing tie-in with Study 1, on dialogue acts. Choe (2018) and Tannen (1984) point out how an interruption can also be perceived differently depending on whether it references previous context or whether it is completely unrelated and intends to advance the conversation in a new direction. Choe (2018) uses this idea to evaluate "listenership" in

a multi-party instant message conversation, and finds that people naturally find ways of adapting strategies from spoken conversation into text-based conversation in order to demonstrate their level of involvement and engagement. Given these findings from previous research, it will be especially fruitful to evaluate typing patterns in overlapping turns, not only because these turn types constitute the majority of a conversation, but also because different behaviors in response to interruptions could be telling about the cultural expectations of a user.

Before delving into the individual results, the differences between the VADER sentiment model, my own manual sentiment annotation, and other algorithmic models should be further clarified. The primary difference between the VADER ratings and the manual ratings is that the VADER ratings were determined algorithmically while the annotation ratings were done by (a small number of) human annotators. In my experiments the annotation scores were held as a "gold standard" because the human annotators were influenced by not only the explicit lexical content but also implicit connotations of a message, which could be missed by an algorithmic sentiment determination (see Appendix F for the full guidelines, based on Mohammad (2016)). Although it is possible or probable that VADER missed some implicit meanings, the advantage that VADER has over similar software is that VADER was tuned for microblog-like language (Twitter) and its results generalize better over multiple social media domains (Hutto and Gilbert, 2014, p. 216). Since the spontaneous conversations in my experiments use an informal language style that is more similar to social media posts (as compared to the periodicals that similar sentiment libraries were trained on), VADER seemed to be the most germane library for my purposes.

Creating the optimal keystroke model was interesting, specifically which keystroke predictors were the most informative. For example, the optimal model used the dwell time of only the content words. However, in building the optimal model, an earlier model used all dwell times. While overall dwell time was a significantly informative predictor, it also made the by-user random effect less informative. However, when looking at only the dwell time within content words, the by-user random effect was much more significant and the overall model fit improved. This seems in line with similar findings from Lee et al. (2014) and Lee et al. (2015b) which found that dwell is significantly

correlated to emotional arousal at a unique user-by-user level. This makes sense in that emotional arousal is more related to content words (Niederhoffer and Pennebaker, 2002), and so limiting dwell times, connected to emotion, to only content words, should be more informative and more tailored to individuals. In the future if researchers are building an agent that is informed about user emotions from keystrokes, the predictors of the optimal model point to the notion that not all keystroke patterns will be helpful. For example, if dwell times in function words are not unique to individuals then incorporating these dwell times into a user template might make the template less tailored to that particular user.

It should also be noted that the optimal keystroke model created for these experiments was created using only keystroke features. It is possible that when keystroke metrics are used as predictors along with lexical sentiment scores, then a different set of keystroke metrics might be optimal. In other words, there might be significant overlap between keystroke features in the optimal model and lexical sentiment scores. As such, adding keystrokes to sentiment would be largely redundant. This issue should be addressed in future work that builds an optimal model that is also based on a keystroke metric's low correlation to lexical sentiment scores.

Experiment 2a provided an answer to RQ **2a**, showing that keystroke information can provide additional or complimentary information in addition to the information provided by a lexically-determined sentiment baseline value. These results are seen in Table 6.4, where a combined model that includes both lexical information and keystroke information outperforms a model that only uses a standard sentiment analysis library.

Table 6.5 echoes results in previous literature that found that sentiment analysis tools designed to assess sentiment in isolated text could also be useful for conversational dyadic text (Ojamaa et al., 2015; Zhang et al., 2021). Since there exist only a handful of studies that perform sentiment analysis on dialogue, as compared to monologue, this additional validation is also valuable to the general research community.

It is difficult to compare the results of my models to those used for e.g. training VADER (Hutto and Gilbert, 2014), because the datasets were different sizes and the tasks were different.

Nonetheless, it appears that the VADER sentiment analysis library that was trained on isolated data will still provide information on sentiment in conversations.

However, beyond demonstrating the transferability of lexical models, my experiment also illustrated the utility of keystroke models in determining conversation-based sentiment. This can be seen as an extension of the plethora of literature reviewed in Yang and Qin (2021) which has demonstrated that keystroke information can be useful for detecting sentiment in isolated settings. Taken as a whole, Experiment 2a demonstrated that lexical and keystroke sentiment models designed for isolated text can be extended to lexical and keystroke information produced in a conversational setting.

Experiment 2a also provided a significant contribution to existing research by showing that aside from keystroke information being useful for determining the sentiment of a turn *per se*, keystroke information is also useful for predicting how sentiment values change from one user's to the other user's turn. In fact, keystroke information provides the most statistically significant amount of additional information for predicting change, as compared to the other prediction tasks in Experiment 2a. This is similar to the neural network model built in Hazarika et al. (2018), which predicted emotions of utterances based on lexical content as well as language production patterns. Whereas Hazarika et al. (2018) used audio and visual features of a user involved in a spoken conversation, I was able to derive similar results using keystroke patterns. As mentioned previously, monitoring keystrokes is significantly less intrusive than a video camera and microphone, which points to an advantage of the approach taken in this study.

Sentiment change is especially tricky to keep track of and predict within conversations: Text written in isolation proceeds in a somewhat linear and logical fashion; conversely, sentiment in a conversation can jump around, where sentiment is constantly dependent on changing context as well as more or less recent context (see Zhang et al., 2021). When designing a chatbot or an augmented text chat platform, if a human or computer agent could take advantage of keystroke information from an interlocutor to detect when sentiment has shifted, then this could act as a trigger. The

human or computer agent would have evidence that the interlocutor is not on the same page as themselves, and so a shift in tone or content is necessary.

To investigate an additional source of variability in dialogues, Experiment 2b sought to answer the final research question, RQ **2b**. Whereas Experiment 2a demonstrated that keystrokes patterns are informative about sentiment, Experiment 2b answers whether user's opinions of their partner and the conversation itself also affect typing patterns in addition to the sentiment of an utterance.

Table 6.6 shows that the user's opinions are strongly correlated with the way that they type; specifically, opinion scores are related to typing patterns in a way that is independent of the sentiment of the text they are typing. As a concrete example, a user could type something very positive or very negative, and that positive sentiment would be reflected by typing information. However, the sentiment of the text would not be the only factor that contributes to the way they type: the user's opinions of their partner would also independently be associated with the typing metric. This seems to answer RQ **2b**, in that typing patterns are independently sensitive to both the specific utterance sentiment and overall user opinions.

This finding makes sense in light of findings such as those in Gidron et al. (2020) and Barnett et al. (2018), which showed that the rapport between a participant and an experimenter can effect executive function and experimental test results. In my own experiments, then, it would also stand to reason that opinions would effect typing patterns. The reason for this is that typing execution is governed by multiple cognitive processes, including motor execution, lexical recall and executive function (Dagum, 2018; Logan and Crump, 2011; Rumelhart and Norman, 1982). As such, if rapport can interfere with executive function, then changes in rapport levels can likely also change typing patterns. Note, however, that this conclusion does not go so far as to hypothesize about the direction of effect of a rapport-typing connection.

This finding is important if researchers intend to use typing patterns to derive underlying or unobservable motivations. In a spoken conversation, speakers can use auditory cues to infer some of this information, as in Ondobaka et al. (2017). But this information is not readly available in

text-based dialogue. Nonetheless, Experiment 2b seems to suggest that this information is present in the typing signal, and is therefore recoverable and can be made manifest.

Based on the results, it seems that the user's opinions most strongly affect the broad typing patterns, i.e. overall inter-keystroke intervals (IKIs) and overall dwell times. It is possible, though, that my data should have been subset in a different fashion, and so opinions actually affect a different and more specific subset of typing patterns. This should be investigated in future work.

Another interesting result was that multiple user opinions affect the pause time before a user sends a message, i.e. their hesitancy before pressing ENTER. This pause time is the gap between when a user finishes composing a message and they then transmit the message. Since pauses in typing can be a sign of uncertainty (Schilperoord, 2002), a delay in this location could represent uncertainty or a lack of confidence in whether a user wants to send the message they have just composed.

It would be interesting in future work to also further investigate the relationship between a user's opinion of themselves (self-opinions in Table 6.6) and the hesitancy in sending a message. As shown in work such as Jokinen (2010), in spoken language a connection exists between hesitation and uncertainty; it should be investigated as to whether this same relationship holds in text-based communication as well, where a lower self opinion is correlated with greater hesitation in transmitting a message.

Pauses before sending a message are also interesting because my findings leave open the question of whether this hesitation is a "verbal cue" or not. Kalman (2007, 42) points out that, "To be considered a nonverbal cue in text-based CMC, an element must be expressed differentially by different writers or in different contexts, and this variance must be communicated to the reader." Although my experiments used a by-subject random effect, Table 6.5 shows how this random effect was not always significantly influential on the predictive task. Therefore, a more uniform experiment should be performed where more subjects produce a more uniform quantity of samples, which could provide evidence as to how consistent a user is about using this pause, but also how consistently different each user is. This is important because the goal of my research is not just to

elucidate cognitive processes and how they differ under different circumstances, but also which processes are intended to *communicate* underlying emotions, rather than just being a reflection of those motivational differences.

The question still remains though, *Why does all of this matter?* When building computer agents that can form trustworthy and natural-feeling relationships with humans, it is important to not just match wording or sentiment, but also use timing indications such as response times to gauge a user's motivations. Li et al. (2017b), for example, shows the importance of combining these factors to make a human-like robot. Zhao et al. (2016) also shows how rapport in virtual agents is partially based on timing patterns.

The importance of this is highlighted in Bothe et al. (2017), which brings up a critical point relating to emotion detection in dialogue: as humans and computers interact in more high-risk environments, such as a car assembly lines or repairing a satellite in orbit, the detection and conveyance of emotion in language production becomes crucial. As an example of applying my own research, one applicable domain is in telehealth. For a remote healthcare provider, it is important to understand not only what a patient is saying, but also the emotion valence of how they are saying it. Keystroke analysis of dialogues could be an important conveyance of this emotional content.

As a similar example, an important underlying trait that can be detected in typing patterns is cognitive load (Brizan et al., 2015, inter alia). Khawaji et al. (2014) goes even further, and shows how keystroke and mouse movements reflect both trust and cognitive load. These two human factors could be important in a setting where e.g. a repair-person is delivering complex directions via CMC or a doctor is involved in a difficult discussion: it is important to be able to judge both the opinion of an interlocutor as well as the degree to which they can comprehend the conversation, where an understandable conversation could be represented by a lighter cognitive load.

These findings are also important in light of findings such as those in Bos et al. (2002), which examined trust development in four different communication settings: face-to-face, video, audio, and text chat. They found that the emergence of trust was the worst and slowest in text chat. However, findings such as those in my Experiment 2b could be used to make user trust more salient

and thereby make trust development or lack of trust more obvious to a user or their conversational partner. This could allow the user to consciously or intentionally improve their level of trust.

Synthesizing the findings of both experiments, though, it seems that keystroke patterns provide information about underlying social motivations. This information includes both a more nuanced sentiment, as compared to what can be detected lexically, as well as overarching opinions of a conversational partner which may never be overtly realized in a conversation.

6.4.1 Unexpected findings

The abundance of neutral messages was surprising, in that I had anticipated a much more "lively" discussion about TV and movie preferences since they can be very personal and strongly held beliefs. However, this distribution was similar to that in Ojamaa et al. (2015), which also found that the majority of the turns in their data had neutral sentiment. For the sake of Study 2, the abundance of neutral sentiment was less than ideal, since it made the examples of strong sentiment more rare. In retrospect this makes sense, in that the goal of most turns is simply to advance the conversation, e.g. forward functioning dialogue acts in Study 1 (Chapter 5).

Similarly, as illustrated in Figure 6.3a (algorithmic sentiment analysis) and Figure 6.3b (manual sentiment analysis), a large proportion of turns have positive sentiment, and a majority of survey opinions are positive. As noted by Svennevig (2014, 302), the goal of a first-encounter conversation between strangers (like those in my experiment) is to establish common ground and establish a relationship of cooperativeness. Thus this large positive skew should have been anticipated, and a more deliberate methodology should have been employed to elicit negative utterances and negative partner opinions.

6.5 Future work

One of the most important next steps in this research is to establish not just a correlation between underlying motivations and typing patterns, but to also establish a direction of causality. Similar to a

study such as Liebman and Gergle (2016b), an intervention should be added to these conversations so that a conclusion can demonstrate whether changing sentiment causes changing typing patterns, or whether a (preconceived) opinion causes changes in typing patterns. In my own study, it seems safe to conclude that both of these factors are correlated to changes in typing patterns, although the experiments lack the ability to say whether sentiment or opinion causes changes in typing.

As workplaces become more distributed (Pew Research Center, 2020; Teevan et al., 2022), it will also be important to study how sentiment and opinion influence typing in larger groups, such as a meeting with more than two participants. In addition, it is important to study typing patterns in situations where power dynamics are more prominent. For example, Willemyns et al. (2003) studied boss-employee trust relationships through the lens of communication accommodation theory. It points to the necessity of more conscientious discourse management in order to establish and build this type of relationship, especially in a distributed environment. In general, this type of study should move past first-encounter dialogues (Ojamaa et al., 2015), and be extended to longitudinal studies that investigate how relationships and typing change over time, as familiarity between users develops.

Finally, the research in Study 2 must be extended to overlapping turns. As mentioned in numerous studies such as Gravano and Hirschberg (2011), the majority of turns in a dialogue overlap one another. Therefore, limiting a study or (eventual) tool to only "proper," non-overlapping turns artificially eliminates a large proportion of a dialogue.

Part of my rationale for not studying overlapping turns in this initial study is that if a user is typing, and then sees a typing indicator pop up for their interlocutor, this may be a distraction or cause the user to stop typing and wait for their interlocutor. However, a future study should investigate user behavior when a typing indicator appears. Tegos et al. (2020) showed that most users rely on a typing indicator popping up as proof that their interlocutor is engaged in the conversation. One could imagine in the case of investigating typing and underlying relationships that if rapport level is high, then a user will not feel to stop typing and wait for their interlocutor to send a message,

whereas if a user is less confident in their relationship, they might stop and wait to see what their interlocutor says before proceeding.