“Love ya, jerkface”: using Sparse Log-Linear Models to Build Positive (and Impolite) Relationships with Teens

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, Justine Cassell
School of Computer Science, Carnegie Mellon University
{yww, slfink, aeo, awb, justine}@cs.cmu.edu

Abstract

One challenge of implementing spoken dialogue systems for long-term interaction is how to adapt the dialogue as user and system become more familiar. We believe this challenge includes evoking and signaling aspects of long-term relationships such as rapport. For tutoring systems, this may additionally require knowing how relationships are signaled among non-adult users. We therefore investigate conversational strategies used by teenagers in peer tutoring dialogues, and how these strategies function differently among friends or strangers. In particular, we use annotated and automatically extracted linguistic devices to predict impoliteness and positivity in the next turn. To take into account the sparse nature of these features in real data we use models including Lasso, ridge estimator, and elastic net. We evaluate the predictive power of our models under various settings, and compare our sparse models with standard non-sparse solutions. Our experiments demonstrate that our models are more accurate than non-sparse models quantitatively, and that teens use unexpected kinds of language to do relationship work such as signaling rapport, but friends and strangers, tutors and tutees, carry out this work in quite different ways from one another.

1 Introduction and Related Work

Rapport, the harmonious synchrony between interlocutors, has numerous benefits for a range of dialogue types, including direction giving (Cassell et al., 2007) or contributing to patient recovery (Vowles and Thompson, 2012). In peer tutoring, an educational paradigm in which students of similar ability tutor one another, friendship among tutors and tutees leads to better learning (Gartner et al., 1971). With the burgeoning use of spoken dialogue systems in education, understanding the process by which two humans build and signal rapport during learning becomes a vital step for implementing spoken dialogue systems (SDSs) that can initiate (and, as importantly, maintain) a successful relationship with students over time. However, implementing a tutorial dialogue system that appropriately challenges students in the way that peers do so well (Sharpley et al., 1983), while still demonstrating the rapport that peers can also provide, calls for understanding the differences in communication between peer tutors just meeting and those who are already friends.

The Tickle-Degnen and Rosenthal (1990) model provides a starting point by outlining the components of rapport, including the finding that positivity decreases over the course of a relationship. The popularity of this model, however, has not diminished the disproportionate attention that positivity and politeness receive in analyses of rapport (Brown and Levinson, 1978), including in the vast majority of computational approaches to rapport-building in dialogue (Stronks et al., 2002; Johnson and Rizzo, 2004; Bickmore and Picard, 2005; Gratch et al., 2006; McLaren et al., 2007; Cassell et al., 2007; Baker et al., 2008; Bickmore et al., 2011). The creation and expression of rapport is complex, and can also be signaled through negative, or impolite, exchanges (Straehle, 1993; Watts, 2003; Spencer-Oatey, 2008) that communicate affection and relationship security among intimates who can flout common social norms (Culpeper, 2011; Kienpointner, 1997).

However, it is an open question as to whether such rudeness is likely to impress a new student on the first day of class. We must better understand how and when impoliteness and other negative dialogue moves can contribute to the development and expression of the rapport that is so important in educational relationships. In this analysis, then, we begin with a corpus of tutoring chat data annotated with a set of affectively-charged linguistic devices (e.g. complaining, emoticons), and then differentiate between the linguistic devices that friend and stranger interlocutors employ (with friendship standing as a proxy for pre-existent rapport) and the resulting social effects or functions of those devices on the partners.

Since our ultimate goal is to build an SDS that can adapt to the user’s language in real time, we also automatically extract lexical and syntactic features from the conversations. And, in order to determine what the system should say to evoke particular responses, we predict social effects in partner two from the use of the linguistic devices in partner one.

Since we want to understand how the system can deal with newly met peers as well as peers who have become friends, we develop and evaluate our model on dyads of friends and then evaluate the same model with dyads of strangers, to examine whether dyads with less a priori rapport react differently to the same linguistic devices.

Of course, in addition to understanding the phenomenon of rapport in all of its complexity, a major challenge for building rapport-signaling SDS is to construct a compact feature space that capture only reliable rapport signals and generalizes well across different speakers. Of course phenomena such as insults, complaints and pet names, no matter how important, appear relatively rarely in data of this sort. Training discriminative models with maximum likelihood estimators (MLE) on such datasets usually results in assigning too much weight on less frequent signals. This standard MLE training method not only produces dense models, but may also overestimates lower frequency features that might be unreliable signals and overfit to a particular set of speakers. In recent studies on speaker state prediction that use lexical features, it has been shown that MLE estimators demonstrate large performance gaps between non-overlapping speaker datasets (Jeon et al., 2010; Wang et al., 2012a).

On the other hand, recent studies on ℓ_1/ℓ_2 based group penalty for evaluating dialogue systems (González-Brenes and Mostow, 2011), structured sparsity for linguistic structure prediction (Martins et al., 2011), and discovering historical legal opinions with a sparse mixed-effects latent variable model (Wang et al., 2012b) have all shown concrete benefits of modeling sparsity in language-related predictive tasks. We therefore apply sparsity-sensitive models that can prevent less frequent features from overfitting. We start with the ℓ_1-regularized Lasso (Tibshirani, 1994) model, since, compared to other covariance matrix based sparse models, such as sparse Principal Component Analysis (PCA) and sparse Canonical Correlation Analysis (CCA), the Lasso model is straightforward and requires fewer computing resources when the feature dimension is high. Hence, we compare the contributions of both automated features and annotated features using the proposed Lasso model to predict impoliteness and positivity.

In addition to Lasso and a logistic regression baseline, we introduce two alternative penalty models: the non-sparse ridge (le Cessie and van Houwelingen, 1992) estimator, and an elastic net model (Zou and Hastie, 2005). The ridge estimator applies a quadratic penalty for feature selection, resulting in a smooth objective function and a non-sparse feature space, which can be seen as a strong non-sparse penalty model. We investigate the elastic net model, because it balances the pros and cons of Lasso and ridge estimators, and enforces composite penalty. In addition to the model comparisons, by varying the different sizes of feature windows (number of turns in the dialogue history), we empirically show that our proposed sparse log-linear model is flexible, enabling the model to capture long-range dependency.

This approach also allows us to extend previous work on speaker state prediction. Although speaker state prediction has attracted much attention in the dialogue research community, most studies have focused on the analysis of anger, frustration, and other classic emotions (Litman and Forbes-Riley, 2004; Liscombe et al., 2005; Devillers and Vidrascu, 2006; Ai et al., 2006; Grimm et al., 2007; Gupta and Nitendra., 2007; Metallinou et al., 2011). Recently, Wang and Hirschberg (2011) proposed a hierarchical model that detects level of interest of speakers in dialogue, using a multistream prediction feedback technique. However, to the best of our knowledge, we are among the first to study the problem of automatic impoliteness and positivity prediction in dialogue. Because our ultimate goal is to build an SDS that responds to users' language use over time, the features from the user's target turn that the model is aiming to predict are not observable, which renders the task more difficult than previous speaker state detection tasks.

Our main contributions are three-fold: (1) analysis of linguistic devices that function to signal rapport among friends - and their effects on non-friend dyads; (2) detailed analyses of language behavior features that predict these rapport behaviors - both impoliteness and positivity - in the next turn of teenagers' peer tutoring sessions; (3) an evaluation of non-sparse and sparse log-linear models for predicting impoliteness and positivity.

By understanding the signals of rapport that a person is likely to display in response to various linguistic devices, we can begin to build an SDS that can anticipate the social response and adapt to the rapport-signaling efforts of its partner, both as a newly introduced technology, and, over time, as a system with whom the user has a rapport.

2 The Corpus

We use the data from a previous study evaluating the impact of a peer tutoring intervention that monitored students' collaboration and in some cases provided adaptive support (Walker et al., 2011). In the intervention, peer tutors observed the work of their tutee and supported them through a chat interface as they completed algebra problems. The system logged all chat and other information about the problem steps. Participants were 130 high school students (81 female) in grades 7-12 from one American high school with some prior knowledge of the algebra material. Participants were asked to sign up for the study with a friend. Those who were interested but were unable to participate with a friend, were matched with another unmatched participant. In an after-school session, participants first took a 20-minute pre-test on the math concepts, and then spent 20 minutes working alone with the computer to prepare for tutoring. One student in each dyad was then randomly assigned the role of tutor, while the other was given the role of tutee, regardless of relative ability. They spent the next 60 minutes engaging in tutoring. Finally, students were given a domain posttest isomorphic to the pretest.

54 dyads signed up as **friends** and 6 were unmatched **strangers**. To compare behavior between friends and strangers in the face of very different data set sizes we use 48 friend dyads for training, and select 6 friend and 6 stranger dyads as two separate test sets. The total number of utterances in the friend training set, friend test set, and stranger test set are 4538, 468 and 402. To perform turn-based prediction experiments, we concatenate the text in the utterances by the same speaker into a single turn, and perform an "OR" operation[1] on features (See Section 3 for details) in multiple utterances of the same speaker to generate the turn-based binary features.

3 Feature Engineering

In this section, we describe both the annotated and automatically extracted features analyzed.

3.1 Annotated Features and Labels[2]

To understand what linguistic devices participated in positivity and impoliteness during tutoring, we annotated all 60 dyads for surface-level language behaviors such as complaints, challenges (Culpeper, 1996) and praise. We also automatically identified chat features that socially color the communication, such as excessive punctuation[P] or capitalization[Ca]. Utterances could receive more than one code, and inter-rater reliability ranged from K=.71 to K=1.

Because these linguistic behaviors may serve a range of different functions in context, such as rude language serving to cement a relationship (Ardington, 2006), or teasing to increase rapport (Straehle, 1993), we also annotate the **social functionality** of each utterance in context, in terms of positivity (K=.79)[3] and impoliteness (K=.76), which are seen as holding down opposite kinds of social functionality (Terkourafi, 2008). Details of annotation can be found in our recent work (Ogan et al., 2012).

Language Behavior Features

Language behavior features were annotated by two raters, based on previous work on impoliteness (Culpeper, 1996), positivity (Boyer et al., 2008), and computer-mediated communication (Herring and Zelenkauskaite, 2009), as follows:.

- Insults[Di] (κ=1): Personalized negative vocatives or references. *eg. "you are so weird."*
- Challenges[Ch] (κ =.91): Directly questioning partner's decision or ability. *eg. Partner 1: "see I am helping", Partner 2: "barely."*
- Condescensions / brags[C] (κ=1): Asserting authority or partner's inferiority. *eg. Tutee: "nothing you have done has affected me what so ever."*
- Message enforcer[Ef] (κ=.85): Emphasizing text or attracting partner's attention. *eg. "Earth to Erin."*
- Dismissal / Silencer / Curse[Cu] (κ =.76): Asserting unimportance of contribution/partner. *eg. "shuttttt up computer."*
- Pet name[Pe] (κ = .9): Vocatives that may or may not be insulting. *eg. "whats up homie?"*
- Criticisms / exclusive complaints[EC] (κ=.8): Negative evaluation of partner. *eg. "You are so bad at this dude."*
- Inclusive complaints[I] (κ=.78): Complaints directed outside the partner, such as at the task, computers, or study. *eg. "This is really dumb, ya think?"*
- Laughter[L] (κ=1): *eg. "haha", "lol"*
- Off-task[O] (κ=.71): Doesn't pertain to or advance tutorial dialogue. *eg. "Coming over after this?"*

Impoliteness and Positivity Labels

While the surface-level features were coded based on a single utterance, context determined the labels for impoliteness and positivity, including the recent tone of the dialogue and the partner's response to the utterance. Utterances were coded as positivity (κ=.79) when they included goals that directly added positive affect into the exchange through praise, empathy, reassurance, cooperative talk (McLaren et al.,

[1] If any of the utterances within one turn has this feature turned on, then we say that we have observed this feature in this turn.

[2] We thank Erin Walker for data collection and annotation.

[3] We use Cohen's kappa in this study.

2011), task enthusiasm, and making or responding to jokes. Impoliteness (κ=.76) included both cooperatively rude utterances such as teasing (typical eg. “hahah you’re the worst tutor ever”) and uncooperatively rude utterances that may cause offense (typical eg. “um why don’t you try actually explainin urself..”) (Kienpointner, 1997).

3.2 Automated Features

To compare the performance between what could be automatically extracted from dialogue and hand annotation, we extracted 2,872 unigram and 12,016 bigram features from the text corpus. Using the Stanford PoS tagger[4] with its attached model, we also extracted 46 common part-of-speech tags from the text. In addition to the above lexical and syntactic features, we automatically extracted the capitalization features[Ca] that have at least one full word (eg. “CALM DOWN”) (Chovanec, 2009). Since a recent text prediction task (Wang and McKeown, 2010) observed benefits from modeling punctuation features[P], we extracted the expressive punctuation that included at least one exclamation point or more than one question-mark (eg. “I don’t get it?!??!”) (Crystal, 2001). We used a smiley dictionary[5] to extract the emoticons[E] that convey emotional states (Sánchez et al., 2006) from text.

4 Sparse Log-Linear Models

We formulate our impoliteness and positivity prediction problems as binary classifications. To do this, we estimate the label $\hat{y}_t \sim Bernoulli(\hat{\theta})$. First, we introduce a standard log-linear parametrization[6] to our predictive tasks:

$$\hat{\theta}_{\vec{y}_t} = \frac{\exp \sum_i \vec{w}_i \vec{f}_i(\vec{y}_t)}{1 + \exp \sum_i \vec{w}_i \vec{f}_i(\vec{y}_t)}, \quad (1)$$

where $\vec{f}(\vec{y}_t)$ is a set of feature functions computed on the observation vector $\vec{y}_t$. The term $\vec{w}_i$ puts a weight on feature i for predicting impoliteness, and our estimation problem is now to set these weights. The log-likelihood and the gradient are:

$$\ell = \sum_t y_t \log \hat{\theta}_{\vec{y}_t} + (1 - y_t) \log(1 - \hat{\theta}_{\vec{y}_t}) \quad (2)$$

$$\frac{\partial \ell}{\partial \vec{w}} = \sum_t \left(\frac{\partial \hat{\theta}_{\vec{y}_t}}{\partial \vec{w}} \right) \left(\frac{y_t}{\hat{\theta}_{\vec{y}_t}} - \frac{1 - y_t}{1 - \hat{\theta}_{\vec{y}_t}} \right) \quad (3)$$

$$\frac{\partial \hat{\theta}_{\vec{y}_t}}{\partial \vec{w}} = \left(\hat{\theta}_{\vec{y}_t} - (\hat{\theta}_{\vec{y}_t})^2 \right) \vec{f}(\vec{y}_t), \quad (4)$$

so the parameters can be set using gradient ascent. To control the overall complexity, we can apply regularized models on the elements of $\vec{w}$. A sparsity-inducing model, such as the Lasso (Tibshirani, 1994) or elastic net (Zou and Hastie, 2005) model, will drive many of these weights to zero, revealing important interactions between the impoliteness/positivity label and other features. Instead of maximizing the log-likelihood, we can minimize the following Lasso model that consists of the negative log-likelihood loss function:

$$\min \Big(-\ell + \sum_i \lambda_1 ||\vec{w}_i|| \Big) \quad (5)$$

Since the Lasso penalty can introduce discontinuities to the original convex function, we can also consider an alternative non-sparse ridge estimator (le Cessie and van Houwelingen, 1992) that has the convex property:

$$\min \Big(-\ell + \sum_i \lambda_2 ||\vec{w}_i||^2 \Big) \quad (6)$$

In addition to the Lasso and ridge estimators, the composite penalty based elastic net model balances the sparsity and smoothness properties of both Lasso and ridge estimators:

$$\min \Big(-\ell + \sum_i \lambda_1 ||w_i|| + \sum_i \lambda_2 ||w_i||^2 \Big) \quad (7)$$

Our log-linear model is quite flexible; by comparing various restrictions, we can test different features when modeling impoliteness and positivity. In addition, the model can incorporate features from previous time windows, which requires much less computational complexity compared to standard high order Markov models. We use the L-BFGS method (Liu and Nocedal, 1989) for the numerical optimization.

5 Empirical Experiments

We predict impoliteness vs. non-impoliteness and positivity vs. non-positivity of an interlocutor in the immediate future turn, given only information from current/previous turns. Because accuracy, precision, recall and F-measure are threshold-based point estimation metrics that might prevent one from observing the big picture of system performance, we consider the Receiver Operating Characteristic (ROC) metric to evaluate the dynamics of the true positive rate vs. the false positive rate (Hanley and McNeil, 1982) in our system. We mainly use Area Under Curve (AUC) as a metric to compare classifiers, since it maps the ROC metric to a single scalar value representing expected performance. A random classifier will have an AUC of 0.5 (Fawcett, 2006).

[4] http://nlp.stanford.edu/software/tagger.shtml

[5] http://www.techdictionary.com/emoticon.html

[6] We thank Jacob Eisenstein for the formulation of logistic regression model.

Models	P	Ca	E	L	O	Ef	Pe	Di	C	EC	Ch	Cu	I
Impoliteness Prediction													
Tr-Te	.44	-1.10	.62	.72	.09	.64	.09	1.29	.96	.89	.69	.77	-0.19
Te-Tr	-2.48	.54	-0.26	0.15	.59	1.62	.24	.22	.89	.72	.75	.04	-0.18
Positivity Prediction													
Tr-Te	-0.87	.19	.36	.55	1.06	-0.62	.69	-1.63	-1.57	.16	-0.41	1.22	.86
Te-Tr	-1.39	-0.46	.70	.48	.46	.33	.62	-0.71	.70	-0.65	-0.47	-0.54	.78

Table 1: Comparing the Learned Weights of Different Features when Predicting the Partner's Impoliteness in a Non-Sparse Log-Linear Model. *Tr-Te: predict tutee turn with tutor turn. Te-Tr: predict tutor turn with tutee turn. For full name of features, see Section 3.*

5.1 Comparing the Learned Weights of Different Features

In our previous analysis of these data (Ogan et al., 2012), a PCA method allowed us to group linguistic behaviors in order to address the issue of data sparsity. With the use of log-linear models, we are able to investigate the contributions of individual language behaviors in one student's turn to the prediction of social functions in their partner's next turn. In this experiment, we evaluate the weights of various linguistic devices in a standard logistic regression model. We found that behaviors commonly associated with impoliteness were predictors of partner impoliteness in the next turn, while positive behaviors such as laughter were predictors of upcoming positivity. SDSs can leverage this knowledge to take the partners lead during a tutoring session, using the partners positivity or impoliteness to determine the affect of the systems upcoming move. As we intend to develop a system that acts as a tutee, however, we further divided the analysis by tutoring role, investigating how partners in different roles employ language features differently, such that the system can act in accordance with its given role. Table 1 shows the results.

Similarly to the collapsed factors in our previous work, we found here that tutors and tutees do in fact use language behaviors differently, and to accomplish different social functions. Effectively, this means that certain language behaviors may instigate impoliteness when said by one partner, but lead to positivity when expressed by the other. For example, tutee bragging predicts a response of positivity on behalf of the tutor ($\vec{w}_C^{(TE)}$ = .7), perhaps because the tutor wants to be supportive of a protégé's self-efficacy and success. Conversely, when the tutor brags during a peer tutoring dialogue, the tutee, who may feel threatened by the tutors bravado, is extremely likely to respond with impoliteness ($\vec{w}_C^{(TR)}$ = .96). In a peer tutoring paradigm, when the more powerful partner (the tutor) expresses dominance through self-inflation, the subordinate partner may use impoliteness to regain some social control. On the other hand, some language behaviors actively work to tear down this power imbalance, such as inclusive complaining, where the partners take an us against the task approach, building solidarity through complaining about the experiment. These utterances predict positivity whether used by the tutor ($\vec{w}_I^{(TR)}$ = .86) or tutee ($\vec{w}_I^{(TE)}$ = .78). Other comparisons between weighted features by role demonstrate similarly theoretically-motivated findings that shed light on how language is used to achieve social functions.

5.2 Comparing the Contributions of Different Features on Friend and Stranger Datasets

A previous study (Ogan et al., 2012) on these same data seemed to indicate that negative conversational strategies composed of linguistic devices such as complaining and insults were correlated with learning in the friend dyads and negatively correlated with learning in strangers. However the small number of stranger dyads prevented them from drawing conclusions about particular linguistic devices from the data. Here, we empirically show the predictive performance of different feature sets on both friend and stranger test sets in Table 2 , using a sparse Lasso model with features from only the current turn. In the impoliteness prediction task, when predicting on the test set that consists of only friends, we observe statistically significant improvement over a random baseline, using surface-level language behavior features, lexical, lexical + syntactic, all automatic, and all features. When combining all features, the best AUC is .621. The automatic features, mainly including n-grams and part-of-speech tags, have emerged as a useful automated feature space. On the other hand, we do not observe any significant results on the stranger datasets, suggesting that strangers do not respond with impoliteness in the same way that friends do. When predicting positivity on the friend dataset, we see that the performance of surface-level language behavior features has dropped from the first task, and the statistical t-test is non-significant when comparing to a random baseline. This is not surprising, because we have shown in the previous section that surface-level language behavior features are strong indicators of impoliteness, but might not have advantages in predicting positivity for friends. Interestingly, the automated features outperform the combination of all features, indicating a promising future for the actual deployment of an SDS that can interact using appropriate positivity and impoliteness.

When predicting positivity in the stranger dataset, we find the opposite trend. In contrast to the impoliteness prediction task, the overall performance on the stranger dataset improved, and the lexical, lexical+syntactic, and all feature combination have significantly outperformed the chance baseline. These results suggest that positivity is a predictable behavior among strangers, who may all express uniform positivity across all dyads, while it is the impoliteness that is predictable among friends. Perhaps it is that through the development of a rapport with a partner, the particular ways in which positivity is expressed becomes personalized to the dyad, and can no longer be applied to other groups who have their own expressions of positivity. In other words, unlike in Tolstoy's world, here unhappy families are all alike; every happy family is happy in its own way. We must look to the easily-predictable impoliteness among friends instead, arguing strongly for the inclusion of impoliteness in a model of rapport.

Feature Sets	F-AUC	p	S-AUC	p
Impoliteness Prediction				
Random	.500	-	.500	-
Behavior	.596	.017	.505	.473
Lex	.599	.014	.435	.819
Lex + POS	.605	.009	.425	.857
All Auto	.591	.022	.451	.751
All Features	.621	.003	.427	.850
Positivity Prediction				
Random	.500	-	.500	-
Behavior	.549	.141	.527	.302
Lex	.623	.003	.601	.025
Lex + POS	.646	.001	.587	.047
All Auto	.651	.001	.577	.070
All Features	.641	.001	.608	.019

Table 2: Comparing contributions of different feature streams on both friend and stranger testsets with Lasso model when predicting impoliteness and positivity of the next turn using only features from the current turn. *(F-: the friend test set. S: the stranger test set. p: one-tailed p-value by comparing to a random classifier. Behavior: detailed surface-level language behavior features defined in Section 3. Lex: unigram and bigram. POS: part-of-speech features. All Auto: all automatically extracted features (Lex + POS + punctuation + caps + emoticons).)*

5.3 Comparing Logistic Regression, Lasso, Ridge, and Elastic Net

While our previous work (Ogan et al., 2012) demonstrated that PCA is a useful feature selection method when there are only a dozen features, in this experiment, the dimension of our feature space is substantially higher, which aligns to the size of vocabulary. Thus, covariance-based feature selection methods, such as PCA, might be too slow. Here we compare the performances of standard MLE trained logistic regression, Lasso, non-sparse ridge, and elastic net models. In particular, we demonstrate the predictive power of Lasso and elastic net models, varying distinct levels of sparsity. In the Figure 1, we show the comparison of three different models in the impoliteness prediction task. The horizontal axis represents different values of regularization coefficient λ. For the Lasso model and the elastic net model, increasing the value λ will result in a sparser feature space, and we set the $\lambda = \lambda_1 = \lambda_2$ in the elastic net model to promote same level of sparsity and smoothness. The result at $\lambda = 0$ represents the standard non-sparse logistic regression model, which obtains an AUC of .563. When introducing penalty for large weights in this standard model, .4 to .5 significant improvements ($p = .003$ for Lasso, $p = .007$ for ridge, and $p = .004$ for elastic net) of AUC are observed from Lasso, ridge and elastic net models when $\lambda = 1$. The elastic net model that balances sparsity and smoothness, has obtain the best result in this experiment. The best result of elastic net model is .63 when $\lambda = 7$. This experiment shows that all three penalty models have outperformed the non-sparse logistic regression model. The elastic net model, which balances sparisty and smoothness, obtains the best results when predicting impoliteness. Figure 2 shows the comparison of three models on the friend dataset in the positivity prediction task. When $\lambda = 0$, the standard logistic regression model has an AUC of .638. When increasing the λ to 1, both Lasso and elastic net models have shown significant improvements (both $p < .001$) in AUC, but not the non-sparse ridge estimator. The Lasso model is found to be the best model in this task: we obtain better results when the model gets sparser until the model is too sparse when $\lambda = 6$. In contrast to the experiment in Figure 1, we see that both the ridge and elastic net models do not very strong advantages in this positivity prediction task. We hypothesize that the reason why Lasso works better in the positivity task is that the frequency of positivity labels is substantially higher than the impoliteness labels in our corpus, so that a Lasso model that enforces full ℓ_1 penalty fits better in this task. In contrast, since the impoliteness label is less frequent, a denser elastic net composite penalty model that preserve critical features, works the best in the impoliteness prediction task. In general, we can see that sparse log-linear models outperform standard log-linear models as well as non-sparse ridge estimators in the two tasks.

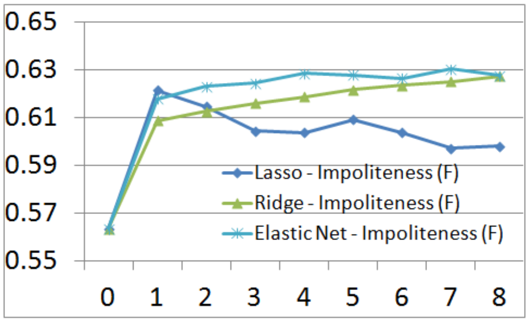

Figure 1: Comparing Impacts of Different Levels of Sparsity on the Friend Dataset When Predicting Impoliteness with Lasso, Ridge, and Elastic Net Models

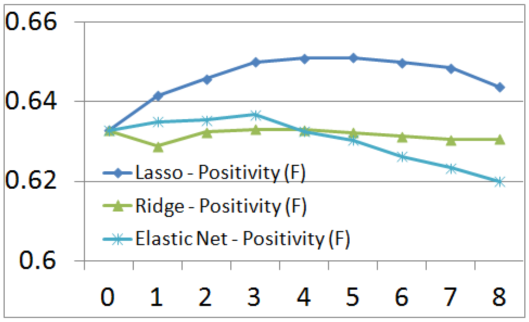

Figure 2: Comparing Impacts of Different Levels of Sparsity on the Friend Dataset When Predicting Positivity with Lasso, Ridge, and Elastic Net Models

5.4 Comparing Impacts of Different Feature Window Sizes

A practical problem for parameter estimation in both generative and discriminative models for dialogue processing is to evaluate how much history the system should take into account, so that it can have enough information to make correct predictions. In this experiment, we investigate the impact of using different feature window sizes using the elastic net model. We compare the two-tailed student t-test between the baseline that only uses features from the current turn and models that use current + previous n turn(s). For the friend dataset, when only using the features from the current turn to predict the impoliteness in the immediate next turn, we observe an AUC of .619. The best result is obtained when we combine the previous two turns together with the current feature turn: an AUC of .635, significantly better ($p = .03$) than only using the current turn window. The patterns on the non-friend dataset are less clear, while the model obtains the best result when window size is +3 previous turns, the improvement is not significant ($p = .962$). In the positivity task, we also observe benefits to incorporating larger feature windows. The AUC on the friend test set starts at .638, when only using the current feature window in the elastic net model. After incorporating larger feature windows, we obtain the best result of .675 at the +4 window ($p = .04$). Similarly, the AUC on non-friend test set initializes at .618, but climbs to .632 at the +4 window.

6 Error Analysis and Discussion

We performed an error analysis to understand the contexts under which our model failed to accurately predict a students' social response, and discuss the implications of these examples based on a theoretical understanding of the roles of tutors and tutees as well as friends and strangers. The following is an example error produced when looking only at the previous turn to predict the current turn:

- Tutee (impolite): "*dude thats def wrong i gotta subract 16m not just 16*" (the current turn)
- Tutor (non-impolite): "*16m is what has to be subtracted from both sides*" (the next turn, predicted incorrectly)

In the segment above the tutee challenges the tutor by pointing out a "def" mistake; the tutor responds with a task-oriented contribution that moves the dialogue forward, but does not escalate the face threat (Ogan et al., 2012). And, in fact, if we look one more turn back in the history, the tutor once again uses calm language: "wait it says youre wrong i dont know why ust wait". The increased window size is implicitly evoking the differential conversational strategies of tutors vs. tutees. And while the current data set is too small to build separate models for tutors and tutees, in this case (and based on the prior work in Ogan et al., 2012), accounting for role distinctions that differentiate strategies taken by tutors and tutees is the likely reason behind the improvement due to window size.

Conversely to the friend data set, the false negatives that occur when predicting impoliteness in the stranger data set are not improved by increasing the window size, as is demonstrated in the following exchange:

- Tutor (non-impolite): “subtract ym from both sides.”
- Tutee (non-impolite): “first step? first Step?”
- Tutor (non-impolite): “*subtract hb from both sides*” (the current turn)
- Tutee (impolite): “*first step? FIRST STEP??????????*” (the next turn, predicted incorrectly)

The impolite tutee utterance at turn 4 is predicted to be non-impolite when analysis is limited to the previous turn, as is also shown in the first example in this section. However, unlike the previous example which improved with an expanding window size, looking back to turns 1 and 2 does not improve the model. While we do not have enough stranger dyads to completely explore this phenomenon, it seems clear that strangers’ responses do not follow the same patterns as friends. The current unpredictability of strangers can be due to a number of social phenomena, such as less affect (both positive and negative) overall, which results in a different conversational flow. Less overall affect means that there is less likely to be useful information in the previous utterances. This is an important distinction between designing models for dyads with rapport and those without, which is a primary concern in the development of social SDSs. Among strangers, other techniques may need to be used to increase model accuracy, such as looking at the content of the utterances to determine whether or not a speaker had been repeating themselves, as is shown in this example, which could likely be an indicator of rudeness.

As a final example of how the error analysis can reveal important phenomena for future study, when examining the prediction of positivity on the stranger test set, we first observe that emoticons are useful indicators of positivity. However, sometimes emoticons serve quite different social functions, which leads to false positives:

- Tutor (non-positivity): “*Simplify ! :)*” (the current turn)
- Tutee (non-positivity): “*y didnt it chang*” (the next turn, predicted incorrectly)

Here, the smiley face is used by the tutor primarily to mitigate the face threat of an impolite command. However, since the experiment reported in Section 6.1 shows that our model attributes more weight to emoticons when predicting positivity, the model errs on this utterance. Here the error analysis suggests that in fact we might need to investigate more complicated latent variable models to capture the subtle social functionality of some language use in context.

7 Conclusion

Long-term relationships involve the expression of both positive and negative sentiments and, paradoxically, both can serve to increase closeness. In this paper, we have addressed the novel task of predicting impoliteness and positivity in teenagers’ peer tutoring conversations, and our results shed light on what kinds of behaviors evoke these social functions for friends and for strangers, and for tutors and tutees. Our investigation has successfully predicted impoliteness and positivity on the basis of both annotated and automatically extracted features, suggesting that a dialogue system will one day be able to employ analyses such as these to signal relationships with users. And while social features such as those we annotated are naturally quite rare in dialogue, our quantitative experiments have demonstrated the capabilities of modeling sparsity in log-linear models: elastic net and Lasso models outperformed standard logistic regression model and the non-sparse ridge penalty model.

We found that positivity is much more predictable for strangers than is impoliteness, while the opposite was true for friends. This could lend support for the importance of positivity as a rapport-signaling function in the early stages of a relationship (as in (Tickle-Degnen and Rosenthal, 1990)), and indicating the need for further research on the increasing importance of impoliteness as a rapport signal over the course of relationship development.

We also found that performance on the prediction tasks increased with larger feature window sizes, particularly for impoliteness among friends and positivity among strangers. From our error analysis, we see that this improvement may arise because different behaviors predict impoliteness and positivity based on the social role of the speaker. Thus tutee bragging predicts positivity in tutors, while tutor bragging negatively predicts positivity among tutees. The power differential between the two may lead tutees to want to take tutors “down a peg” while tutors struggle to maintain the position of power in the dyad.

While results such as these may seem specific to teenage peer tutors, the general conclusion remains, that linguistic devices have different social functions in different contexts, and dialogue systems that intend to spend a lifetime on the job will do well to adapt their language to the stage of relationship with a user, and the social role they play.

References

Hua Ai, Diane J. Litman, Kate Forbes-Riley, Mihai Rotaru, Joel Tetreault, and Amruta Purandare. 2006. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006)*.

Angela M. Ardington. 2006. Playfully negotiated activity in girls talk. *Journal of Pragmatics*, 38(1):73 – 95.

Rachel E. Baker, Alastair J. Gill, and Justine Cassell. 2008. Reactive redundancy and listener comprehension in direction-giving. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*.

Timothy Bickmore, Laura Pfeifer, and Daniel Schulman. 2011. Relational agents improve engagement and learning in science museum visitors. In *Proceedings of the 10th international conference on Intelligent virtual agents*, IVA'11.

Kristy Elizabeth Boyer, Robert Phillips, Michael Wallis, Mladen Vouk, and James Lester. 2008. Balancing cognitive and motivational scaffolding in tutorial dialogue. In *Proceedings of the 9th international conference on Intelligent Tutoring Systems*, ITS '08.

Penelope Brown and Stephen Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*.

Justine Cassell, Alastair J. Gill, and Paul A. Tepper. 2007. Coordination in conversation and rapport. In *Proceedings of the Workshop on Embodied Language Processing*, EmbodiedNLP '07, pages 41–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jan Chovanec. 2009. Simulation of spoken interaction in written online media texts. *Brno Studies in English*.

David Crystal. 2001. Language and the internet. *Cambridge University Press*.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. In *Journal of Pragmatics*.

Jonathan Culpeper. 2011. Impoliteness: Using language to cause offence.

Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech 2006)*.

A Gartner, M Kohler, and F Riessman. 1971. Children teach children: Learning by teaching. In *New York and London: Harper and Row*.

José González-Brenes and Jack Mostow. 2011. Which system differences matter? using l1/l2 regularization to compare dialogue systems. In *Proceedings of the SIGDIAL 2011 Conference*, pages 8–17, Portland, Oregon, June. Association for Computational Linguistics.

Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J. van der Werf, and Louis-Philippe Morency. 2006. Virtual rapport. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA 2006)*.

M. Grimm, E. Mower K. Kroschel, and S. Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. In *Speech Communication*.

P. Gupta and R. Nitendra. 2007. Two-stream emotion recognition for call center monitoring. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*.

Susan C. Herring and Asta Zelenkauskaite. 2009. Symbolic capital in a virtual heterosexual market. In *Written Communication*.

Je Hun Jeon, Rui Xia, and Yang Liu. 2010. Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, INTERSPEECH 2010.

W. Lewis Johnson and Paola Rizzo. 2004. Politeness in tutoring dialogs: run the factory, thats what id do. In *Intelligent Tutoring Systems*, Lecture Notes in Computer Science.

Manfred Kienpointner. 1997. Varieties of rudeness: types and functions of impolite utterances. In *Functions of Language*.

S. le Cessie and J.C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Jackson Liscombe, Julia Hirschberg, and Jennifer J. Venditti. 2005. Detecting certainness in spoken tutorial dialogues. In *Proceedings of the 6th Annual Conference of the International Speech Communication Association (Interspeech 2005)*.

D. Litman and K. Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45:503–528.

Andre Martins, Noah Smith, Mario Figueiredo, and Pedro Aguiar. 2011. Structured sparsity in structured prediction. In *Proceedings of the 2011 Conference on*

Empirical Methods in Natural Language Processing, pages 1500–1511, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Bruce M. McLaren, Sung-Joo Lim, David Yaron, and Ken Koedinger. 2007. Can a polite intelligent tutoring system lead to improved learning outside of the lab? In *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*.

Bruce McLaren, DeLeeuwm Krista E., and Richard E. Mayer. 2011. Polite web-based intelligent tutors: Can they im-prove learning in classrooms? In *Computers and Education*.

Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth S. Narayanan. 2011. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*.

Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012. Rudeness and rapport: Insults and learning gains in peer tutoring. In *Proceedings of the 11 International Conference on Intelligence Tutoring Systems (ITS 2012)*.

J. Alfredo Sánchez, Norma P. Hernández, Julio C. Penagos, and Yulia Ostróvskaya. 2006. Conveying mood and emotion in instant messaging by using a two-dimensional model for affective states. In *Proceedings of VII Brazilian symposium on Human factors in computing systems*, IHC '06, pages 66–72, New York, NY, USA. ACM.

A. Sharpley, J. Irvine, and C. Sharpley. 1983. An examination of the effectiveness of a cross-age tutoring program in mathematics for elementary school children. In *American Educational Research Journal*.

Helen Spencer-Oatey. 2008. Face (im)politeness and rapport. In *Culturally Speaking: Culture, Communication and Politeness Theory*.

Carolyn A. Straehle. 1993. "samuel?" "yes dear?" teasing and conversatrion rapport. In *Framing in Discourse*.

Bas Stronks, Anton Nijholt, Paul van Der Vet, Dirk Heylen, and Aaron Machado. 2002. Designing for friendship: Becoming friends with your eca. In *Proceedings of Embodied conversational agents - let's specify and evaluate (AAMAS)*.

Marina Terkourafi. 2008. Toward a unified theory of politeness, impoliteness, and rudeness. *Impoliteness in language: studies on its interplay with power in theory and practice*.

Robert Tibshirani. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. In *Psychological Inquiry*.

Kevin E. Vowles and Miles Thompson. 2012. The patient-provider relationship in chronic pain. In *Psychiatric Management of Pain*.

Erin Walker, Nikol Rummel, and Kenneth R. Koedinger. 2011. Is it feedback relevance or increased accountability that matters? In *Proceedings of the 10th International Conference on Computer-Supported Collaborative Learning (CSCL 2011)*.

William Yang Wang and Julia Hirschberg. 2011. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proceedings of the 12th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011)*, Portland, OR., USA, June. ACL.

William Yang Wang and Kathleen McKeown. 2010. "got you!": Automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1146–1154, Beijing, China, August. Coling 2010 Organizing Committee.

William Yang Wang, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg. 2012a. Automatic detection of speaker state: Lexical, prosodic, and phonetic approaches to level-of-interest and intoxication classification. *Computer Speech & Language*.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012b. Historical analysis of legal opinions with a sparse mixed-effects latent variable model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*.

Richard J. Watts. 2003. Politeness. *Cambridge University Press*.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.