

Understanding the Perception of Courteous and Humorous Behavior using Prosodic and Lexical Features

Sidd Jagadish*, Ranjay Krishna[‡], Gabriele Carotti-Sha[‡]

siddj@stanford.edu, rak248@stanford.edu, gcarotti@stanford.edu

*Department of Statistics, Stanford University

[‡]Department of Computer Science, Stanford University

[‡] Symbolic Systems, Stanford University

Abstract

Computationally understanding the perception of a speaker’s characteristics is an important task speech processing, behavioral outcomes and dialogue systems. Our work focuses on the classification and understanding the perception of two such characteristics: courtesy and humor. We analyze ratings surveyed from the SpeedDate corpus to build models using prosodic and lexical features of the speakers. We find that lexical features like *Hedge-words* and prosodic features like *Intensity Min SD* are important for detecting courteous while *Exclamations* and listener’s *Intensity* are important for funny. Using an AdaBoost Classifier, we detect the funny with an accuracy of SHIT GOES HERE and courteous at SHIT GOES HERE.

Index Terms: Prosody, Speed Date, AdaBoost, Gender differences in prosody, factor analysis, courteous, funny

1. Introduction

Efforts in linguistic understanding encounter inevitable difficulties whether the interpretative entity is a human or a machine. For example, any communicative exchange across agents runs the risk of misrepresentation of one’s self or of their interlocutor. It has been shown that given a dialogue between just two agents, human judgments about each other can be subject to great discrepancies (R. Ranganath et al. 2009), and this is potentially due to a failure in recognizing each other’s affective cues. To face this issue, it is useful to invoke the notion of *interpersonal stances*, as described by Schaefer (2000, 2003). An interpersonal stance is the way in which interlocutors pose themselves with respect to other agents within a given exchange. In particular, this notion incorporates affective stances, such as flirtatiousness, awkwardness, or courteousness, whose expression can be detected via subtle cues from a speaker’s voice or words. This observation leads to the definition of two sets of communicative features: prosodic, pertaining to the physical characteristics of the vocal signal, and lexical, pertaining to the semantic content of the words pronounced.

Examples of prosodic features are the fundamental frequency contour, or pitch, of a given speaker’s voice, or the mean intensity they use to convey a particular emotive state (anger, excitement, humor). On the other hand, the number of times a given word is repeated, either throughout an entire conversation or only in response to a preceding speech segment, are considered lexical features. Thanks to the detailed and insightful research conducted by R. Ranganath et al. (2013), it has been possible to extract both sets of features from a series of speed

date encounters, which were recorded and transcribed. Despite the relative artificiality of conversations within this particular context, this very artificiality can be favorable in an empirical study: it poses constraints on how much time interlocutors have to establish an exchange, and also generates very specific interpersonal stances (listed in the following sections), all of which are relevant to the date being either a success or a failure. By applying NLP classification methods to the extracted feature sets, it is possible to generate a fairly accurate predictor of whether, given a pair of interlocutors, one of the two found the other to be flirtatious, intelligent, etc. and vice versa.

We build our own models upon the previously conducted research on the same dataset, and apply the same techniques in order to detect stances such as courteousness, sincerity, and humor.

2. Related Works

3. The SpeedDate Dataset

We utilize the SpeedDate dataset introduced by [1] to build our model. The dataset contains approximately 1100 heterosexual 4-minute speed dates. Each date is stored as a wav file recording of the speed date along with text files of the dates annotations. On average, each date contains 812 words, with an average of 406 words per speaker. These words are divided into an average of 93 turns per date where the speaker changes to the other.

3.1. Feature Extraction

We use OpenSMILE to extract Prosodic features from audio files of the speed date. Lexical features are extracted from the transcripts with the help of the LIWC dictionary.

3.1.1. Prosodic Feature

Prosodic features describe the speech of the speaker. In our model, we include features such as *F0*, *intensity*, *RMS Amplitude*, *turn duration* and *pitch*. For each of these different features, we calculate the *min*, *max*, *sd*, *range*, *sd sd* and *mean*.

3.1.2. Lexical Feature

Using LIWC, introduced by [2], we extract *hedge words* like *sort of* and *I guess* that indicate uncertainty. We also gather *meta* words that represent common words in the given speed date scenario and *academic* words like *PhD* and *research*. Finally, we record occurrences of discourse markers.

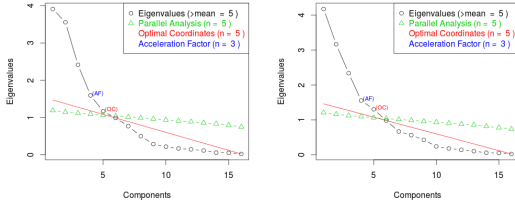


Figure 1: *Determining Number of Factors*

3.1.3. LIWC Feature

We use the LIWC software to classify words into specific topic groups like *love*, *hate*, *food*, *negate* and *drink*. We count the word occurrences within each speed date for each of these word topics.

3.1.4. Accommodation Feature

[3], [4] and [5] demonstrate through their work the convergence of vocal intensities between interlocutors. To capture this convergence, we extract features from a speaker that accommodate the previous speaker's speech. These features include the *Rate of Speech* of the two speakers over time, the number of *functional* and *content* words also used in the other's previous turn and *laughter* that immediately proceeds the other's laugh.

3.2. Feature Normalization

Before we use the features in our model, we normalize the relevant lexical feature with the the total word count in the conversation. We then normalize all the features such that they all have zero mean and unit variance.

4. Exploratory Analysis

Before attempting to classify speakers as funny or not, we examine the underlying dimensions along which the data varies. To do, so we use exploratory factor analysis. To determine the number of factors, we use a scree test and various non-graphical measures, including parallel analysis and an optimal coordinates test, as described in (CITE). We find an optimal number of factors $k = 5$ for both males and females, conducted separately. Figure 1 shows the two scree plots.

Interestingly, although we find that the various factors for male and female speech are similar, they explain different amounts of variation in the data. For males, the first factor reflects high intensity values and low intensity variation, explaining

5. Evaluation Approach

5.1. Methodology

We divide the dataset into two separate groups based on gender. For each gender, we train a classifier for detecting the perception of funny and courteous characteristics.

We create a training set for a given characteristic for a given gender by splitting the dataset into deciles. We use the top decile as positive and the bottom decile as negative examples.

5.2. Classifiers

We start by using an SVM with a linear kernel. To improve our classifications, we next enforce an L1 norm on the SVM to compensate for the large number of lexical and prosodic features we extract. In case our decision boundaries are non-linear, we also train a RBF kernelized SVM. However, we lose interpretability of the features by using an RBF. So, we train a fourth Adaboost Classifier with decision stumps of unary depth to maintain interpretability while also allowing for a non-linear decision boundary.

6. Classification Results

7. Analysis

7.1. Adaboost Weights

By limiting the depth of the trees in Adaboost to 1, we are able to extract the decision boundaries

8. Future Work

The model we use has a lot of features and reducing dimensionality of this data by learning hidden variables from an sparse auto-encoder might give us more insight into how we can create a shared representation of the lexical with the prosodic features.

The SpeedDate dataset we use also contains ratings of people's intelligence, sincerity and other characteristics. An interesting area to explore would find correlations between these characteristics and their corresponding prosodic and lexical features.

Adding another source of features, such as video from video speed dates is another extension to this project. It would be insightful to see how the contribution of the current features will change if visual features are also added.

9. Conclusions

10. Acknowledgements

We thank Dan Jurafsky for giving us access to the Speed Dating Dataset and for his continued guidance during this project.

11. References

- [1] Ranganatha, R., Jurafsky, D., McFarland, D.A., "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates", in *Computer Speech & Language* Volume 27, Issue 1, Pages 89-115, 2013
- [2] Pennebaker, J. W., Booth, R., Francis, M., "Linguistic inquiry and word count: LIWC2007" operators manual. Tech. rep., University of Texas, 2007
- [3] Natale, M., "Convergence of mean vocal intensity in dyadic communication as a function of social desirability", in *Journal of Personality and Social Psychology* 32 (5), 790-804, 1975
- [4] Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., Pennebaker, J.W., "Language style matching predicts relationship initiation and stability", in *Psychological Science* 22 (1), 39, 2011
- [5] Levitan, R., Hirschberg, J., "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions", in *ACL* 2011