Look Who's Talking: Inferring Speaker Attributes from Personal Longitudinal Dialog

Charles Welch, Verónica Pérez-Rosas,
Jonathan K. Kummerfeld, and Rada Mihalcea

University of Michigan
{cfwelch,vrncapr,jkummerf,mihalcea}@umich.edu

Abstract. We examine a large dialog corpus obtained from the conversation history of a single individual with 104 conversation partners. The corpus consists of half a million instant messages, across several messaging platforms. We focus our analyses on seven speaker attributes, each of which partitions the set of speakers, namely: gender; relative age; family member; romantic partner; classmate; co-worker; and native to the same country. In addition to the content of the messages, we examine conversational aspects such as the time messages are sent, messaging frequency, psycholinguistic word categories, linguistic mirroring, and graph-based features reflecting how people in the corpus mention each other. We present two sets of experiments predicting each attribute using (1) short context windows; and (2) a larger set of messages. We find that using all features leads to gains of 9-14% over using message text only.

Keywords: longitudinal dialog analysis, natural language processing

1 Introduction

People spend a significant amount of time using social media services such as instant messaging to communicate and keep in touch with others. Over time, conversation history can grow quickly, thus becoming an abundant source of personal data that provides the opportunity to study an individual's communication patterns and social preferences. Analyzing conversations from a single individual rather than conversations from multiple individuals can enable identification of social behaviors that are specific to that individual. Moreover, longitudinal analyses can help us better understand an individual's social interactions and how they develop over time.

In this work we look at a collection of personal conversations of one of this paper authors' over a five-year span, consisting of nearly half a million messages shared with 104 conversation partners. To address data privacy issues, during the experiments and analyses presented in this paper, the actual message content is only accessible to its owner. We focus our analyses on seven speaker attributes: a ternary attribute for relative age (younger, older, or same age); and six binary attributes reflecting whether somebody is the same gender; a family member; a romantic partner; a classmate; a co-worker; and a native of the same country.

We explore the classification of speaker attributes, i.e, the group(s) the speaker belongs to, using a variety of linguistic features, message and time frequency features, stylistic and psycholinguistic features, and graph-based features. In addition, we examine the performance increase gained by using six of the attributes as features to try to classify the seventh.

We analyze linguistic variation in messages exchanged between the author and the other speakers. We also conduct analyses that look at speaker interaction behaviors, considering aspects such as time, messaging frequency, turn-taking, and linguistic mirroring. Next, we apply graph-based methods to model how people interact with each other by representing people as nodes and speaker mentioning each other as directed edges. We then apply clustering methods to identify groups that naturally occur in the graph. Finally, we conduct several classification experiments to quantify the impact of features derived from these analyses on our ability to determine who a speaker is.

Identifying speaker attributes has important applications within the areas of personalization and recommendation [14, 4]. While a large number of conversations that occur online are short, such as interactions on Twitter, there are also many social media platforms where personal dialog may span thousands of utterances. For this reason, we conduct evaluations at the level of small context windows, as well as at the speaker level using a large set of messages from each speaker. To the best of our knowledge, this is the first study on speaker attribute prediction using personal longitudinal dialog data that focuses on one person's dialog interactions with many other speakers.

2 Related Work

Our work is related to three main directions of research: authorship attribution, discourse analysis, and speaker attribute classification from social media.

On authorship attribution, there have been several studies focusing on inferring author's characteristics from their writing, including their gender, age, educational, cultural background, and native language [6, 10]. This work has considered linguistic features to capture lexical, syntactic, structural, and style differences between individuals [10]. A recent study in this area analyzed language use in social media to identify aspects such as gender, age, and personality by looking at group differences on language usage in words, phrases, and topics discussed by Facebook users [15].

Discourse analysis approaches have been used to examine language to reveal social behavior patterns. Holmer [7] applied discourse structure analysis to chat communication to identify and visualize message content and interaction structures. He focused on visualizing aspects such as conversation complexity, overlapping turns, distance between messages, turn changes, patterns in message production and references. In addition, he also proposed graph-based methods for showing coherence and thread patterns during the messaging interaction. Tuulos [17] inferred social structures in chat-room conversations, using heuristics based on participants' references, message response time and dialog sequences and represented social structure using graph-based methods. Similarly, Jing [9] looked at extracting networks of biographical facts from speech transcripts that characterize the relationships between people and organizations.

Work in classifying user attributes has used both message content and other meta-features. Rao [14] looked at classifying gender, age (older or younger than 30), political leaning, and region of origin (north or south India) as binary variables using a few hundred or a few thousand tweets from each user. They used the number of followers and following users as network information to look at frequency of tweets, replies, and retweets as communication-based features but found no differences between classes. Hutto [8] analyzed sentiment, topic focus, and network structure in tweeting behavior to understand aspects such as social behavior, message content and following behavior. Other work has derived useful information from Twitter profiles, such as Bergsma [2] who focused on gender classification using features derived from usernames, and Argamon [1] who found differences in part of speech and style when examining gender in the British National Corpus.

3 Conversation Dataset

We use a corpus of text messages from one author's personal conversations on Google Hangouts, Facebook Messenger, and SMS text messages. The message set contains nearly half a million messages from conversations held between the author and 104 individuals. Aggregate statistics describing the corpus are shown in Table 1.

Table 1. Distribution of messages and tokens (words, punctuation, emoticons) in the conversations between the author and other individuals.

	Author	Others	All
Total Messages	237,300	216,766	454,066
Unique Messages	165,536	168,041	326,243
Total Tokens	1,370,916	1,602,607	2,973,523
Unique Tokens	38,937	48,005	68,985
Average Tokens / Message	5.78	7.39	6.55

We use seven attributes that describe the relationship between the author and their conversation partner. Table 2 shows the distribution of people and messages for each attribute in the dataset. They were annotated by the author and interpreted as follows:

Family: This person is related to the author.

Romantic Relationship (Rom. Rel.): This person's relationship with the author was at some point not platonic.

Table 2. Distribution of speakers and messages in the corpus by speaker attributes (% of corpus). The values for *Age* represent 'younger', 'older', and 'same age', while the values for the other attributes represent 'yes' and 'no'.

	Family	Rom. Rel.	Rel. Age	Child. Co.	Gender	School	Work
	Y/N	Y/N	Y/O/S	Y/N	Y/N	Y/N	Y/N
%Speakers	6/94	9/91	26/30/44	78/20	51/49	62/38	33/67
%Messages	8/92	22/78	24/24/52	88/11	53/47	75/25	54/46

Relative Age (Rel. Age): This person the same age (±1.5 years), is older, or is younger than the author.

Childhood Country (Child. Co.): This person grew up in the same country as the author.

Gender: This person has the same gender as the author.

School: This person and the author met attending school.

Work: This person and the author know each other because they worked together.

4 Message Content

We start by exploring linguistic differences in the messages exchanged between the author and each of the groups defined by the seven attributes described above. We obtain the most dominant semantic word classes [13] in messages exchanged with people sharing each attribute using the LIWC [16] lexicon, which contains psycholinguistic categories of words. The top ten dominant classes for each attribute-value pair are shown in Table 3.

Not surprisingly, the 'Family=Yes' group talks more about family and home than the 'Family=No' group. Interestingly, people who are not family members seem to use more emotion related words. Word categories related to feelings are also very dominant for the 'Romantic Relationship=Yes', 'Relative Age=Same', 'Childhood Country=Same' and 'Gender=No' groups; however they seem to focus on negative emotions such as anxiety and sadness. In fact, those two are in the top three classes for conversations with romantic partners 'Romantic Relationship=Yes', which also includes death words (words related to death are often used in hyperbole, e.g. "I didn't eat lunch and I'm dying"). This suggests that more serious conversations occur between the author and this group as compared to the 'Romantic Relationship=No' group.

Several of the attributes clearly separate the set of speakers into those who speak about work and those who do not. People who talk the most about work are those who grew up in other countries ('Childhood Country=Other'), people from work ('Work=Yes'), people older than the author ('Relative Age=Older'), people with the same gender ('Gender=Yes') and people from school ('School=Yes').

Table 3. Dominant LIWC word classes for each attribute/value pair. The top ten classes are listed for each attribute in decreasing order.

Attribute	Top Classes
Family	**Yes:** Family, Money, Home, Swear, Death, Leisure, Filler, Anger, Female, Health
	No: Anxious, Insight, Feel, Risk, Sad, Positive Emotion, Non-fluencies, Causality, Affect, Work
Romantic Relationship	**Yes:** Anxious, Death, Sad, Feel, Body, Filler, You, Family, Perception, Health
	No: Swear, Female, Money, Friend, Anger, She-He, Work, Leisure, Informal, Male
Relative Age	**Younger:** Netspeak, Ingest, Swear, Friend, Biological, Home, Anger, Informal, Body, Leisure
	Same: Female, Swear, Anger, She-He, Anxious, Negative Emotion, Friend, Sad, Negate, Money
	Older: See, We, Work, Number, Article, Home, Perception, Space, Motion, Relativity
Childhood Country	**Same:** Death, Family, Anger, Swear, Feel, Female, Negative Emotion, Body, Anxious, Health
	Other: We, Work, You, Male, Focus Future, Social, Affiliation, Friend, Assent, Time
Gender	**Yes:** Money, Female, Swear, Work, Friend, Netspeak, She-He, Article, Power
	No: Sad, Anxious, Family, Health, Death, Body, Biological, Negative Emotion, Ingest, Home
School	**Yes:** Work, Non-fluencies, Insight, Risk, Anxious, Quantify, Focus Past, Causality, Tentative, Compare
	No: Family, Money, Health, Home, Netspeak, Death, Swear, Leisure, Biological, Anger
Work	**Yes:** Work, Article, Number, We, Non-fluencies, Quantify, Compare, Insight, Achievement, Assent
	No: Family, Health, Money, Death, Anger, Swear, Anxious, Home, Biological, Sad

However, there are some differences between these groups which can be seen mostly in the family, health, time, and gender specific words they use.

People from school use more words referring to the past, while people from other countries focus more on the future. Interestingly, people not from work ('Work=No') and the people not from school ('School=No') are very similar, and both use a lot of family, health, and money words. The similarity of these two attributes is also interesting in that people from work ('Work=Yes') and/or school ('School=Yes') use more quantifying words (e.g. sampling, percent, average) and disfluencies (e.g. umm, hmm, sigh). We also see that those who grew up in other countries use more male words, while speakers that are the same age, from the same country, or of the same gender use more female words.

5 Groups Over Time

To understand the role that time has in the author's interactions with different groups we look at patterns in message volume over different intervals. Most notably, we find interaction differences given the day of the week, and the hour of the day. In Figure 1 we plot the attribute/value pairs that differ the most from the trend over all people, marked 'All'. The difference was calculated as the sum of differences on each of the seven days of the week and each of the 24 hours of the day.

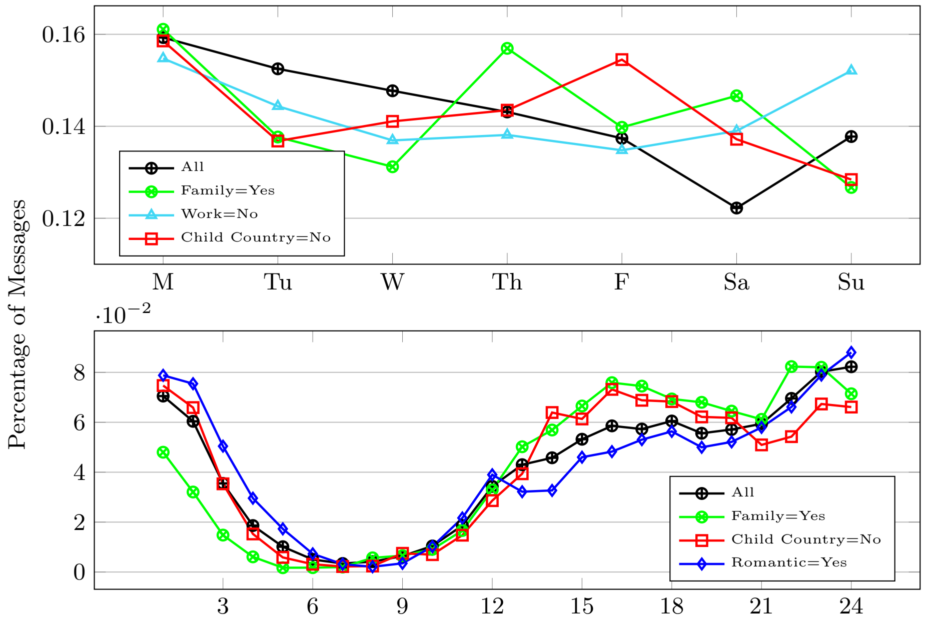

Fig. 1. Distribution of messages over time. The top shows the distribution over the day of the week and the bottom shows hour of the day. The groups shown are those that vary the most from the aggregate trend over all speakers.

We see that the overall trend for the day-of-week plot (top) is that there are more conversations during the first days of the week. The number of conversations drops until Sunday where it jumps back up and peaks on Monday. Throughout the week, most of the conversations occur between family members and people that grew up in other countries (co-workers mainly). In contrast, there are many more conversations with people outside of work on the weekend.

The hour-of-day plot (bottom) indicates that most of the interactions happen between 9AM and 6PM. Though this is a trend aggregated over all days in the corpus it shows that the author is least likely to be talking to people in the 7-8AM range. The author tends to speak more to people later in the day, with a peak at midnight. People who grew up in other countries converse more with the author during the day. The dominant 'Work' category for 'Childhood Country=Other' in Table 3 shows this trend, as this group may converse with the author more about work during work hours. We also find that family members speak to the author more during the day and romantic partners speak to the author more after midnight but before noon.

6 Conversation Interaction

Linguistic mirroring is a behavior in which one person subconsciously imitates the linguistic patterns of their conversation partner. Increased linguistic mirroring can be an indicator of an individual building rapport with others and thus forming better interpersonal relationships. We study linguistic mirroring in our dataset to analyze how relationships change over time. We calculate linguistic style matching (LSM) as the similarity of the normalized counts of nine types of function words [5], as the main metric for our analyses. In Figure 2 we show style matching over the first 5,000 messages with people in five specific groups. We see that although the general trend is to match language style more over time, this trend levels off after 3k messages, potentially because at this point relationships start to consolidate.

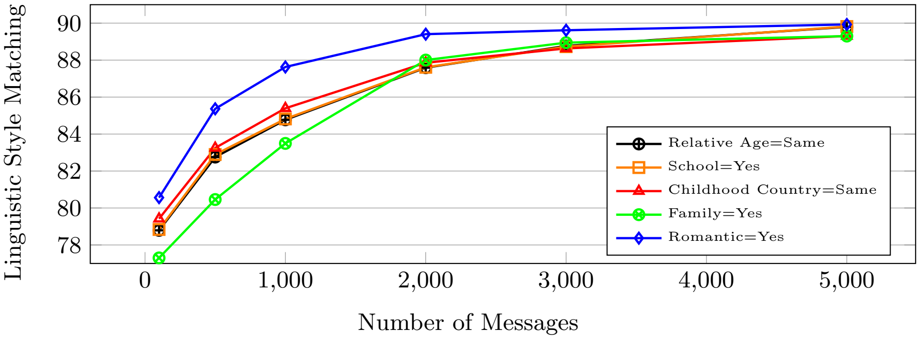

Fig. 2. Language mirroring as a function of the number of messages exchanged within groups. Mirroring is shown over the first 5,000 messages averaged over people in each of the listed groups.

Next, we examine interactions between groups of people by constructing a graph where nodes represent speakers and edges between nodes represent speakers mentioning each other. Speakers who mention each other also tend to know each other. They might mention another person when planning to meet up with others or when talking about an interaction they had with this person in the past. We clustered the graph of people using Louvain clustering [3] to maximize

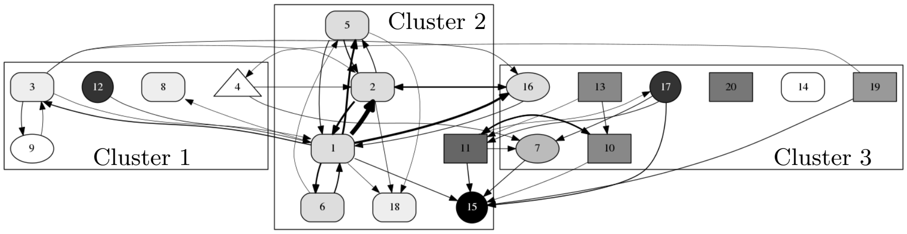

Fig. 3. Speaker references for the top 20 conversation partners. The graph shows interactions with people from different groups: high school (rectangles), college (triangles), graduate school (rounded rectangles), family members (circles), and other people (ellipses). Shading is proportional to how long ago the author met the person. Edges below a threshold of 25 mentions are removed. Note that the clustering uses all 104 people, but only 20 are shown here.

the modularity of the network. This gave four clusters, one of which only contained two people. The remaining clusters roughly evenly split the set of people. The top twenty most frequent conversation partners are shown in Figure 3. Interestingly, the clusters resemble groups of speakers that the author spoke most to at three periods of time contained in the corpus i.e, conversations before attending graduate school (Cluster 3), the beginning of graduate school (Cluster 2), and later in graduate school (Cluster 1). We also see that people who spoke to the author more at a particular time were also more likely to know each other.

Table 4. Two examples of five-message context windows (ctx_1 and ctx_2) taken from the data.

Message Number	Time	Message
$ctx_1 msg_0$	15:45:06	Participant: Wanna grab coffee?
$ctx_1 msg_1$	15:45:20	Author: yeah
$ctx_1 msg_2$	15:45:25	Participant: Sweet!!!!
$ctx_1 msg_3$	15:45:29	Participant: Meet in the lobby?
$ctx_1 msg_4$	15:45:52	Author: okay
$ctx_2 msg_0$	12:21:00	Participant: Perfect!!
$ctx_2 msg_1$	15:56:22	Participant: Wanna go to get Thai?
$ctx_2 msg_2$	16:01:18	Participant: I'll take it you're sleeping lol
$ctx_2 msg_3$	16:19:59	Author: Yeah
$ctx_2 msg_4$	16:20:08	Author: I mean yeah I was sleeping

7 Model

Using the messages in a conversation between two speakers, we wish to be able to identify the value of each of the speaker attributes of whom the author is conversing with. In order to do this, we can encode part of the conversation and additional features and output the value of an attribute. In text messaging, it is often not clear what a conversation is about by just examining individual messages. Thus, we decide to conduct our analysis on small sequences of message exchanges between speakers. Throughout the rest of the paper, we will refer to each of these sequences as a *context window*, which consists of five messages exchanged between the author and another speaker[1]. Two sample context windows are shown in Table 4.

During our experiments, we use a bidirectional long-short term memory network (BiLSTM) as our baseline model. The input for this model is a dialog context window, in which all utterances are concatenated but one token is used to represent the beginning of an author utterance and another token is used to represent the beginning of any other speaker's utterance. We use the same implementation to incorporate additional features.

The model architecture is shown in Figure 4. As shown, the context encoder takes the concatenated window of length n and generates the encoding ρ_1. In the baseline case the feature encoders are not used and the context encoding is passed directly to an attribute decoder. A separate attribute decoder is used for each speaker attribute and has k outputs, where k is two for every case except 'relative age', which has three possible values.

When using additional features, we take the BiLSTM output, representing the encoded context window, and append it to a normalized vector representation of each additional feature set, ρ_i. A feed-forward layer is then used to encode each feature set separately. The hidden size s for both the feature encoders and attribute decoders were manually tuned in preliminary experiments.

We use a hidden size of 64 for experiments in this paper. In our models that use one or more feature encoders, the concatenated ρ vector is used for decoding. The feature encoder sizes t will vary depending on which feature set is being encoded. The word embedding inputs to the context encoder are 300 dimensional.

8 Features

Word Embeddings: We obtain word vector representations for each message using the GloVe Common Crawl pre-trained model [12]. We chose GloVe over other frequently used embeddings because its training data is more similar to our data and we observed a higher token coverage rate than embeddings such as word2vec trained on GoogleNews [11].

[1] Context window size is fixed in our experiments but future work could explore prediction accuracy as a function of this variable.

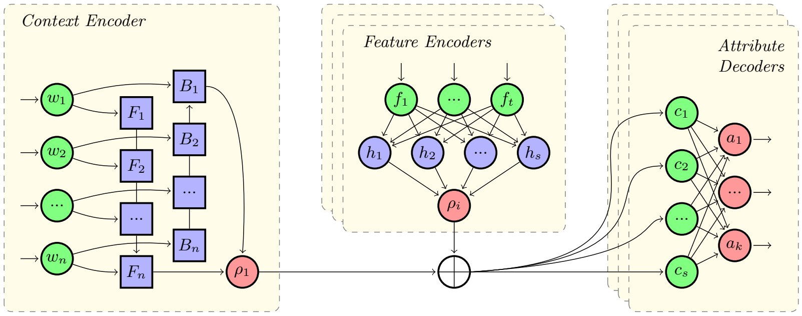

Fig. 4. The model architecture encodes a context window as a sequence of tokens w_1 to w_n using a BiLSTM which is represented with forward and backward cells. The encoding is then used in combination with our other feature sets for decoding. A separate decoder is used for each speaker attribute.

LIWC: To calculate these features, we obtain the normalized counts of 73 LIWC categories. The feature set includes the vectors obtained from messages of individual conversation participants, the cosine similarity between them, and the vector sum of both speakers.

Time: These features include the time elapsed during the context window, the number of seconds between each of the messages, and the day, month, year, season (winter, fall, summer, spring), and hour of the day of the last message.

Messaging frequency: This set of features includes the number of messages exchanged between conversation participants in the past day, week, month, and from all time. The vector also includes a list of binary values representing the turn change sequence in the context window.

Style Matching: Looks at the similarity of the ratios of function word usage between the two speakers. This set of features includes the LSM score for the last hundred messages exchanged by the conversation participants, as well as the change in style matching over the context window by subtracting the final and initial LSM scores.

Graph-based: Uses the training set of messages to generate a graph where nodes represent people and weighted, directed edges represent how often that person mentions another person when speaking to the author. This graph is used to generate features by finding the shortest path between users where edge weights are smaller when they have more mentions. We then use the adjacency matrix to find the shortest paths between nodes and use each row as a feature set, representing a speaker i conversing with this person. Given a graph of mentions, where $M_{i,j}$ represents how often person i mentions person j, we compute weights using the following equation:

$$W_{i,j} = 1 - \frac{w_{max} - M_{i,j}}{w_{max} - w_{min}}$$

Speaker Attributes: When we are predicting one of the seven speaker attributes this feature set represents the values of the other six attributes. Note that we cannot use this feature when training joint models.

9 Experiments

Using the features described in Section 8 we run experiments using leave-one-speaker-out cross validation. We take the 104 speakers in our dataset and hold out all context windows containing dialog with one of the speakers as a test set and use the rest for training and validation with a 90% and 10% split. This means that we train and tune parameters on context windows from all 103 other speakers and update the model based on its predictions on each individual context window. During test time we examine the context-level and speaker-level accuracy. Context-level accuracy is calculated by macro-averaging the context window accuracy over all speakers. To calculate accuracy at speaker level, we first obtain the attribute prediction at context-window level for the held-out speaker and assign the attribute value most frequently predicted by the classifier.

We run experiments using a baseline model which only uses word embeddings and compare it to a model that uses all of our features. Additionally, we perform an ablation to examine the effectiveness of each feature set for predicting each speaker attribute by running the model using the word embeddings plus one of the other feature sets at a time. While we vary the number of feature encoders we use (see Figure 4), each model always uses one attribute decoder. The loss for each model is calculated as the cross-entropy loss for that model's attribute decoder.

Since this evaluation is computationally expensive we run our experiments on a subset of the original corpus. Thus, we obtain a sample of 27,316 context windows, distributed as evenly as possible, from each speaker in the dataset to ensure that all people and attributes are represented. Experiments using this dataset took 3-4 days to run on a cluster with 12 NVIDIA GeForce GTX TITAN X GPUs.

During our experiments we consider single attribute models, which use only one attribute decoder, and joint models, which learn to predict all attributes at the same time using all decoders. In the single attribute setting we train a separate model for each attribute and calculate the cross-entropy loss for the decoder, while in the joint case we take the sum of the losses for all decoders.

10 Results

The results obtained for each attribute, when using different combinations of features are shown in Table 5 and Table 6. The first table shows accuracies at the person-level while the latter shows performance macro-averaged over context-windows. Overall, the combination of all features improves the prediction performance for all the attributes over a baseline model that only uses word embeddings, with the exception of the gender attribute. The largest context-window

Table 5. Results are shown for the accuracy per person using leave-one-speaker-out cross validation. Individual models learn to classify each attribute in all cases except for the two 'Joint' rows, which jointly classify attributes. Feature ablations are shown for each of the single feature types, and compared to the model that uses all features, as well as the baselines obtained using the majority class or message embeddings (Emb) only. Additional improvements are shown when training single attribute classifiers and using the other six attributes as features.

		Rom.	Rel.	Child.			
	Family	Rel.	Age	Co.	Gender	School	Work
Baselines							
Majority Class	94.2	91.3	44.2	77.9	51.0	61.5	67.3
Emb	**94.2**	**91.3**	45.2	79.8	**86.5**	73.1	80.8
Single Attribute Decoder Ablation							
Emb + Time	94.2	91.3	44.2	79.8	85.6	76.0	85.6
Emb + LIWC	94.2	91.3	46.2	80.8	82.7	73.1	84.6
Emb + Style	94.2	91.3	49.0	78.8	86.5	76.0	85.6
Emb + Frequency	94.2	91.3	44.2	80.8	83.7	75.0	86.5
Emb + Graph	93.3	91.3	43.3	77.9	80.8	76.0	**87.5**
Single Attribute Decoder All Features vs Joint Decoder Models							
All Features	92.3	91.3	45.2	81.7	76.0	**76.9**	83.7
Joint + Emb	94.2	91.3	48.1	78.8	85.6	71.2	83.7
Joint + All	92.3	91.3	**51.9**	**84.6**	77.9	75.0	84.6
Single Attribute Decoder with Attribute Features							
Emb + Attributes	94.2	91.3	48.1	87.5	83.7	73.1	84.6
All + Attributes	93.3	91.3	50.0	***88.5***	78.8	***78.8***	85.6

level improvements are obtained for the *Relative age*, *Childhood country*, *Gender* and *Work* attributes. The largest speaker-level improvements are similar with the addition of *School* and without *Gender*.

Although in some cases the accuracy of attribute prediction at speaker-level is not improved by the different set of features, we still observe an improvement on the prediction accuracy at the context window level. For instance, the *Family* and *Romantic* attributes improve by 2.1% and 6% respectively. We also see that the *Gender* attribute improves up to 6.8% by this metric.

Using the other six speaker attributes as features to classify the seventh proved to be beneficial in all cases. The graph features also proved useful for all attributes showing gains of up to 6.7% in speaker-level performance and up to 7% in context-window level performance. The frequency features gave the biggest performance increase to the *Romantic*, *Childhood country*, and *Work* attributes. Time features improve performance most on *Romantic*, *Gender*, *School*, *Work*.

The overall trend we found in Section 6 showed that the most distinct groups when looking at language mirroring were 'Family=Yes' and 'Romantic=Yes'.

Table 6. Accuracy on context windows macro-averaged over speakers. The individual, joint, single attribute, and baseline models are defined the same way as in Table 5.

		Rom.	Rel.	Child.			
	Family	Rel.	Age	Co.	Gender	School	Work
Baselines							
Majority Class	94.2	91.3	44.2	77.9	51.0	61.5	67.3
Emb	92.0	86.0	39.2	75.7	63.7	64.6	69.5
Single Attribute Decoder Ablation							
Emb + Time	91.7	86.8	40.5	77.4	63.4	64.4	73.1
Emb + LIWC	91.9	86.4	39.6	76.7	62.6	63.8	69.4
Emb + Style	92.0	86.0	38.9	76.2	62.8	65.1	69.2
Emb + Frequency	91.3	87.9	39.2	76.0	62.4	65.5	71.3
Emb + Graph	92.1	86.2	41.7	76.9	61.4	67.2	73.3
Single Attribute Decoder All Features vs Joint Decoder Models							
All Features	92.0	88.1	42.7	78.9	61.2	67.0	76.0
Joint + Emb	**93.9**	**90.9**	43.4	78.0	**64.2**	65.5	69.3
Joint + All	92.1	90.2	**47.2**	**80.8**	61.8	**68.7**	**78.4**
Single Attribute Decoder with Attribute Features							
Emb + Attributes	92.6	86.4	41.5	84.1	***68.6***	72.7	78.4
All + Attributes	92.0	88.2	44.3	***85.7***	67.1	***74.3***	***83.4***

However, we found that the language mirroring features that we used, which use a sliding window, were most useful for *Relative age*, *School*, and *Work*. Similarly, LIWC features help for *Relative age* and *Work*, but they also improve prediction performance for *Childhood country* and *Gender*.

At the speaker level, classification is more difficult and we do not see improvement for all attributes when using the additional features or joint decoders. However, at the context-window level we found that joint decoders improved over single attribute decoders in all cases, though using the additional features did not help for *Romantic*, *Family*, and *Gender*. When using single attribute decoding with the other attributes as features we found even higher performance for four of the attributes. Interestingly, *Gender* still does not benefit from using extra features and simply knowing the values of the other speaker attributes gives the best result. The lowest accuracy overall is obtained for relative age, this can be partly explained by the lower baseline as compared to the other attributes, which is influenced by the fact that it has three possible values instead of two.

11 Conclusion

In this paper, we addressed the task of classifying the attributes of an individual based on their conversations in a longitudinal dataset. We conducted analyses of several interaction aspects, including message content, speaker groups over time, and interaction during the conversation. We developed a bidirectional LSTM architecture that, in addition to message content, includes a variety of features derived from our analyses, covering the time-stamp of the messages, messaging frequency, psycholinguistic word categories, linguistic mirroring, and graph-based representations of interactions between people. Additionally, to account for scenarios where some attributes are known, we present experiments that evaluate the use of the other six speaker attributes when classifying the seventh.

Our experiments evaluate the accuracy of predictions at the context-window level, which uses only a sequence of five messages for message content, as well as at the speaker level using a larger set of context windows from each speaker. We observed improvements in speaker level accuracy up to 8.7% and up to 13.9% accuracy on context windows. We explore the usefulness of each feature with an ablative study and compare two different methods of decoding. For the case of predicting someone's relative age or whether or not they are a co-worker, classmate, or native from the same country, we see improvement at both levels. Our evaluations show improvement over a system that only uses one of these features at a time, as well as over a baseline system that relies exclusively on message content.

To the best of our knowledge, this is the first study on speaker attribute prediction using personal longitudinal dialog data that focuses on one persons' interactions with many users. The code used to extract the conversations from social media, to interactively annotate speakers, and to perform the experiments presented in this paper is publicly available[2], so others can conduct analyses on their own data.

Acknowledgments

This material is based in part upon work supported by the Michigan Institute for Data Science, by the National Science Foundation (grant #1815291), by the John Templeton Foundation (grant #61156), and by DARPA (grant #HR001117S0026-AIDA-FP-045). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Michigan Institute for Data Science, the National Science Foundation, the John Templeton Foundation, or DARPA.

References

1. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, genre, and writing style in formal written texts. Text-Interdisciplinary Journal for the Study of Discourse **23**(3), 321–346 (2003)
2. Bergsma, S., Dredze, M., Van Durme, B., Wilson, T., Yarowsky, D.: Broadly improving user classification via communication-based name and location clustering

[2] `https://github.com/cfwelch/longitudinal_dialog`

on twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1010–1019 (2013)

3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
4. Garera, N., Yarowsky, D.: Modeling latent biographic attributes in conversational genres. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 710–718. Association for Computational Linguistics (2009)
5. Gonzales, A.L., Hancock, J.T., Pennebaker, J.W.: Language style matching as a predictor of social dynamics in small groups. Communication Research (2009)
6. Hirst, G., Feiguina, O.: Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing **22**(4), 405–417 (2007)
7. Holmer, T.: Discourse structure analysis of chat communication. Language@ Internet **5**(10) (2008)
8. Hutto, C.J., Yardi, S., Gilbert, E.: A longitudinal study of follow predictors on twitter. In: Proceedings of the SIGCHI conference on human factors in computing systems. pp. 821–830. ACM (2013)
9. Jing, H., Kambhatla, N., Roukos, S.: Extracting social networks and biographical facts from conversational speech transcripts. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 1040–1047 (2007)
10. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the Association for Information Science and Technology **60**(1), 9–26 (2009)
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
12. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014), `http://www.aclweb.org/anthology/D14-1162`
13. Pulman, S., Mihalcea, R.: Linguistic ethnography: Identifying dominant word classes in text pp. 595–602 (2009)
14. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of the 2nd international workshop on Search and mining user-generated contents. pp. 37–44. ACM (2010)
15. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73791 (2013)
16. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology **29**(1), 24–54 (2010)
17. Tuulos, V.H., Tirri, H.: Combining topic models and social networks for chat data mining. In: Proceedings of the 2004 IEEE/WIC/ACM international Conference on Web intelligence. pp. 206–213. IEEE Computer Society (2004)