Using Latent Communication Styles to Predict Individual Characteristics

Jordan Bates*, Jennifer Neville*, and Jim Tyler**
*Computer Science Department, **Psychological Sciences Department
Purdue University
West Lafayette, IN USA
[jtbates | neville]@cs.purdue.edu

ABSTRACT

Data in online social network and social media systems provides a significant source of information about individual attitudes, preferences, and relationships. While there is a large body of work using statistical and machine learning techniques to predict the characteristics of text and users from documents, these efforts typically do not try to exploit the text to infer personality traits and understand interpersonal communications. However, some recent work in social psychology has focused on the aspects of writing and speech that are suggestive of personal characteristics. These efforts differ from much of the current work in text mining in that they focus on "style" rather than content—by modeling usage patterns involving the most frequent words (i.e., function words such as "the", "is", "a"). In this work, we identify stylistic patterns in communication data through the use of latent semantic analysis on function words. We use the discovered topics in logistic regression models and show that *style* is more predictive than *content* for several classification tasks focusing on personal traits such as gender, political party affiliation, verbal aggressiveness, and sentiment.

1. INTRODUCTION

Over the last ten years, online social networks (OSNs) and social media have become an integral aspect of the social fabric that today has far-reaching influence on nearly all aspects of our lives. The data in these online systems provides a significant source of information about individual attitudes and preferences, as well as social relationships. If machine learning methods can be developed to analyze this social and relational information, it will enable us to better understand how individual and peer characteristics interact to influence subsequent behavior. Ultimately, understanding these social processes may offer us mechanisms with which to automatically predict behavior from observational data in social systems.

There is a large body of work using statistical and machine learning techniques to predict the characteristics of text and users from documents. For example, researchers have used automated methods for analyzing sentiment [25], predicting reading level [11], detecting gender [35, 18], discriminating authors [4], evaluating text quality [1], and predicting genre category [30]. However, since these efforts are primarily investigated by the information retrieval and natural language processing communities, the focus is generally on automatically understanding textual content to drive the development of methods to organize, rank, and retrieve documents in an online setting. While there has been growing interest in methods that can automatically personalize systems to individual preferences and characteristics [32, 17], these efforts rarely try to exploit the text to infer personality traits and understand interpersonal communications.

In contrast, much of the research in social psychology studying personality has focused on the collection and analysis of self-report data. There has been some recent work that shows implicit behavioral traces communicate personal characteristics [22, 14]. Initially this work focused on physical traces that could be found in bedrooms and offices (e.g., books). However, the recent explosive growth in the use of the World Wide Web has produced a wealth of electronic traces in webpages and Facebook pages, which exhibit aspects of personality as well [34, 5]. Indeed, recent work in this direction suggests that OSN profile and activity data is correlated with each of the Big Five personality factors [15, 13] and that Twitter activity data is correlated with at least three of the five personality factors [28]. This related work suggests that communicative acts contain implicit behavioral information that is likely to provide a rich source of data to automatically identify and predict personal traits.

A related line of work in social psychology has focused on the aspects of writing and speech that are suggestive of personal characteristics [8, 24, 7, 10, 20]. These efforts differ from much of the current work in text mining in that they focus on "style" rather than content. Specifically, the social psychologists have focused on analyzing writing "style" through the use of functions words. There are less than 400 function words (pronouns, prepositions, articles, conjunctions, and auxiliary verbs), yet they account for over 50% of the words used in daily speech [29]. These function words correspond to what is conventionally found on "stop word" lists in the information retrieval community. Stop word lists are comprised of the most frequent words in the documents (e.g., "the", "is", "he", "a"). Since the majority of text mining efforts aim to automate Web search and retrieval, as a pre-processing step "stop words" are often dropped from the text before analysis, because they do not contain much

SOMA KDD 2012

information about the content (e.g., topic) of the document.

However, since the same content can be expressed in many different ways, one's *style* is determined through the particular choice and combination of function words in speech/writing. Thus, the way that people use function words can reflect their motives, needs, and important dimensions of personality. While these function words may not be discriminative with respect to content, nor do they help in understanding semantics, recent work in social psychology has shown that the use of function words reveal a range of personal characteristics, including gender, age, status, and self-esteem [10].

In this work, we aim to extend previous research and investigate whether function word patterns can be exploited to automatically predict behavior and traits. The recent work connecting stylistic patterns of writing (both in online and offline settings) to components of the Big Five personality factors [3, 24, 20] is suggestive that the stylistic patterns of communication will be indicative of other personality traits as well. Specifically, we use latent semantic analysis to automatically identify aspects of both style and content in Congressional speeches and inter-personal communication (both online and face-to-face). We then use the latent *style* topics and *content* topics to learn predictive models of individual traits. Our analysis shows that through the use of style topics we can significantly improve the predictive accuracy for several classification tasks including: political party, bill sentiment, gender, and verbal aggressiveness.

2. BACKGROUND AND RELATED WORK

Recent work that has examined individual characteristics in the context of social media and online social networks (OSNs) has examined the degree to which OSN information correlates with certain personality characteristics [15, 13, 28]. Gosling et al. [15] study how personality is reflected in OSNs by examining Facebook usage patterns. The authors find several connections between the Big Five personality traits (i.e., openness, conscientiousness, extraversion, agreeableness, and neuroticism) and profile information and Facebook-related behaviors. For example, both frequency of Facebook usage and engagement in the site were correlate with extraversion. The social engagement that extraverts seek out leaves behind a behavioral residue in the form of friends lists and picture postings. Golbeck et al. [13] gather all the public Facebook profile data from a set of 300 subjects and used this information to learn predictive models for each of the Big Five personality traits. Using 161 features describing the Facebook profile of each user, they are able to learn models that can predict each of the five personality factors to within 11% of the actual values. Quercia et al. [28] analyze the relationship between personality and different types of Twitter users, using a set of 300 subjects. The authors show that they can predict a user's personality simply based on three profile features: counts of following, followers, and listed. The results show that these three quantities can be used to predict the user's five personality traits with RMSE less than one on a 1-5 scale. Although these investigations have identified the information in online profiles and activity that can be used to predict personality traits, they are limited in their focus on individual level profile information rather than textual communication information.

On the other hand, previous research has shown that linguistic patterns are correlated with personality traits [8, 10, 20]. Campbell & Pennebaker [8] examine writing samples from undergraduates describing traumatic events to investigate the relationship between writing style and health practices. Using latent semantic analysis (LSA), the authors find that the use of function words (particularly personal pronouns) was related to positive health outcomes. Chung & Pennebaker [10] discuss research that has found (1) the use of first person singular is associated with negative affective states, (2) the combined use of first person singular pronouns and exclusive words is associated with honesty, and (3) the relative use of first person singular pronouns in dyads is indicative of the relative status of two people (the person with fewer "I" words tends have higher status). Furthermore, they report that there are sex differences in the use of virtually all function words: pronouns, prepositions, articles, and auxiliary verbs. Mairesse et al. [20] also analyze writing samples from undergraduates and learn classification, regression, and ranking models to predict the Big Five personality traits. The models are built from a wide range of linguistic features that include frequency counts of 88 word categories, features from the MRC Psycholinguistic database, and utterance types. The results confirm previous findings linking language and personality, while revealing many new linguistic markers.

Related work on learning descriptive models of documents or communications, including networks of linked documents, focuses primarily on the use of latent variable models (e.g., LSA [12], LDA [6]) to cluster the terms and/or documents into topics, roles, or groups. For example, McCallum et al. [21] learn topic distributions based on a network of email messages sent/received among users, where the textual content is on the links of the graph. The model assumes the following generative process: users have (latent) roles in the network (e.g., faculty, secretary), when a user writes an email, a language model is chosen based on the user's current role as well as the role of the recipient, the selected language model then determines the observed words in the email message. A complementary set of works consider similar latent variable models to understand patterns in hyperlinked document networks [9, 23, 31]. Here the textual content is on the nodes of the graph (i.e., documents) and the generative process assumes that documents are members of (latent) groups representing particular topics, then the latent group memberships determines both the content of the document (i.e., words) and the hyperlinks to other documents (i.e., references). The primary aim of these past modeling efforts has been on discovering clusters of documents and users to describe the overall patterns in the data, rather than to predict particular characteristics of the users. In our work, we will focus on developing predictive models of user traits (i.e., gender, political party) that can incorporate individual patterns of communication style.

Our work is based on findings in social psychology that have correlated aspects of writing and speech to personal characteristics [8, 24, 7, 10, 20]. This analysis diverges from conventional text mining and information retrieval in that it focuses more on linguistic style than content. The majority of natural language processing methods for Web search and retrieval remove "stop words" from the text as a preprocessing step before analysis. Stop word lists are comprised of the most frequent words in the documents since these words are thought to obscure the content information in less frequent words. The social psychologists however, observe that the same content can be expressed in many

different ways. Thus one's "style", as reflected through the particular choice and combination of function words in communication, is often indicative of personal traits and behavior. In this work, we aim to extend these findings and investigate whether patterns in functions word usage can be exploited to automatically infer personal traits—both in longer communications (e.g., speeches) and in shorter group communications (e.g., online chats and verbal discussion).

3. DATA

3.1 Speech corpus

Our first corpus consists of the ConVote dataset [33], which we refer to as the Speech corpus. ConVote is comprised of the set of speeches from the 2005 U.S. House record concerning a bill that went up for vote and received at least 20% 'Yes' and at least 20% 'No' votes. A speech is an uninterrupted utterance by a single Representative. In the Speech corpus documents correspond to unannotated speeches from ConVote ("data_stage_three"). The speeches are already tokenized with Penn Treebank tokenization [19]. We remove speeches with less than 150 words. The Speech corpus contains 365 Representatives, 53 bills, and 2071 speeches with a mean document length of 569 words. There are 24,194 distinct words in the corpus.

We use the speaker's party affiliation (Democrat/Republican) and the speaker's vote (Yes/No) on the bill discussed in the speech as prediction targets. The one Independent Representative, Bernie Sanders, caucuses with the Democrats, so we used this as his party affiliation. The Republicans were the majority party in 2005. There is good evidence to suggest that training a text classifier to predict party with speeches from a single session will produce a classifier that is more sensitive to indicators of party status (majority/minority) than of ideology [16].

3.2 Communications corpora

The second corpus consists of data gathered in a psychology laboratory experiment which focused on understanding the communication differences between distributed and co-located groups. During these experiments undergraduate students were assigned to teams of 3-5 individuals and given the task of working together to solve a complex task. Participants in the first phase of the experiment were assigned to *distributed* teams and the groups communicated using an online chat room. Participants in the second phase of the experiment were assigned to *collocated* teams and the groups communicated verbally (i.e., face-to-face) within a single room. The Phase II verbal communications were videotaped and then later transcribed to electronic form. We will hereby refer to Phase I as Chat and Phase II as Face-to-face.

Each team was given a logic problem to solve as a group during a 45 minute time period. An example of one such puzzle is given a set of names, occupations, and companies identify the occupation of each person and what company they work for using a set of constraints. After the session participants would complete surveys detailing the performance of each member including themselves and the performance of the group as a whole. The categories in which members were evaluated included involvement, trustworthiness, respectability, likability, conflict with the team, competence, task versus social (focus on the task compared to interested in conversation), dominance/assertiveness, nervousness, and productivity. The participants evaluated the performance of the group in the categories of trustworthiness, cohesion, satisfaction, productivity, and performance/effectiveness. Individuals also rated themselves in terms of communication anxiety, verbal aggressiveness, and self-esteem. Each response was given on a scale of 1-7 except for the self-esteem which was on a scale of 0-4. The transcripts of Face-to-face tended to be longer than those of Chat, with the participants of Face-to-face speaking an average of 31.2 times while the participants of Chat posted an average of 8.6 messages.

For the experiments in this paper, we created one document per participant consisting of all of that participant's utterances during the session. We tokenize these documents with the Penn Treebank tokenization. We remove documents with less than 25 words. We use the speaker's gender (Male/Female) as a prediction target. We also use Rosenberg self-esteem scores (on a scale 0-40) and verbal aggressiveness scores (0-80). We turn this into binary prediction tasks by removing the middle tertile and predicting whether the score is High or Low. This means that when we predict self-esteem and verbal aggressiveness the corpus size is reduced by one third.

Chat corpus.

There are 500 participants in 150 groups. Mean document length is 222 words. There are 5,149 distinct words.

Face-to-face corpus.

There are 276 participants in 77 groups. Mean document length is 1,077 words. There are 3,486 distinct words.

4. METHODS

We use techniques based on Latent Semantic Analysis (LSA [12]) to extract "style" and "content" topics of the textual data and investigate whether they predict speaker traits. LSA may be used to automatically extract k topics from a semantic space. This can be thought of as a type of dimensionality reduction where the documents are projected into k dimensions (or topics).

4.1 Pre-processing

As described in the data section, a document in the Speech corpus corresponds to a single uninterrupted utterance in the House Record. For the Chat and Face-to-face a document is the concatenation of all a participant's utterances in the study transcript. The documents are tokenized with Penn Treebank tokenization. We discard documents with fewer than 150 words for the Speech corpus and documents with fewer than 25 words for the Chat and Face-to-face corpora.

4.2 Term count model

To represent a semantic space with n words and m documents we use a term count model. This is a vector space representation consisting of an $n \times m$ term-document matrix (TDM) denoted M with element (i, j) representing the frequency of word w_i in document d_j. Column j of M is the document vector d_j.

4.3 Style and content semantic spaces

We split a corpus into two semantic spaces, style and content, represented by term-document matrices M_s and M_c.

The content semantic space is formed by removing all words in our set of style words S. Conversely, the style semantic space is formed by by removing all words *except* those in the style set. Thus, the rows of M_s are some subset of the rows of M corresponding to words in S and M_c is the complement. We use as our style set a list of function words that is the union of the WordNet stop list [26] and the function words defined by the English dictionary of the Linguistic Inquiry and Word Count software [27].

4.4 Latent semantic analysis

In latent semantic analysis, dimensionality reduction to k latent topics is accomplished via a truncated singular value decomposition. We use LSA on the training documents and this gives us their latent representation

$$M_k = U_k \Sigma_k V_k^T \tag{1}$$

where M_k is the rank k approximation of M with minimal error. Columns of the matrix V_k^T, called the right singular vectors, represent the length k latent feature vectors of the training documents. For document j the right singular vector is $\hat{d}_j = \Sigma_k^{-1} U_k^T d_j$. The left singular vectors (the columns of the $m \times k$ matrix U_k) represent the topics and the singular values (the diagonal entries of the $k \times k$ square diagonal matrix Σ_k) represent the topics' relative importance. The topics (left singular vectors) are a word combination pattern over all m words.

We keep the topics and their singular values from training and use them to transform test documents into the latent feature space via the equation

$$\hat{q} = \Sigma_k^{-1} U_k^T q \tag{2}$$

Words in the test documents which are not in the training set have no representation in the term count model created from the training set and therefore have no contribution to the latent features of the test documents.

We do LSA with $k = 10$ topics independently on the style semantic space and the content semantic space. This gives us the latent style space and the latent content space.

4.5 Evaluation method

To evaluate the efficacy of style and content features in predicting author attributes we use logistic regression. We first form the latent semantic spaces using LSA on the training set. This gives us the k left singular vectors (topics) in U_k, the k singular values in the diagonal matrix Σ_k, and the n right singular vectors (document latent feature representations) in V_k^T. A logistic regression model is trained on the 10 continuous features from the latent feature representations of the training documents to predict a binary class label. We find the accuracy of the trained logistic regression model on the test documents by classifying their latent feature representations, found with Equation 2.

We predict party (Republican/Democrat) and vote (Yes/-No) for the Speech corpus. We predict gender (Male/Female), self-esteem (High/Low), and verbal aggressiveness (High/Low) for the Chat and Face-to-face corpora. High and Low correspond to the top and bottom tertiles of the score. We do not use documents with scores in the middle tertile when training and evaluating predictive models of self-esteem and verbal aggressiveness.

We perform our evaluation with stratified 10-fold cross-validation. We compare style and content trained classifiers against one another and against a baseline majority classifier. We compare accuracies across the 10 folds with a paired t-test. Our hypothesis is that style topics will more accurately model speaker attributes than content topics because personal attributes influence the speaking style more directly and reliably than the content of the speech.

5. EXPERIMENTAL RESULTS

We present the experimental results for each of the 12 prediction tasks. For each fold we use the training set to find style and content topics and the majority class. We use logistic regression to learn a style based predictive model and a content based predictive model which are then evaluated on the test set along with a baseline majority classifier. We display the mean accuracy in a graph and discuss significance in the text.

To get an idea of what the latent topics for each corpus are, we do LSA on the style and content semantic spaces of the entire corpus and train a classifier for one prediction task. We determine the most predictive topics by the absolute value of the regression coefficients. We then present the top 7 positively contributing words and top 7 negatively contributing words for the three most discriminative style topics and the three most discriminative content topics. Note that multiplying a left singular value (topic vector), the corresponding latent feature (row of V_k^T), and corresponding regression coefficient by -1 leaves the semantic space and predictive model unchanged. We present the most predictive topics so that they are all oriented towards the same class label (the regression coefficients have the same sign).

5.1 Speech results

Baseline, style, and content predictive models of party and vote are evaluated for the Speech corpus. Their mean accuracies are presented in Figure 1. Style significantly outperforms baseline ($p < .0001$) and content ($p = .02$) in predicting the speaker's party affiliation. Style also significantly outperforms baseline ($p < .0001$) and content ($p = .01$) in predicting vote on the bill under discussion. Content is significantly better than the baseline for both party ($p = .002$) and vote ($p = .02$).

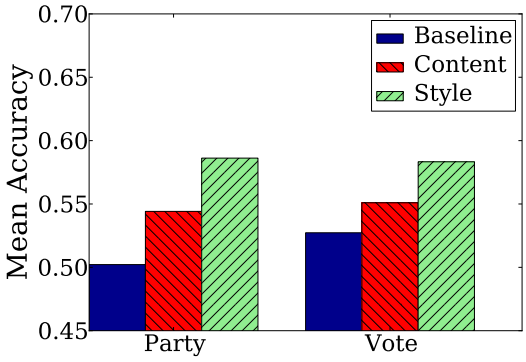

Figure 1: Results on Speech data.

Example style and content topics are learned for the entire Speech corpus. We train regression models on the resulting style and content latent feature spaces to predict party. The

three most discriminative topics for predicting party are presented for the style model in Figure 2 and for the content model in Figure 3. Latent topics do not necessarily (and often do not) have an interpretable meaning. However, it is interesting to note that the second most discriminative style topic in 2 when oriented towards Republicans has positive weights on the words "we" and "our" and negative weights on the word "I" and "my". This makes sense in light of the fact that the Republicans were the majority party at the time and are generally a more cohesive group than the Democrats. It is also noteworthy that the words "Mr.", "Speaker", and "Chairman" have high positive weights in the content topics in Figure 3. These words are honorifics and their use might be indicative of a more respectful tone. This points to the difficulty of separating notions of style and content which we will discuss later.

The three most discriminative style topics for predicting party, oriented with positive towards Republican		
Topic 8	Topic 9	Topic 4
.49∗"and"	.58'∗"we"	.71∗";"
.34∗"of"	.22∗"a"	.45∗"&"
.34∗"that"	.22∗"in"	.14∗"("
.26∗"i"	.18∗"to"	.14∗")"
.16∗"a"	.14∗"of"	.14∗"."
.14∗"was"	.12∗"our"	.13∗"that"
.10∗"in"	.11∗"have"	.12∗"is"
⋮	⋮	⋮
−.10∗";"	−.10∗"my"	−.06∗"$"
−.11∗"is"	−.12∗"not"	−.07∗"in"
−.13∗"the"	−.14∗"and"	−.09∗"for"
−.16∗"for"	−.19∗"it"	−.11∗"'s"
−.19∗"$"	−.24∗"is"	−.15∗"to"
−.22∗","	−.31∗"i"	−.15∗"and"
−.40∗"to"	−.37∗"this"	−.23∗","

Figure 2: Example style topics for Speech data

The three most discriminative content topics for predicting party, oriented with positive towards Republican		
Topic 6	Topic 9	Topic 4
.25∗"theresa"	.21∗"trade"	.21∗"law"
.21∗"mr."	.19∗"security"	.21∗"bill"
.16∗"michael"	.17∗"bankruptcy"	.15∗"act"
.15∗"court"	.17∗"committee"	.14∗"committee"
.15∗"trade"	.15∗"united"	.10∗"security"
.15∗"schiavo"	.12∗"homeland"	.08∗"chairman"
.15∗"speaker"	.10∗"debtors"	.08∗"energy"
⋮	⋮	⋮
−.13∗"programs"	−.14∗"budget"	−.16∗"care"
−.13∗"funding"	−.17∗"schiavo"	−.21∗"health"
−.14∗"h.r."	−.19∗"court"	−.21∗"debtors"
−.14∗"safety"	−.20∗"michael"	−.24∗"tax"
−.24∗"health"	−.21∗"bill"	−.26∗"percent"
−.25∗"bill"	−.23∗"tax"	−.29∗"bankruptcy"
−.26∗"budget"	−.31∗"theresa"	−.34∗"medical"

Figure 3: Example content topics for Speech data

5.2 Chat results

Baseline, style, and content predictive models of gender, self-esteem, and verbal aggressiveness are evaluated for the Chat corpus. Their mean accuracies are presented in Figure 4. Style outperforms content in all three prediction tasks. This advantage is weakly significant for gender ($p = .08$) and verbal aggressiveness ($p = .06$). However, neither style nor content performs significantly better than the baseline on any of the prediction tasks for the Chat corpus.

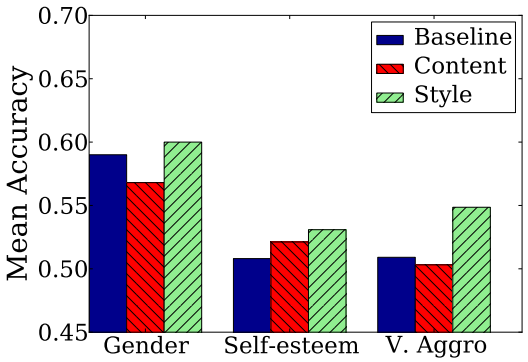

Figure 4: Results on Chat data.

The three most discriminative style topics for predicting gender, oriented with positive towards Male		
Topic 3	Topic 9	Topic 5
.37∗"the"	.39'∗"a"	.54∗"."
.33∗"i"	.38∗"is"	.52∗"?"
.18∗"you"	.29∗"to"	.26∗"the"
.14∗"it"	.20∗"you"	.16∗"you"
.13∗"to"	.20∗"it"	.14∗"!"
.08∗"of"	.19∗"!"	.09∗"do"
.08∗"in"	.13∗"this"	.05∗"of"
⋮	⋮	⋮
−.12∗"or"	−.11∗"have"	−.08∗"not"
−.14∗":"	−.21∗"on"	−.08∗"at"
−.15∗"be"	−.21∗"so"	−.09∗"it"
−.20∗"."	−.24∗"i"	−.14∗"and"
−.28∗"so"	−.24∗"that"	−.15∗"so"
−.37∗"?"	−.24∗"we"	−.25∗"i"
−.50∗"is"	−.27∗":"	−.34∗"is"

Figure 5: Example style topics for Chat data

We perform LSA on the style and content semantic spaces of the entire Chat corpus. We use these latent feature representations to train style and content logistic regression models for predicting gender. The three most discriminative topics with respect to gender are presented for the style model in Figure 5 and for the content model in Figure 6. We see words related to the logic problem such as days, names, and professions are prominently featured in the content topics in Figure 6. We also see punctuation such as "..." that our style list missed. Interestingly, there are highly weighted words such as "ok" and "haha" which are not function words per se, but might still be considered a style word by merit of revealing relatively little about content and relatively more

The three most discriminative content topics for predicting gender, oriented with positive towards Male		
Topic 2	Topic 6	Topic 8
.29∗“–”	.31‘∗“started”	.24∗“last”
.26∗“engineer”	.27∗“new”	.23∗“names”
.24∗“10”	.25∗“person”	.21∗“name”
.18∗“dr.”	.24∗“10”	.16∗“frank”
.17∗“11”	.20∗“dr.”	.15∗“new”
.16∗“05”	.16∗“11”	.15∗“wild”
.16∗“times”	.16∗“worked”	.14∗“got”
⋮	⋮	⋮
−.09∗“cynthia”	−.10∗“frank”	−.19∗“ok”
−.10∗“west”	−.14∗“engineer”	−.19∗“wednesday”
−.10∗“north”	−.17∗“lol”	−.21∗“started”
−.12∗“ok”	−.17∗“sales”	−.23∗“monday”
−.12∗“started”	−.19∗“manager”	−.25∗“haha”
−.14∗“lol”	−.19∗“yeah”	−.25∗“thursday”
−.59∗“. . . ”	−.28∗“ok”	−.27∗“tuesday”

Figure 6: Example content topics for Chat data

about style.

5.3 Face-to-face corpus

We train predictive models of gender, self-esteem, and verbal aggressiveness with the Face-to-face corpus. For each prediction task we train a baseline majority model and two logistic regression models, one on the latent style features and one on the latent content features. Evaluation of these three models is performed across a stratified 10-fold cross-validation. The mean accuracies of each predictive model are presented in Figure 7. Style is significantly better at predicting gender than both content ($p = .03$) and baseline ($p = .02$). Style has a weakly significant advantage over content for predicting verbal aggressiveness ($p = .08$). However, style is not significantly better than the baseline for aggressiveness nor for self-esteem. Content does not perform significantly better than the baseline on any task for the Face-to-face corpus.

We use LSA to find the latent style and latent content semantic spaces of the entire Face-to-face corpus. We use these latent features to train style and content logistic regression models for predicting gender. The three most discriminative style topics with respect to gender are presented in Figure 8 and the three most discriminative content topics are shown in Figure 9. We see some similarities to the topics for the Chat corpus in Figures 5 and 6. For example, in both the style and content models the most discriminative topic when oriented towards predicting Male has “engineer” with a very positive weight and “. . . ” with a very negative weight.

6. DISCUSSION AND CONCLUSION

Function word usage, or writing style more generally, reveals useful information about the author. Our results are consistent with our hypothesis that stylistic features are more useful than content features for text analysis tasks involving author attributes. For such tasks, the information discarded with standard stop word lists may be more revealing than what is left behind.

We obtain the most significant results on the Speech corpus which is the largest corpus. We do not obtain significant

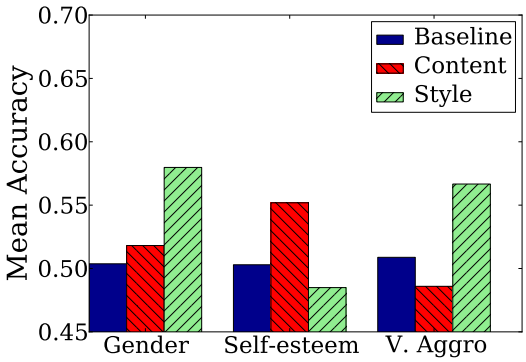

Figure 7: Results on Face-to-face data.

The three most discriminative style topics for predicting gender, oriented with positive towards Male		
Topic 8	Topic 2	Topic 7
.36∗“not”	.41‘∗“.”	.41∗“?”
.34∗“'s”	.35∗“,”	.33∗“'s”
.32∗“she”	.29∗“-”	.24∗“n't”
.31∗“and”	.09∗“'s”	.21∗“on”
.30∗“that”	.05∗“no”	.18∗“be”
.27∗“a”	.05∗“like”	.18∗“so”
.19∗“on”	.04∗“all”	.15∗“ca”
⋮	⋮	⋮
−.12∗“who”	−.11∗“for”	−.10∗“and”
−.13∗“has”	−.12∗“that”	−.11∗“we”
−.14∗“,”	−.14∗“so”	−.11∗“have”
−.15∗“it”	−.17∗“be”	−.15∗“was”
−.17∗“be”	−.23∗“on”	−.16∗“.”
−.21∗“:”	−.27∗“n't”	−.25∗“i”
−.26∗“?”	−.55∗“the”	−.48∗“the”

Figure 8: Example style topics for Face-to-face data

results on the Chat corpus which is the smallest. We obtain significant results for gender on the Face-to-face corpus but not for self-esteem or verbal aggressiveness, where we the corpus size is reduced by removing participants with scores middle tertile. This is consistent with previous work concluding that the corpus must be sufficiently large to discern author attributes [3].

Function word lists may be incomplete. Our style set missed some punctuation, such as “. . . ” and “–”. More importantly there other words like “yeah” and “oh” that are not strictly speaking function words but that also carry little information about the content of the text while still being indicative of its style. This can be coped with by defining the set of style words as the most frequent words in the corpus [8]. We found that this approach was helpful with the Speech corpus but not with the Chat or Face-to-face corpora. We believe this is because the participants in these experiments were all discussing the same logic problem. Therefore words related to the logic problem are among the most frequent words, but their use is not indicative of style.

Although there are words which we can consider “style” words because they bear little relation to any particular topic, words can also be indicative of both content and style.

The three most discriminative style topics for predicting gender, oriented with positive towards Male		
Topic 4	Topic 7	Topic 5
.29∗“know”	.37‘∗“name”	.34∗“yeah”
.28∗“engineer”	.36∗“last”	.20∗“oh”
.25∗“frank”	.28∗“silvia”	.17∗“right”
.23∗“sales”	.26∗“elliot”	.15∗“okay”
.20∗“manager”	.22∗“south”	.15∗“works”
.19∗“elliot”	.17∗“wild”	.15∗“wednesday”
.16∗“test”	.15∗“frank”	.14∗“north”
⋮	⋮	⋮
−.15∗“oh”	−.12∗“okay”	−.14∗“job”
−.17∗“person”	−.15∗“manager”	−.15∗“start”
−.17∗“wait”	−.17∗“sales”	−.18∗“started”
−.18∗“yeah”	−.19∗“know”	−.24∗“person”
−.22∗“started”	−.21∗“thursday”	−.26∗“worked”
−.24∗“. . . ”	−.21∗“tuesday”	−.26∗“know”
−.27∗“okay”	−.27∗“started”	−.35∗“new”

Figure 9: Example content topics for Face-to-face data

For example the words hood, trunk, and muffler indicate that a text is about cars, but also that it's author is American because the words bonnet, boot, and silencer were not used instead. Style is counterfactual in this way; given a particular denotational meaning (content), style is characterized in relation to the set of possible expressions of that meaning [2]. Splitting a semantic space into style and content using a simple word list is therefore insufficient. Further work is needed to develop more sophisticated techniques to jointly determine content and stylistic features thereby making it possible to distinguish an author's tendency to choose a particular subject matter from the tendency to use a specific vocabulary and style when discussing that subject.

Acknowledgements

This research is supported by NSF under contract number SES-0823313. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of NSF or the U.S. Government.

7. REFERENCES

[1] B. Amento, L. Terveen, and W. Hill. Does authority mean quality? predicting expert quality ratings of web documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 296–303, New York, NY, USA, 2000. ACM.

[2] S. Argamon and M. Koppel. The rest of the story: Finding meaning in stylistic variation. In *The Structure of Style: Algorithmic Approaches to Understanding Manner and Meaning*, volume 1, pages 79–112. Springer-Verlag, New York, NY, 2010.

[3] S. Argamon and S. Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 ACH/ALLC Conference*, Victoria, BC, 2005.

[4] S. Argamon, M. Šarić, and S. S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 475–480, New York, NY, USA, 2003. ACM.

[5] M. D. Back, J. M. Stopfer, S. Vazire, S. Gaddis, S. C. Schmukle, B. Egloff, and S. D. Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, 21(3):372, 2010.

[6] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] M. D. Boni, M. Dias, and R. Hurling. Automatically predicting dominant and submissive personality types from text. In *Proceedings of the IADIS International Conference Applied Computing*, 2006.

[8] R. S. Campbell and J. W. Pennebaker. The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, 14:60–65, 2003.

[9] J. Chang. Relational topic models for document networks. In *In Proc. of Conf. on AI and Statistics (AISTATS*, 2009.

[10] C. Chung and J. Pennebaker. The psychological function of function words. *Social communication: Frontiers of social psychology*, pages 343–359, 2007.

[11] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *J. Am. Soc. Inf. Sci. Technol.*, 56(13):1448–1462, Nov. 2005.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[13] J. Golbeck, C. Robles, and K. Turner. Predicting personality with social media. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, pages 253–262, New York, NY, USA, 2011. ACM.

[14] S. D. Gosling. *Snoop: What Your Stuff Says About You*. Basic Books, first trade paper edition edition, May 2009.

[15] S. D. Gosling, A. A. Augustine, S. Vazire, N. Holtzman, and S. Gaddis. Manifestations of Personality in Online Social Networks: Self-Reported Facebook-Related Behaviors and Observable Profile Information. *Cyberpsychology, behavior and social networking*, Jan. 2011.

[16] G. Hirst, Y. Riabinin, and J. Graham. Party status as a confound in the automatic classification of political speech by ideology. In *Proceedings of JADT 2010*, 2010.

[17] J. Koren, Y. Zhang, and X. Liu. Personalized interactive faceted search. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 477–486, New York, NY, USA, 2008. ACM.

[18] H. Liu and R. Mihalcea. Of men, women, and computers: Data-driven gender modeling for improved user interfaces. In *International Conference on Weblogs and Social Media*, 2007.

[19] R. MacIntyre. Penn treebank tokenization, 1995.

[20] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, Vol*, 30:457–501, 2007.

[21] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30(1):249–272, Oct. 2007.

[22] M. R. Mehl, S. D. Gosling, and J. W. Pennebaker. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90(5):862–877, 2006.

[23] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[24] J. Oberlander and S. Nowson. Whose thumb is it anyway?: classifying author personality from weblog text. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 627–634, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[25] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, Jan. 2008.

[26] T. Pedersen. WordNet stop list. http://www.d.umn.edu/~tpederse/Group01/WordNet/wordnet-stoplist.html. [Online; accessed 2012-05-11].

[27] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. *The development and psychometric properties of LIWC2007*. LIWC.net, Austin, TX, 2007.

[28] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our Twitter profiles, our selves: Predicting personality with Twitter. In *Proceedings of the IEEE Third International Conference on Social Computing (SocialCom)*, pages 180–185. IEEE, Oct. 2011.

[29] E. Rochon, E. M. Saffran, R. S. Berndt, and M. F. Schwartz. Quantitative analysis of aphasic sentence production: further development and new data. *Brain Lang*, 72(3):193–218, 2000.

[30] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 808–814, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[31] Y. Sun, J. Han, J. Gao, and Y. Yu. itopicmodel: Information network-integrated topic modeling. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, ICDM '09, pages 493–502, Washington, DC, USA, 2009. IEEE Computer Society.

[32] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 449–456, New York, NY, USA, 2005. ACM.

[33] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, 2006.

[34] S. Vazire and S. D. Gosling. e-Perceptions: Personality Impressions Based on Personal Websites. *Journal of Personality and Social Psychology*, 87(1):123–132, July 2004.

[35] X. Yan and L. Yan. Gender classification of weblog authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 228–230, 2006.