Person Identification from Text and Speech Genre Samples

Jade Goldstein-Stewart
U.S. Department of Defense

jadeg@acm.org

Ransom Winder
The MITRE Corporation
Hanover, MD, USA
rwinder@mitre.org

Roberta Evans Sabin
Loyola University
Baltimore, MD, USA
res@loyola.edu

Abstract

In this paper, we describe experiments conducted on identifying a person using a novel unique correlated corpus of text and audio samples of the person's communication in six genres. The text samples include essays, emails, blogs, and chat. Audio samples were collected from individual interviews and group discussions and then transcribed to text. For each genre, samples were collected for six topics. We show that we can identify the communicant with an accuracy of 71% for six fold cross validation using an average of 22,000 words per individual across the six genres. For person identification in a particular genre (train on five genres, test on one), an average accuracy of 82% is achieved. For identification from topics (train on five topics, test on one), an average accuracy of 94% is achieved. We also report results on identifying a person's communication in a genre using text genres only as well as audio genres only.

1 Introduction

Can one identify a person from samples of his/her communication? What common patterns of communication can be used to identify people? Are such patterns consistent across varying genres?

People tend to be interested in subjects and topics that they discuss with friends, family, colleagues and acquaintances. They can communicate with these people textually via email, text messages and chat rooms. They can also communicate via verbal conversations. Other forms of communication could include blogs or even formal writings such as essays or scientific articles. People communicating in these different "genres" may have different stylistic patterns and we are interested in whether or not we could identify people from their communications in different genres.

The attempt to identify authorship of written text has a long history that predates electronic computing. The idea that features such as average word length and average sentence length could allow an author to be identified dates to Mendenhall (1887). Mosteller and Wallace (1964) used function words in a groundbreaking study that identified authors of *The Federalist Papers*. Since then many attempts at authorship attribution have used function words and other features, such as word class frequencies and measures derived from syntactic analysis, often combined using multivariable statistical techniques.

Recently, McCarthy (2006) was able to differentiate three authors' works, and Hill and Provost (2003), using a feature of co-citations, showed that they could successfully identify scientific articles by the same person, achieving 85% accuracy when the person has authored over 100 papers. Levitan and Argamon (2006) and McCombe (2002) further investigated authorship identification of *The Federalist Papers* (three authors).

The genre of the text may affect the authorship identification task. The attempt to characterize genres dates to Biber (1988) who selected 67 linguistic features and analyzed samples of 23 spoken and written genres. He determined six factors that could be used to identify written text. Since his study, new "cybergenres" have evolved, including email, blogs, chat, and text messaging. Efforts have been made to characterize the linguistic features of these genres (Baron, 2003; Crystal, 2001; Herring, 2001; Shepherd and Watters, 1999; Yates, 1996). The task is complicated by the great diversity that can be exhibited within even a single genre. Email can be business-related, personal, or spam; the style

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 336–344,
Athens, Greece, 30 March – 3 April 2009. ©2009 Association for Computational Linguistics

can be tremendously affected by demographic factors, including gender and age of the sender. The context of communication influences language style (Thomson and Murachver, 2001; Coupland, et al., 1988). Some people use abbreviations to ease the efficiency of communication in informal genres – items that one would not find in a formal essay. Informal writing may also contain emoticons (e.g., ":-)" or "☺") to convey mood.

Successes have been achieved in categorizing web page decriptions (Calvo, et al., 2004) and genre determination (Goldstein-Stewart, et al., 2007; Santini 2007). Genders of authors have been successfully identified within the British National Corpus (Koppel, et al., 2002). In authorship identification, recent research has focused on identifying authors within a particular genre: email collections, news stories, scientific papers, listserv forums, and computer programs (de Vel, et al., 2001; Krsul and Spafford, 1997; Madigan, et al., 2005; McCombe, 2002). In the KDD Cup 2003 Competitive Task, systems attempted to identify successfully scientific articles authored by the same person. The best system (Hill and Provost, 2003) was able to identify successfully scientific articles by the same person 45% of the time; for authors with over 100 papers, 85% accuracy was achieved.

Are there common features of communication of an individual across and within genres? Undoubtedly, the lack of corpora has been an impediment to answering this question, as gathering personal communication samples faces considerable privacy and accessibility hurdles. To our knowledge, all previous studies have focused on individual communications in one or possibly two genres.

To analyze, compare, and contrast the communication of individuals across and within different modalities, we collected a corpus consisting of communication samples of 21 people in six genres on six topics. We believe this corpus is the first attempt to create such a correlated corpus.

From this corpus, we are able to perform experiments on person identification. Specifically, this means recognizing which individual of a set of people composed a document or spoke an utterance which was transcribed. We believe using text and transcribed speech in this manner is a novel research area. In particular, the following types of experiments can be performed:

- Identification of person in a novel genre (using five genres as training)
- Identification of person in a novel topic (using five topics as training)
- Identification of person in written genres, after training on the two spoken genres
- Identification of person in spoken genres, after training on the written genres
- Identification of person in written genres, after training on the other written genres

In this paper, we discuss the formation and statistics of this corpus and report results for identifying individual people using techniques that utilize several different feature sets.

2 Corpus Collection

Our interest was in the research question: can a person be identified from their writing and audio samples? Since we hypothesize that people communicate about items of interest to them across various genres, we decided to test this theory. Email and chat were chosen as textual genres (Table 1), since text messages, although very common, were not easy to collect. We also collected blogs and essays as samples of textual genres. For audio genres, to simulate conversational speech as much as possible, we collected data from interviews and discussion groups that consisted of sets of subjects participating in the study. Genres labeled "peer give and take" allowed subjects to interact.

Such a collection of genres allows us to examine both conversational and non-conversational genres, both written and spoken modalities, and both formal and informal writing with the aim of contrasting and comparing computer-mediated and non-computer-mediated genres as well as informal and formal genres.

Genre	Computer-mediated	Peer Give and Take	Mode	Conversational	Audience
Email	yes	no	text	yes	addressee
Essay	No	no	text	no	unspec
Interview	No	no	speech	yes	interviewer
Blog	yes	yes	text	no	world
Chat	yes	yes	text	yes	group
Discussion	No	yes	speech	yes	group

Table 1. Genres

In order to ensure that the students could produce enough data, we chose six topics that were controversial and politically and/or socially relevant for college students from among whom the subjects would be drawn. These six topics were chosen from a pilot study consisting of twelve topics, in which we analyzed the amount of information that people tended to "volunteer" on the topics as well as their thoughts about being able to write/speak on such a topic. The six topics are listed in Table 2.

Topic	Question
Church	Do you feel the Catholic Church needs to change its ways to adapt to life in the 21st Century?
Gay Marriage	While some states have legalized gay marriage, others are still opposed to it. Do you think either side is right or wrong?
Privacy Rights	Recently, school officials prevented a school shooting because one of the shooters posted a myspace bulletin. Do you think this was an invasion of privacy?
Legalization of Marijuana	The city of Denver has decided to legalize small amounts of marijuana for persons over 21. How do you feel about this?
War in Iraq	The controversial war in Iraq has made news headlines almost every day since it began. How do you feel about the war?
Gender Discrimination	Do you feel that gender discrimination is still an issue in the present-day United States?

Table 2. Topics

The corpus was created in three phases (Goldstein-Stewart, 2008). In Phase I, emails, essays and interviews were collected. In Phase II, blogs and chat and discussion groups were created and samples collected. For blogs, subjects blogged over a period of time and could read and/or comment on other subjects' blogs in their own blog. A graduate research assistant acted as interviewer and discussion and chat group moderator.

Of the 24 subjects who completed Phase I, 7 decided not to continue into Phase II. Seven additional students were recruited for Phase II. In Phase III, these replacement students were then asked to provide samples for the Phase I genres. Four students fully complied, resulting in a corpus with a full set of samples for 21 subjects, 11 women and 10 men.

All audio recordings, interviews and discussions, were transcribed. Interviewer/moderator comments were removed and, for each discussion, four individual files, one for each participant's contribution, were produced.

Our data is somewhat homogeneous: it samples only undergraduate university students and was collected in controlled settings. But we believe that controlling the topics, genres, and demographics of subjects allows the elimination of many variables that effect communicative style and aids the identification of common features.

3 Corpus Statistics

3.1 Word Count

The mean word counts for the 21 students per genre and per topic are shown in Figures 1 and 2, respectively. Figure 1 shows that the students produced more content in the directly interactive genres – interview and discussion (the spoken genres) as well as chat (a written genre).

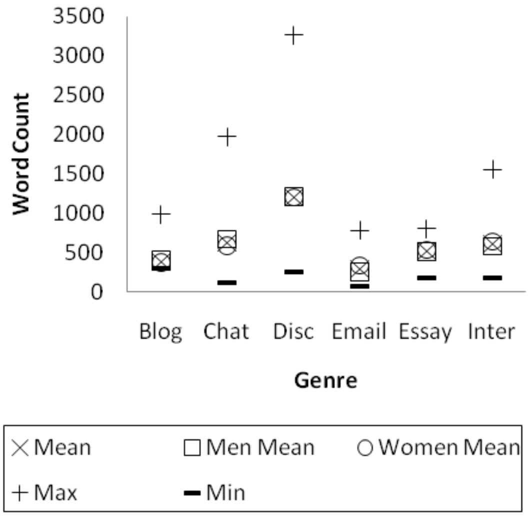

Figure 1. Mean word counts for gender and genre

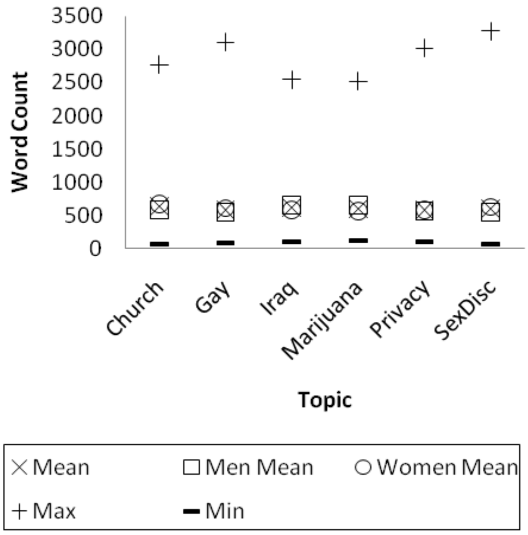

Figure 2. Mean word counts for gender and topic

The email genre had the lowest mean word count, perhaps indicating that it is a genre intended for succinct messaging.

3.2 Word Usage By Individuals

We performed an analysis of the word usage of individuals. Among the top 20 most frequently occurring words, the most frequent word used by all males was "the". For the 11 females, six most frequently used "the", four used "I", and one used "like". Among abbreviations, 13 individuals used "lol". Abbreviations were mainly used in chat. Other abbreviations were used to varying degrees such as the abbreviation "u". Emoticons were used by five participants.

4 Classification

4.1 Features

Frequencies of words in word categories were determined using Linguistic Inquiry and Word Count (LIWC). LIWC2001 analyzes text and produces 88 output variables, among them word count and average words per sentence. All others are percentages, including percentage of words that are parts of speech or belong to given dictionaries (Pennebaker, et al., 2001). Default dictionaries contain categories of words that indicate basic emotional and cognitive dimensions and were used here. LIWC was designed for both text and speech and has categories, such negations, numbers, social words, and emotion. Refer to LIWC (www.liwc.net) for a full description of categories. Here the 88 LIWC features are denoted feature set L.

From the original 24 participants' documents and the new 7 participants' documents from Phase II, we aggregated all samples from all genres and computed the top 100 words for males and for females, including stop words. Six words differed between males and females. Of these top words, the 64 words with counts that varied by 10% or more between male and female usage were selected. Excluded from this list were 5 words that appeared frequently but were highly topic-specific: "catholic", "church", "marijuana", "marriage", and "school."

Most of these words appeared on a large stop word list (www.webconfs.com/stop-words.php). Non-stop word terms included the word "feel", which was used more frequently by females than males, as well as the terms "yea" and "lot" (used more commonly by women) and "uh" (used more commonly by men). Some stop words were used more by males ("some", "any"), others by females ("I", "and"). Since this set mainly consists of stop words, we refer to it as the functional word features or set F.

The third feature set (T) consisted of the five topic specific words excluded from F.

The fourth feature set (S) consisted of the stop word list of 659 words mentioned above.

The fifth feature set (I) we consider informal features. It contains nine common words not in set S: "feel", "lot", "uh", "women", "people", "men", "gonna", "yea" and "yeah". This set also contains the abbreviations and emotional expressions "lol", "ur", "tru", "wat", and "haha". Some of the expressions could be characteristic of particular individuals. For example the term "wat" was consistently used by one individual in the informal chat genre.

Another feature set (E) was built around the emoticons that appeared in the corpus. These included ":)", ":(", ":-(", ";)", ":-/", and ">:o)".

For our results, we use eight feature set combinations: 1. All 88 LIWC features (denoted L); 2. LIWC and functional word features, (L+F); 3. LIWC plus all functional word features and the topic words (L+F+T); 4. LIWC plus all functional word features and emoticons (L+F+E); 5. LIWC plus all stop word features (L+S); 6. LIWC plus all stop word and informal features (L+S+I); 7. LIWC supplemented by informal, topic, and stop word features, (L+S+I+T). Note that, when combined, sets S and I cover set F.

4.2 Classifiers

Classification of all samples was performed using four classifiers of the Weka workbench, version 3.5 (Witten and Frank, 2005). All were used with default settings except the Random Forest classifier (Breiman, 2001), which used 100 trees. We collected classification results for Naïve-Bayes, J48 (decision tree), SMO (support vector machine) (Cortes and Vapnik, 1995; Platt, 1998) and RF (Random Forests) methods.

5 Person Identification Results

5.1 Cross Validation Across Genres

To identify a person as the author of a text, six fold cross validation was used. All 756 samples were divided into 126 "documents," each consisting of all six samples of a person's expression in a single genre, regardless of topic. There is a baseline of approximately 5% accuracy if randomly guessing the person. Table 3 shows the accuracy results of classification using combinations of the feature sets and classifiers.

The results show that SMO is by far the best classifier of the four and, thus, we used only this classifier on subsequent experiments. L+S performed better alone than when adding the informal features – a surprising result.

Table 4 shows a comparison of results using feature sets L+F and L+F+T. The five topic words appear to grant a benefit in the best trained case (SMO).

Table 5 shows a comparison of results using feature sets L+F and L+F+E, and this shows that the inclusion of the individual emoticon features does provide a benefit, which is interesting considering that these are relatively few and are typically concentrated in the chat documents.

Feature	SMO	RF100	J48	NB
L	52	30	15	17
L+F	60	44	21	25
L+S	71	42	19	33
L+S+I	71	39	17	33
L+S+I+T	71	40	17	33

Table 3. Person identification accuracy (%) using six fold cross validation

Feature	SMO	RF100	J48	NB
L+F	60	44	21	25
L+F+T	67	40	21	25

Table 4. Accuracy (%) using six fold cross validation with and without topic word features (T)

Feature	SMO	RF100	J48	NB
L+F	60	44	21	25
L+F+E	65	41	21	25

Table 5. Accuracy (%) using six fold cross validation with and without emoticon features (E)

5.2 Predict Communicant in One Genre Given Information on Other Genres

The next set of experiments we performed was to identify a person based on knowledge of the person's communication in other genres. We first train on five genres, and we then test on one – a "hold out" or test genre.

Again, as in six fold cross validation, a total of 126 "documents" were used: for each genre, 21 samples were constructed, each the concatenation of all text produced by an individual in that genre, across all topics. Table 6 shows the results of this experiment. The result of 100% for L+F, L+F+T, and L+F+E in email was surprising, especially since the word counts for email were the lowest. The lack of difference in L+F and L+F+E results is not surprising since the emoticon features appear only in chat documents, with one exception of a single emoticon in a blog document (":-/"), which did not appear in any chat documents. So there was no emoticon feature that appeared across different genres.

SMO	HOLD OUT (TEST GENRE)						
Features	A	B	C	D	E	S	I
L	**60**	76	52	43	76	81	29
L+F	**75**	81	57	48	**100**	90	71
L+F+T	**76**	86	62	52	**100**	86	71
L+F+E	**75**	81	57	48	**100**	90	71
L+S	**82**	81	67	67	86	90	**100**
L+S+I	**79**	86	52	57	86	90	**100**
L+S+I+T	**81**	86	52	67	90	90	**100**

Table 6. Person identification accuracy (%) training with SMO on 5 genres and testing on 1. A=Average over all genres, B=Blog, C=Chat, D=Discussion, E=Email, S=Essay, I=Interview

Train	Test	L+F	L+F+T
CDSI	Email	67	95
BDSI	Email	71	52
BCSI	Email	76	100
BCDI	Email	57	90
BCDS	Email	57	81

Table 7. Accuracy (%) using SMO for predicting email author after training on 4 other genres. B=Blog, C=Chat, D=Discussion, S=Essay, I=Interview

We attempted to determine which genres were most influential in identifying email authorship, by reducing the number of genres in its training set. Results are reported in Table 7. The difference between the two sets, which differ only in five topic specific word features, is more marked here. The lack of these features causes accuracy to drop far more rapidly as the training set is reduced. It also appears that the chat genre is important when identifying the email genre when topical features are included. This is probably not just due to the volume of data since discussion groups also have a great deal of data. We need to investigate further the reason for such a high performance on the email genre.

The results in Table 6 are also interesting for the case of L+S (which has more stop words than L+F). With this feature set, classification for the interview genre improved significantly, while that of email decreased. This may indicate that the set of stop words may be very genre specific – a hypothesis we will test in future work. If this in indeed the case, perhaps certain different sets of stop words may be important for identifying certain genres, genders and individual authorship. Previous results indicate that the usage of certain stop words as features assists with identifying gender (Sabin, et al., 2008).

Table 6 also shows that, using the informal words (feature set I) decreased performance in two genres: chat (the genre in which the abbreviations are mostly used) and discussion. We plan to run further experiments to investigate this. The sections that follow will typically show the results achieved with L+F and L+S features.

Train\Test	B	C	D	E	S	I
Blog	100	14	14	76	57	5
Chat	24	100	29	38	19	10
Discussion	10	5	100	5	10	29
Email	43	10	5	100	48	0
Essay	67	5	5	33	100	5
Interview	5	5	5	5	5	100

Table 8. Accuracy (%) using SMO for predicting person between genres after training on one genre using L+F features

Table 8 displays the accuracies when the L+F feature set of single genre is used for training a model tested on one genre. This generally suggests the contribution of each genre when all are used in training. When the training and testing sets are the same, 100% accuracy is achieved. Examining this chart, the highest accuracies are achieved when training and test sets are textual. Excluding models trained and tested on the same genre, the average accuracy for training and testing within written genres is 36% while the average accuracy for training and testing within spoken genres is 17%. Even lower are average accuracies of the models trained on spoken and tested on textual genres (9%) and the models trained on textual and tested on spoken genres (6%). This indicates that the accuracies that feature the same mode (textual or spoken) in training and testing tend to be higher.

Of particular interest here is further examination of the surprising results of testing on email with the L+F feature set. Of these tests, a model trained on blogs achieved the highest score, perhaps due to a greater stylistic similarity to email than the other genres. This is also the highest score in the chart apart from cases where train and test genres were the same. Training on chat and essay genres shows some improvement over the baseline, but models trained with the two spoken genres do not rise above baseline accuracy when tested on the textual email genre.

5.3 Predict Communicant in One Topic Given Information on Five Topics

This set of experiments was designed to determine if there was no training data provided for a certain topic, yet there were samples of communication for an individual across genres for other topics, could an author be determined?

SMO	**HOLD OUT (TEST TOPIC)**					
Features	**Avg**	Ch	Gay	Iraq	Mar	Pri	Sex
L+F	**87**	81	95	86	95	100	67
L+F+T	**65**	76	71	86	29	62	67
L+F+E	**87**	81	95	86	95	95	67
L+S	**94**	95	95	81	100	100	95

Table 9. Person identification accuracy (%) training with SMO on 5 topics and testing on 1. Avg = Average over all topics: Ch=Catholic Church, Gay=Gay Marriage, Iraq=Iraq War, Mar=Marijuana Legalization, Pri=Privacy Rights, Sex=Sex Discrimination

Again a total of 126 “documents” were used: for each topic, 21 samples were constructed, each the concatenation of all text produced by an individual on that topic, across all genres. One topic was withheld and 105 documents (on the other 5 topics) were used for training. Table 9 shows that overall the L+S feature set performed better than either the L+F or L+F+T sets. The most noticeable differences are the drops in the accuracy when the five topic words are added, particularly on the topics of marijuana and privacy rights. For L+F+T, if “marijuana” is withheld from the topic word features when the marijuana topic is the test set, the accuracy rises to 90%. Similarly, if “school” is withheld from the topic word features when the privacy rights topic is the test set, the accuracy rises to 100%. This indicates the topic words are detrimental to determining the communicant, and this appears to be supported by the lack of an accuracy drop in the testing on the Iraq and sexual discrimination topics, both of which featured the fewest uses of the five topic words. That the results rise when using the L+S features shows that more features that are independent of the topic tend to help distinguish the person (as only the Iraq set experienced a small drop using these features in training and testing, while the others either increased or remained the same). The similarity here of the results using L+F features when compared to L+F+E is likely due to the small number of emoticons observed in the corpus (16 total examples).

5.4 Predict Communicant in a Speech Genre Given Information on the Other

One interesting experiment used one speech genre for training, and the other speech genre for testing. The results (Table 10) show that the additional stop words (S compared to F) make a positive difference in both sets. We hypothesize that the increased performance of training with discussion data and testing on interview data is due to the larger amount of training data available in discussions. We will test this in future work.

Train	Test	L+F	L+S
Inter	Disc	5	19
Disc	Inter	29	48

Table 10. Person identification accuracy (%) training and testing SMO on spoken genres

5.5 Predict Authorship in a Textual Genre Given Information on Speech Genres

Train	Test	L+F	L+S
Disc+Inter	Blog	19	24
Disc+Inter	Chat	5	14
Disc+Inter	Email	5	10
Disc+Inter	Essay	10	29

Table 11. Person identification accuracy (%) training SMO on spoken genres and testing on textual genres

Table 11 shows the results of training on speech data only and predicting the author of the text genre. Again, the speech genres alone do not do well at determining the individual author of the text genre. The best score was 29% for essays.

5.6 Predict Authorship in a Textual Genre Given Information on Other Textual Genres

Table 12 shows the results of training on text data only and predicting authorship for one of the four text genres. Recognizing the authors in chat is the most difficult, which is not surprising since the blogs, essays and emails are more similar to each other than the chat genre, which uses abbreviations and more informal language as well as being immediately interactive.

Train	Test	L+F	L+S
C+E+S	Blog	76	86
B+E+S	Chat	10	19
B+C+S	Email	90	81
B+C+E	Essay	90	86

Table 12. Person identification accuracy (%) training and testing SMO on textual genres

5.7 Predict Communicant in a Speech Genre Given Information on Textual Genres

Training on text and classifying speech-based samples by author showed poor results. Similar to the results for speech genres, using the text genres alone to determine the individual in the speech genre results in a maximum score of 29% for the interview genre (Table 13).

Train	Test	L+F	L+S
B+C+E+S	Discussion	14	23
B+C+E+S	Interview	14	29

Table 13. Person identification accuracy (%) training SMO on textual genres and testing on speech genres

5.8 Error Analysis

Results for different training and test sets vary considerably. A key factor in determining which sets can successfully be used to train other sets seems to be the mode, that is, whether or not a set is textual or spoken, as the lowest accuracies tend to be found between genres of different modes. This suggests that how people write and how they speak may be somewhat distinct.

Typically, more data samples in the training tends to increase the accuracy of the tests, but more features does not guarantee the same result. An examination of the feature sets revealed further explanations for this apart from any inherent difficulties in recognizing authors between sets. For many tests, there is a tendency for the same person to be chosen for classification, indicating a bias to that person in the training data. This is typically caused by features that have mostly, but not all, zero values in training samples, but have many non-zero values in testing. The most striking examples of this are described in 5.3, where the removal of certain topic-related features was found to dramatically increase the accruacy. Targetted removal of other features that have the same biasing effect could increase accuracy.

While Weka normalizes the incoming features for SMO, it was also discovered that a simple initial normalization of the feature sets by dividing by the maximum or standardization by subtracting the mean and dividing by the standard deviation of the feature sets could increase the accuracy across the different tests.

6 Conclusion

In this paper, we have described a novel unique corpus consisting of samples of communication of 21 individuals in six genres across six topics as well as experiments conducted to identify a person's samples within the corpus. We have shown that we can identify individuals with reasonably high accuracy for several cases: (1) when we have samples of their communication across genres (71%), (2) when we have samples of their communication in specific genres other than the one being tested (81%), and (3) when they are communicating on a new topic (94%).

For predicting a person's communication in one text genre using other text genres only, we were able to achieve a good accuracy for all genres (above 76%) except chat. We believe this is because chat, due to its "real-time communication" nature is quite different from the other text genres of emails, essays and blogs.

Identifying a person in one speech genre after training with the other speech genre had lower accuracies (less than 48%). Since these results differed significantly, we hypothesize this is due to the amount of data available for training – a hypothesis we plan to test in the future.

Future plans also include further investigation of some of the suprising results mentioned in this paper as well investigation of stop word lists particular to communicative genres. We also plan to investigate if it is easier to identify those participants who have produced more data (higher total word count) as well as perform a systematic study the effects of the number of words gathered on person identificaton.

In addition, we plan to investigate the efficacy of using other features besides those available in LIWC, stopwords and emoticons in person identification. These include spelling errors, readability measures, complexity measures, suffixes, and content analysis measures.

References

Naomi S. Baron. 2003. Why email looks like speech. In J. Aitchison and D. M. Lewis, editors, *New Media Language*. Routledge, London, UK.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press, Cambridge, UK.

Leo Breiman. 2001. Random forests. Technical Report for Version 3, University of California, Berkeley, CA.

Rafael A. Calvo, Jae-Moon Lee, and Xiaobo Li. 2004. Managing content with automatic document classification. *Journal of Digital Information*, 5(2).

Corinna Cortes and Vladimir Vapnik. 1995. Support vector networks. *Machine Learning*, 20(3):273-297.

Nikolas Coupland, Justine Coupland, Howard Giles, and Karen L. Henwood. 1988. Accommodating the elderly: Invoking and extending a theory, *Language in Society*, 17(1):1-41.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press, Cambridge, UK.

Olivier de Vel, Alison Anderson, Malcolm Corney, George Mohay. 2001. Mining e-mail content for author identification forensics, In *SIGMOD: Special Section on Data Mining for Intrusion Detection and Threat Analysis*.

Jade Goldstein-Stewart, Gary Ciany, and Jaime Carbonell. 2007. Genre identification and goal-focused summarization, In *Proceedings of the ACM 16th Conference on Information and Knowledge Management (CIKM) 2007*, pages 889-892.

Jade Goldstein-Stewart, Kerri A. Goodwin, Roberta E. Sabin, and Ransom K. Winder. 2008. Creating and using a correlated corpora to glean communicative commonalities. In *LREC2008 Proceedings*, Marrakech, Morocco.

Susan Herring. 2001. Gender and power in online communication. Center for Social Informatics, Working Paper, WP-01-05.

Susan Herring. 1996. Two variants of an electronic message schema. In Susan Herring, editor, *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. John Benjamins, Amsterdam, pages 81-106.

Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations*. 5(2):179-184.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computation*. 17(4):401-412.

Ivan Krsul and Eugene H. Spafford. 1997. Authorship analysis: Identifying the author of a program. *Computers and Security* 16(3):233-257.

Shlomo Levitan and Shlomo Argamon. 2006. Fixing the federalist: correcting results and evaluating editions for automated attribution. In *Digital Humanities*, pages 323-328, Paris.

LIWC, Linguistic Inquiry and Word Count. http://www.liwc.net/

David Madigan, Alexander Genkin, David Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. *Proc. of the Meeting of the Classification Society of North America.*

Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, and Danielle S. McNamara. 2006. Analyzing writing styles with Coh-Metrix, In *Proceedings of AI Research Society International Conference (FLAIRS)*, pages 764-769.

Niamh McCombe. 2002. Methods of author identification, Final Year Project, Trinity College, Ireland.

Thomas C. Mendenhall. 1887. The characteristic curves of composition. *Science*, 9(214):237-249.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Boston.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Lawrence Erlbaum Associates, Mahwah, NJ.

John C. Platt. 1998. Using sparseness and analytic QP to speed training of support vector machines. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge, Mass.

Roberta E. Sabin, Kerri A. Goodwin, Jade Goldstein-Stewart, and Joseph A. Pereira. 2008. Gender differences across correlated corpora: preliminary results. *FLAIRS Conference 2008*, Florida, pages 207-212.

Marina Santini. 2007. *Automatic Identification of Genre in Web Pages*. Ph.D., thesis, University of Brighton, Brighton, UK.

Michael Shepherd and Carolyn Watters. 1999. The functionality attribute of cybergenres. In *Proceedings of the 32nd Hawaii International Conf. on System Sciences (HICSS1999)*, Maui, HI.

Rob Thomson and Tamar Murachver. 2001. Predicting gender from electronic discourse. *British Journal of Social Psychology*. 40(2):193-208.

Ian Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition).* Morgan Kaufmann, San Francisco, CA.

Simeon J. Yates. 1996. Oral and written linguistic aspects of computer conferencing: a corpus based study. In Susan Herring, editor, *Computer-mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives.* John Benjamins, Amsterdam, pages 29-46.