Identifying Emotions, Intentions, and Attitudes in Text Using a Game with a Purpose

Lisa Pearl
Department of Cognitive Sciences
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697, USA
lpearl@uci.edu

Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine
3151 Social Science Plaza
Irvine, CA 92697, USA
msteyver@uci.edu

Abstract

Subtle social information is available in text such as a speaker's emotional state, intentions, and attitude, but current information extraction systems are unable to extract this information at the level that humans can. We describe a methodology for creating databases of messages annotated with social information based on interactive games between humans trying to generate and interpret messages for a number of different social information types. We then present some classification results achieved by using a small-scale database created with this methodology.

1 Introduction

A focus of much information extraction research has been identifying surface-level semantic content (e.g., identifying who did what to whom when). In recent years, research on sentiment analysis and opinion mining has recognized that more subtle information can be communicated via linguistic features in the text (see Pang and Lee (2008) for a review), such as whether text (e.g., a movie review) is positive or negative (Turney 2002, Pang, Lee, and Vaithyanathan 2002, Dave, Lawrence, and Pennock 2003, Wiebe et al. 2004, Kennedy and Inkpen 2006, Agarwal, Biadsy, and Mckeown 2009, Greene and Resnik 2009, among many others). However, other subtle information available in text, such as a speaker's emotional states (e.g., anger, embarrassment), intentions (e.g., persuasion, deception), and attitudes (e.g., disbelief, confidence), has not been explored as much, though there has been some work in detecting emotion (e.g., Subasic and Huettner 2001, Alm, Roth, and Sproat 2005, Nicolov et al. 2006, Abbasi 2007) and detecting deception (e.g., Annolli, Balconi, and Ciceri 2002, Zhou et al. 2004, Gupta and Skillicorn 2006, Zhou and Sung 2008). This latter kind of social information is useful for identifying the "tone" of a message, i.e., for understanding the underlying intention behind a message's creation, and also for predicting how this message will be interpreted by humans reading it.

A technical barrier to extracting this kind of social information is that there are currently no large-scale text databases that are annotated with social information from which to learn the relevant linguistic cues. That is, there are few examples of social information "ground truth" - text annotated with human perceptions of the social information contained within the text. Given the success of sentiment analysis, we believe this social information could also be retrievable once the relevant linguistic cues are identified.

One way to create the necessary annotated data is to draw from computational social science (Lazer et al. 2009), and make use of human-based computation (Kosurokoff 2001, von Ahn 2006, among others) since humans are used to transmitting social information through language. In this paper, we describe a methodology for creating this kind of database, and then present the results from a small-scale database created using this methodology[1]. In addition, we show one example of us-

[1]The database can be obtained by downloading it from http://www.socsci.uci.edu/~lpearl/CoLaLab/projects.html or contacting Lisa Pearl at lpearl@uci.edu.

ing this database by training a Sparse Multinomial Logistic Regression classifier (Krishnapuram et al. 2005) on these data.

2 Reliable databases of social information

2.1 The need for databases

In general, reliable databases are required to develop reliable machine learning algorithms. Unfortunately, very few databases annotated with social information exist, and the few that do are small in size. A recent addition to the Linguistic Data Consortium demonstrates this: The Language Understanding Annotation Corpus (LUAC) by Diab et al. (2009) includes text annotated with *committed belief*, which "distinguishes between statements which assert belief or opinion, those which contain speculation, and statements which convey fact or otherwise do not convey belief." This is meant to aid in determining which beliefs can be ascribed to a communicator and how strongly the communicator holds those beliefs. Nonetheless, this is still a small sample of the possible social information contained in text. Moreover, the LUAC contains only about 9000 words across two languages (6949 English, 2183 Arabic), which is small compared to the corpora generally available for natural language processing (e.g., the English Gigaword corpus (Graff 2003) contains 1756504 words).

Another tack taken by researchers has been to use open-source data that are likely to demonstrate certain social information by happenstance, e.g., online gaming forums with games that happen to involve the intent to deceive (e.g., Zhou and Sung 2008: Mafia game forums). While these data sets are larger in size, they do not have the breadth of coverage in terms of what social information they can capture because, by nature, the games only explicitly involve one kind of social information (e.g., intentions: deception); other social information cannot reliably be attributed to the text. In general, real world data sets present the problem of ground truth, i.e., knowing for certain which emotions, intentions, and attitudes are conveyed by a particular message.

However, people can often detect social information conveyed through text (perhaps parsing it as the "tone" of the message). For example, consider the following message: "*Come on...you have to buy this.*" From only the text itself, we can readily infer that the speaker intends to persuade the listener. Human-based computation can leverage this ability from the population, and use it to construct a reliable database of social information. Interestingly, groups of humans are sometimes capable of producing much more precise and reliable results than any particular individual in the group. For example, Steyvers et al. (2009) has shown that such "wisdom of crowds" phenomena occur in many knowledge domains, including human memory, problem solving, and prediction. In addition, Snow et al. (2008) have demonstrated that a relatively small number of non-expert annotations in natural language tasks can achieve the same results as expert annotation.

2.2 Games with a purpose

One approach is to use a *game with a purpose* (GWAP) (von Ahn and Dabbish 2004, von Ahn 2006, von Ahn, Kedia, and Blum 2006) that is designed to encourage people to provide the information needed in the database. GWAPs are currently being used to accumulate information about many things that humans find easy to identify (see *http://www.gwap.com/gwap/* for several examples), such as objects in images (von Ahn and Dabbish 2004), the musical style of songs, impressions of sights and sounds in videos, and common sense relationships between concepts (von Ahn, Kedia, and Blum 2006). In addition, as the collected data comes from and is vetted by a large number of participants, we can gauge which messages are reliable examples of particular social information and which are confusing examples.

2.3 A GWAP for social information in text

We designed a GWAP to create a database of messages annotated with social information, where unpaid participants provide knowledge about the social information in text. The GWAP encourages participants to both generate messages that reflect specific social information and to label messages created by other participants as reflecting specific social information. Participants are given points for every message they create that is correctly labeled by another participant, and for every message created by another participant that they correctly label.

Message generators were instructed to generate a

message expressing some particular social information type (such as *persuading*), and were allowed to use a displayed picture as context to guide their message, so they would not need to rely completely on their own imaginations. All context pictures used in our GWAP were meant to be generic enough that they could be a basis for a message expressing a variety of social information types. Context pictures were randomly assigned when participants were asked to generate messages; this meant that, for example, a picture could be used to generate a persuasive message and be used again later to generate a deceptive message. Generators were also warned not to use "taboo" words that would make the social information too easy to guess [2], but were encouraged to express the social information as clearly as possible. The generator was told that if another participant perceived the correct social information type from the message, the generator would be rewarded with game points.

Message annotators were instructed to guess which social information type was being expressed by the displayed message. They were also shown the image the generator used as context for the message, and were rewarded with points for successful detection of the intended social information.

As an example of the GWAP in action, one participant might generate the message "*Won't you consider joining our campaign? It's for a good cause.*" for the social information of *persuading*; a different participant would see this message and might label it as an example of *persuading*. A participant can only label a message with one social information type (e.g., a participant could not choose both *persuading* and *formal* for the same message).[3]

With enough game players, many messages are created that clearly reflect different social information. Without any of the participants necessarily having expert knowledge or training, we expect that the cumulative knowledge to be quite reliable (for example, see Steyvers et al. (2009) and work by von Ahn (von Ahn and Dabbish 2004, von Ahn 2006, von Ahn, Kedia, and Blum 2006) for other successful cases involving the "wisdom of the crowds", and Snow et al. (2008) for non-expert annotation in natural language tasks such as affect recognition). Because the same text can be evaluated by many different people, this can reduce the effect of idiosyncratic responses from a few individuals.

An advantage of this kind of database is that many different kinds of social information can be generated and labeled by the participants so that the database contains examples of many different kinds of social information in text, even if only a single label is given to a particular message (perhaps expressing that message's most obvious social information from the perspective of the labeler). We can gauge how clearly a message reflects social information by how often it is labeled by others as reflecting that social information. In addition, by the very nature of the GWAP, we can also assess which social information is easily confused by humans, e.g., politeness with embarrassment, or confidence with deception. This can aid the development of models that extract social information and could also identify messages likely to be ambiguous to humans.

2.4 A GWAP study

Below we report data from an offline GWAP that involves eight types of social information indicative of several social aspects that we thought would be of interest: politeness (indicates emotional state, attitude), rudeness (indicates emotional state, attitude), embarrassment (indicates emotional state), formality (indicates attitude), persuading (indicates intent), deception (indicates intent), confidence (indicates emotional state, attitude), and disbelief (indicates attitude). Fifty eight English-speaking adults participated in the GWAP, consisting of a mix of undergraduate students, graduate students, the authors, friends of the students, and friends of the authors, in order to simulate the varied mix of participants in an online GWAP. The undergraduate students were compensated with course credit. Together, these 58 participants created 1176 messages and made 3198 annotations. Note that a participant would label

[2] Taboo words were chosen as morphological variants of the social information type description. For example, *persuade*, *persuades*, *persuaded*, and *persuading* were considered taboo words for "persuading". Future versions of the GWAP could allow the taboo word list to be influenced by which words are often associated with a particular social information type.

[3] We note that this is a restriction that might be relaxed in future versions of the GWAP. For instance, participants might decide whether a message expresses a social information type or not from their perspective, so the task is more like binary classification for each social information type.

more messages than that participant would be asked to generate, and more than one participant would label the same message (though no participant would label a message that s/he created, nor would any participant label the same message more than once). Participants were encouraged to play the GWAP multiple times if they were inclined, to simulate the experience of playing a favorite game. There was no limit on message length, though most participants tended to keep messages fairly brief. Some sample messages (with the participants' own spelling and punctuation) that were correctly and incorrectly labeled are shown in Table 1.

Social Information Generated *Labeled*	**Message**
deception *deception*	"Oh yeah...your hair looks really great like that...yup, I love it...it, uh, really suits you..."
embarrassment *embarrassment*	"Oh... we're not dating. I would never date him... he's like a brother to me.."
disbelief *disbelief*	"Are you and him really friends?"
rudeness *persuading*	"James, Bree doesn't like you. She never did and never will!"
deception *persuading*	"I wasn't going to take anything from your storeroom, I swear! Really, I won't try to get inside again!'
politeness *deception*	"Your orange hair matches your sweater nicely"

Table 1: Sample messages from the offline GWAP.

The GWAP as currently designed allows us to gauge two interesting aspects of social information transmission via text. First, we can assess our non-expert participants' performance. Second, we can assess the messages themselves.

For the participants, we can gauge their accuracy as message generators by measuring how often a message they created was successfully perceived as expressing the intended social information type (that is, their "expressive accuracy"). On average, message generators were able to generate reliable messages 56% of the time. Figure 1 displays the expressive accuracy of participants, while also showing how many messages participants generated. Most participants created less than 30 messages, and were accurate more than half the time.

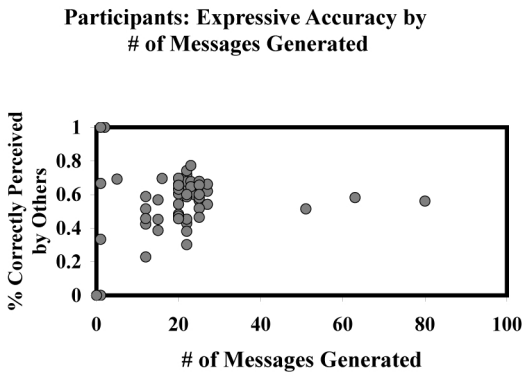

Figure 1: Expressive accuracy of GWAP participants.

At the same time, we can also gauge the accuracy of the participants as non-expert annotators by measuring how often a participant perceived the intended social information (that is, their "perceptive accuracy"). On average, annotators were able to perceive the intended social information 58% of the time. Figure 2 displays the perceptive accuracy, while also showing how many messages participants annotated. Most participants annotated around 20 messages or between 80 and 100 messages and were accurate more than half the time. Average inter-annotator agreement was 0.44, calculated using Fleiss' Kappa (Fleiss 1971), suggesting moderate agreement.

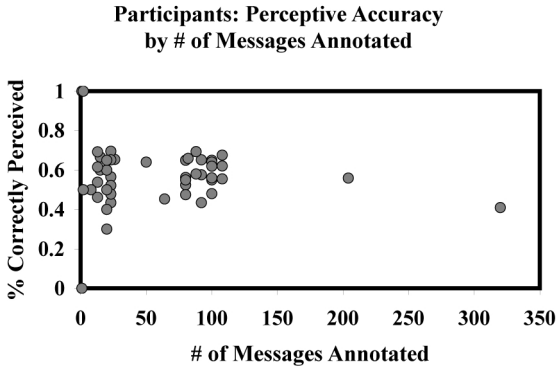

Figure 2: Perceptive accuracy of GWAP participants.

Turning to the messages, we can gauge how often messages were able to successfully express a particular social information type, and how often they were confused as expressing some other type. Table 2 shows a confusion matrix of social information de-

rived from this database.

	deception	politeness	rudeness	embarrassment	confidence	disbelief	formality	persuading
deception	**.37**	.07	.10	.03	.09	.10	.04	.20
politeness	.05	**.53**	.05	.02	.03	.01	.20	.10
rudeness	.04	.01	**.78**	.02	.04	.04	.03	.03
embarrassment	.07	.09	.05	**.56**	.02	.13	.05	.03
confidence	.04	.04	.03	.01	**.67**	.05	.02	.13
disbelief	.10	.05	.05	.04	.07	**.62**	.02	.06
formality	.02	.34	.04	.02	.06	.03	**.39**	.10
persuading	.09	.06	.03	.01	.12	.03	.04	**.61**

Table 2: Confusion matrix for the human participants. The rows represent the intended social information for a message while the columns represent the labeled social information, averaged over messages and participants.

The matrix shows the likelihood that a message will be labeled as expressing specific social information (in a column), given that it has been generated with specific social information in mind (in a row), averaged over messages and participants. In other words, we show the probability distribution $p(labeled|generated)$. The diagonal probabilities indicate how often a message's social information was correctly labeled for each social information type; this shows how often social information transmission was successful. Messages were perceived correctly by human participants about 57% of the time. More particular observations about the data in Table 2 are that people are more likely to correctly identify a message expressing rudeness ($p = .78$) and confidence ($p = .67$) and less likely to correctly identify a message expressing deception ($p = .37$) or formality ($p = .39$). Also, we can see that a deceptive message can often be mistaken for a persuading message ($p = 0.20$), a formal message mistaken for a polite message ($p = 0.34$), a message expressing disbelief mistaken for a message expressing deception ($p = .10$), and a persuading message mistaken for a deceptive message ($p = .09$) or confidence ($p = .12$), among other observations. Some of these may be expected, e.g., confusing confidence with persuading since someone who is trying to persuade will likely be confident about the topic, or formality with politeness since many formal expressions are used to indicate politeness (e.g., *"if you would be so kind"*). Others may be unexpected a priori, such as mistaking disbelief for deception.

2.5 Human reliability and message reliability

Given that humans were believed to be good at identifying social information in text, the low perceptive accuracy rates for participants and low annotation accuracy rates for messages may seem unexpected. However, we believe it indicates that some messages are better than others at expressing social information in a way obvious to humans. That is, messages confusing to human participants (e.g., the lower three examples in Table 1, as well as the confusing messages represented by the probabilities in Table 2) would be consistently mislabeled.

It may be that some messages are created such that many annotators agree with each other, but they all perceive a social information type other than the one intended.[4] In a similar vein, messages with low inter-annotator agreement may simply be poorly generated messages that should be removed from the database. To this end, we can assess how often majority annotator agreement correlates with perception of the message's intended social information type. Table 3 shows the confusion matrix for messages where over 50% of the annotators agreed with each other on which social information type was intended, and at least two annotators labeled the message. A total of 866 messages satisfied these criteria.

The confusion matrix, as before, shows the likelihood that a message will be labeled as expressing specific social information (in a column), given that it has been generated with specific social information in mind (in a row), averaged over messages and participants. The diagonal probabilities indicate how often a message's social information was correctly labeled for each social information type; this shows how often social information transmission was successful. The messages in this subset were perceived correctly by human participants about 71% of the time, a significant improvement over 57%. This demonstrates how even a modest pooling of non-expert opinion can significantly in-

[4]Messages consistently perceived as expressing a different social information type than intended should perhaps be considered as actually expressing that social information type rather than the intended one.

	deception	politeness	rudeness	embarrassment	confidence	disbelief	formality	persuading
deception	**.45**	.05	.10	.01	.07	.07	.03	.21
politeness	.03	**.71**	.03	.00	.01	.00	.13	.09
rudeness	.03	.00	**.92**	.00	.01	.02	.02	.00
embarrassment	.04	.08	.05	**.69**	.00	.11	.01	.02
confidence	.01	.04	.02	.01	**.82**	.01	.01	.09
disbelief	.05	.03	.02	.02	.05	**.82**	.00	.02
formality	.02	.34	.02	.01	.03	.03	**.46**	.10
persuading	.03	.05	.01	.00	.05	.03	.01	**.82**

Table 3: Confusion matrix for the human participants, where the majority of participants agreed on a message's intended social information and at least two participants labeled the message. The rows represent the intended social information for a message while the columns represent the labeled social information, averaged over messages and participants.

crease the accuracy of social information identification in text.

We can observe similar trends to what we saw in Table 2, in many cases sharpened from what they were previously. People are still more likely to identify messages expressing rudeness ($p = .92$) and confidence ($p = .82$), though they are also now more likely to accurately identify persuading ($p = .82$). The ability to identify politeness ($p = .71$) and embarrassment ($p = .69$) has also improved, though a polite message can still be mistaken for a formal message ($p = .13$). Formality ($p = .46$) and deception ($p = .45$) remain more difficult to identify, with formal messages mistaken for politeness ($p = .34$) and deceptive messages mistaken for persuading ($p = .21$) and rudeness (p=.10) [5]. Note, however, that messages of disbelief and persuading are now rarely mistaken for deceptive messages ($p = .05$ and $p = .03$, respectively). It is likely then that the confusions arising in this data set are more representative of the actual confusion humans encounter when perceiving these social information types.

Identifying messages likely to be misperceived by humans is useful for two reasons. First, from a cognitive standpoint, we can identify what features of those messages are the source of the confusion if the messages are consistently misperceived, which tells us what linguistic cues humans are (mistakenly) keying into. This then leads to designing better machine learning algorithms that do not key into those misleading cues. Second, this aids the design of cognitive systems that predict how a message is likely to be interpreted by humans, and can warn a human reader if a message's intent is likely to be interpreted incorrectly.

3 Training a classifier with the database

To demonstrate the utility of the created database for developing computational approaches to social information identification in text, we applied a Sparse Multinomial Logistic Regression (SMLR) classifier (Krishnapuram et al. 2005) to the the subset of messages where two or more participants labeled the message and more than 50% of the participants perceived the intended social information type. This subset consisted of 624 messages (these messages make up the messages in the diagonals of table 3). While we realize that there are many other machine learning techniques that could be used, we thought this classifier would be a reasonable one to start with to demonstrate the utility of the database. As a first pass measure for identifying diagnostic linguistic cues, we examined a number of fairly shallow features:

- unigrams, bigrams, and trigrams
- number of word types, word tokens, and sentences
- number of exclamation marks, questions marks, and punctuation marks
- average sentence and word length
- word type to word token ratio
- average word log frequency for words appearing more than once in the database

[5] We note that people's precision on deceptive messages was higher: 0.67. That is, when they labeled a message as deceptive, it was deceptive 2/3 of the time. However, the probabilities in Table 3 represent deceptive message recall, i.e., how well they were able to label all deceptive messages as deceptive.

The use of shallow linguistic features seemed a reasonable first investigation as prior research involving linguistic cues for identifying information in text has often used word-level cues. For example, positive and negative affect words (e.g., *excellent* vs. *poor*) have been used in sentiment analysis to summarize whether a document is positive or negative (Turney 2002, Pang, Lee, and Vaithyanathan 2002, among others). In deception detection research, informative word-level cues include counting first and third person pronoun usage (e.g., *me* vs. *them*) (Anolli, Balconi, and Ciceri 2002), and noting the number of "exception words" (e.g., *but*, *except*, *without*) (Gupta and Skillicorn 2006). In addition, informative shallow text properties have also been identified (Zhou et al. 2004), such as (a) number of verbs, words, noun phrases, and sentences, (b) average sentence and word length, and (c) word type to word token ratio.

The SMLR classifier model was trained to produce the label (one of eight) corresponding to the generated social information using all the text features as input. Using a 10-fold cross-validation procedure, the model was trained on 90% of the messages and tested on the remaining 10%. The sparse classifier favors a small number of features in the regression solution and sets the weight of a large fraction of features to zero. Some of the non-zero weights learned by the model for each social information type are listed below (though each type has other features that also had non-zero weights). Positive weights indicate positive correlations while negative weights indicate negative correlations. Cues that are negatively correlated are *italicized*. Bigrams and trigrams are indicated by + in between the relevant words (e.g., no+way). BEGIN and END indicate the beginning and the end of the message, respectively.

- **deception**: *#-of-question-marks (-0.5)*, actually (1.4), at+all (0.6), if (0.8), *me (-0.9)*, *my (-0.2)*, not (1.6), of+course (1.1), trying+to (0.8), you+END (1.0)
- **politeness**: BEGIN+please (2.1), help (2.1), may+i (1.2), nice (2.3), nicely+END (1.1), so+sorry (1.5), would+you+like (1.0)
- **rudeness**: annoying (1.2), *good (-1.1)*, *great (-0.6)*, hurry+up (1.0), loud (2.7), mean (0.9), *pretty (-2.0)*, ugly (1.6)
- **embarrassment**: BEGIN+oh (2.0), can't+believe (1.0), can't+believe+i (0.6), forgot (2.1), *good (-.9)*, my (2.0), oh (1.1)
- **confidence**: i+believe (2.1), i+know (2.4), positive (3.5), really+good (2.9), sure (3.3), the+best (2.5), *think (-0.8)*
- **disbelief**: #-of-question-marks (2.4), BEGIN+are (3.8), *like (-0.6)*, never (1.4), no+way (3.0), shocked (1.1), such+a (1.1)
- **formality**: #-of-exclamation-marks (-0.8), BEGIN+excuse (2.1), *don't (-0.8)*, miss (4.1), mr (3.7), please (2.7), sir (5.1), very+nice (1.0)
- **persuading**: BEGIN+if+you (2.3), buy (1.3), come (3.5), have+to (1.6), we+can (1.3), would+look (2.9), you+should (3.4)

Some of the feature-label correlations discovered by the model fit with our intuitions about the social information types. For example, deceptive messages are negatively correlated with some of the first person pronouns (*me*, *my*), in accordance with Anolli, Balconi, and Ciceri (2002)'s results. Several polite and formal words appear correlated with polite and formal messages respectively (*may+i*, *nice*, *so+sorry*, *would+you+like*; *BEGIN+excuse*, *miss*, *mr*, *sir*), and formal messages tend not to include exclamation points. Negative words tend to be associated with rude messages (*annoying*, *loud*, *mean*, *ugly*), while positive words tend to be associated with confident messages (*really+good*, *sure*, *the+best*). Messages conveying disbelief tend to have more question marks and contain expressions of surprise (*never*, *no+way*, *shocked*), and persuasive messages tend to contain coercive expressions (*come*, *have+to*, *you+should*). As this is a relatively small data set, these cues are unlikely to be definitive – however, it is promising for the approach as a whole that the classifier can identify these cues using fairly shallow linguistic analyses.

We can also examine the classifier's ability to label messages, given the features it has deemed diagnostic for each social information type (i.e., those features it gave non-zero weight). For each message

in the dataset, the classifier predicted what the intended social information type was. A correct prediction for a message's type matches the intended type for the message. A confusion matrix for the classifier based on the messages from the 624 message test set is shown in Table 4. Overall, the classifier was able to correctly label 59% of the messages. This is 12% less than humans were able to correctly label, but far better than chance performance (13%) and the performance of a simple algorithm that chooses the most frequent data type in the training set (17%).

The classifier shows some patterns similar to the human participants: (1) deception and formality are harder to detect than other social information types, (2) confidence and embarrassment are easier to detect than other social information types, and (3) formality is often mistaken for politeness ($p = .26$). However, some differences from the human participants are that deception is often mistaken for rudeness ($p = .19$) and politeness is often confused with rudeness and embarrassment, in addition to formality (all $p = .12$).

	deception	politeness	rudeness	embarrassment	confidence	disbelief	formality	persuading
deception	**.36**	.08	.19	.08	.08	.09	.06	.08
politeness	.05	**.49**	.12	.12	.05	.01	.12	.05
rudeness	.06	.06	**.63**	.04	.07	.07	.01	.07
embarrassment	.02	.01	.11	**.76**	.06	.03	.01	.00
confidence	.06	.01	.04	.08	**.68**	.02	.03	.08
disbelief	.08	.03	.08	.02	.09	**.56**	.02	.12
formality	.00	.26	.06	.03	.00	.06	**.43**	.15
persuading	.05	.06	.09	.03	.11	.03	.02	**.61**

Table 4: Confusion matrix for the machine learning classifier. The rows represent the intended social information for a message while the columns represent the labeled social information.

As the classifier's behavior was similar to human behavior in some cases, and the classifier used only these shallow linguistic features to make its decision, this suggests that humans may be keying into some of these shallower linguistic features when deciding a message's social information content. Given this, a classifier trained on such linguistic features may be able to predict which messages are likely to be ambiguous to humans.

4 Conclusion

We have described a methodology using GWAPs to create a database containing messages labeled with social information such as emotions, intentions, and attitudes, which can be valuable to the information extraction research community. Having implemented this methodology on a small scale, we discovered that non-expert annotators were able to identify the social information of interest fairly well when their collective perceptions were combined. However, we also noted that certain social information types are easily confusable by humans. We also used the database created by the GWAP to investigate shallow linguistic cues to social information in text and attempt to automatically label messages as expressing particular social information. The fact that the social information types we used in our GWAP can be identified automatically with some success suggests that these social information types are useful to pursue, though of course there are many other emotional states, attitudes, and intentions that could be explored in future work. In addition, other classifiers, particularly those using deeper-level properties like phrase structure, may be able to identify more subtle cues to social information in text. We also foresee extending the GWAP methodology to create large-scale databases both in English and in other languages in order to continue fostering the development of computational approaches to social information identification.

Acknowledgments

This paper has benefited from discussion and advice from Padhraic Smyth, Pierre Isabelle, and three anonymous reviewers. In addition, this work is supported by NSF grant BCS-0843896 to LP and CORCL grant MI 14B-2009-2010 to LP and MS.

References

Abbasi, A. 2007. Affect intensity analysis of dark web forums. *Proceedings of Intelligence and Security Informatics (ISI)*: 282-288.

Agarwal, A., Biadsy, F., and Mckeown, K. 2009. Contextual Phrase-Level Polarity Analysis using Lexical

Affect Scoring and Syntactic N-grams. *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Green: 24-32.

Alm, C. O., Roth, D., and Sproat, R. 2005. Emotions from text: Machine learning for text-based emotion prediction. *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

Anolli, L., Balconi, M., and Ciceri, R. 2002. Deceptive Miscommunication Theory (DeMiT): A New Model for the Analysis of Deceptive Communication. In Anolli, L., Ciceri, R. and Rivs, G. (eds)., *Say not to say: new perspectives on miscommunication*. IOS Press: 73-100.

Dave, K., Lawrence, S., and Pennock, D. 2003. Mining the peanut gallery: Opinion extraction and semantic classication of product reviews, *Proceedings of WWW*: 519-528.

Diab, M., Dorr, B., Levin, L., Mitamura, T., Passonneau, R., Rambow, O., and Ramshaw, L. 2009. Language Understanding Annotation Corpus. LDC, Philadelphia.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378382.

Graff, D. 2003. English Gigaword. Linguistic Data Consortium, Philadelphia.

Greene, S. and Resnik, P. 2009. More than Words: Syntactic Packaging and Implicit Sentiment. *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, Colorado: 503-511.

Gupta, S. and Skillicorn, D. 2006. Improving a Textual Deception Detection Model, *Proceedings of the 2006 conference of the Center for Advanced Studies on Collaborative research*. Toronto, Canada.

Kennedy, A. and Inkpen, D. 2006. Sentiment classication of movie reviews using contextual valence shifters. Computational Intelligence, 22: 110-125.

Kosorukoff, A. 2001. Human-based Genetic Algorithm. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2001: 3464-3469.

Krishnapuram, B., Figueiredo, M., Carin, L., and Hartemink, A. 2005. Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27: 957-968.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabsi, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Val Alstyne, M. 2009. Computational Social Science, *Science*, 323: 721-723.

Nicolov, N., Salvetti, F., Liberman, M., and Martin, J. H. (eds.) 2006. *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW).* AAAI Press.

Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2(1-2): 1-135.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 79-86.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. 2008. Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 254-263.

Steyvers, M., Lee, M., Miller, B., and Hemmer, P. 2009. The Wisdom of Crowds in the Recollection of Order Information. In J. Lafferty, C. Williams (Eds.) *Advances in Neural Information Processing Systems*, 23, MIT Press.

Subasic, P. and Huettner A. 2001. A?ect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9: 483-496.

Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the Association for Computational Linguistics (ACL)*: 417-424.

von Ahn, L. 2006. Games With A Purpose. *IEEE Computer Magazine*, June 2006: 96-98.

von Ahn, L. and Dabbish, L. 2004. Labeling Images with a Computer Game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Association for Computing Machinery, New York, 2004)*: 319-326.

von Ahn, L., Kedia, M. and Blum, M. 2006. Verbosity: A Game for Collecting Common-Sense Facts, *In Proceedings of the SIGCHI conference on Human Factors in computing systems*, Montral, Quebec, Canada.

Wiebe, J.M., Wilson, T., Bruce, R., Bell, M., and Martin, M. 2004. Learning subjective language. *Computational Linguistics*, 30: 277-308.

Zhou, L., Burgoon, J., Nunamaker, J., and Twitchell, D. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13: 81-106.

Zhou, L. and Sung, Y. 2008. Cues to deception in online Chinese groups. *Proceedings of the 41st Annual Hawaii international Conference on System Sciences*, 146. Washington, DC: IEEE Computer Society.