Linguistic Cues for Deception Detection in Online Mafia Forums

Meredith Fay

May 2009

Lisa Pearl, Faculty Advisor

Department of Cognitive Sciences

Abstract

The detection of dishonesty in computer-mediated communication such as emails, blogs, and online forums is of utmost importance for many governmental, financial, and educational institutions. Several studies have attempted to identify linguistic cues that reliably distinguish deceivers from truth-tellers, but have typically used highly artificial experimental constraints over a very limited period of time. We designed a pilot program to extract and analyze content from a popular online game that includes deception as part of the game premise, and which provides a more robust and ecologically valid data set. While the statistical power of the initial results are limited due to small sample size, this pilot study finds that none of the linguistic features identified in past research are actually reliably useful for detecting deception in this, more realistic, computer-mediated communication data.

1. Introduction

Institutions such as the United States Military and law enforcement agencies have long understood the immense practical value of uncovering dishonesty in interpersonal and organizational communication. The imagination of entertainment media has also been captured by the possibility of reaching truth in deceptive situations using astute observation, as shown in popular programs like *Lie to Me* and *The Mentalist.* Years of research have accumulated pertaining to the validity of various deception "red flags" based on oral, visual, and physiological communication cues [2, 5, 3, 8]. As global interactions come to rely increasingly heavily upon electronic avenues, however, most, if not all, auditory and physical signals become unavailable. What remains is based on the language text available, rather than any non-verbal signals. This domain of communication is referred to as text-based, computer-mediated communication (CMC). This type of communication primarily encompasses the world of e-mails, instant messages, text messages, blogs, and faxes - all communication options increasing in popularity due to their efficiency. In fact, as of 2005 there was an estimated worldwide traffic of 135 billion e-mail messages per day [6]. For the purposes of this discussion, we will focus on a subset of CMC known as TA-CMC, or text-based, asynchronous, computer-mediated communication, which excludes more contemporaneous modes of communication, such as instant messaging.

Deception analysis in the current age of TA-CMC would necessarily be based exclusively upon variables present in the textual information alone. For this purpose, linguistics-based cues (LBCs), which rely solely on the language present in the text, are particularly useful. Examples of such cues include the frequency of passive voice ("he was selected") or of third person reference ("they"). Cues derived solely from language do not presuppose any contextual information about the communication and are thus less subjective and more straightforward to extract for analysis than any other, higher-level context features (such as the more complex construct of sarcasm) in the transmission [11]. While historically, experts have been trained and contracted to recognize dissimulation in others, such training requires a considerable investment of time and money. Thus, calls have intensified for an equally accurate but more cost-efficient automated detection system that could reduce the workload of lie detection experts or even eliminate them altogether.

Previous research on automated TA-CMC deception detection thus far has been limited and dominated by a very select group of researchers (most notably Zhou and colleagues: [9-15]). In general, past experiments have obtained data sets for analysis by pairing a small number of undergraduate students together, and assigning one dyad member to the role of "deceiver," and the other to the role of "truth-teller" [9-13]. Over a short period, ranging from three minutes to three days depending on the study, the pairs work together on an assigned problem solving activity. Typically, the activity is some variant of a military-style decision-making exercise such as the "Desert Survival Problem," in which they must create and come to agreement upon either a prioritized list of items to salvage or a wilderness escape route in the aftermath a hypothetical plane crash that leaves the pair stranded in the desert. The deceiver is instructed to argue for a set of priorities contrary to those they actually hold. The messages between the pair are sent via an adapted e-mail system, and the aggregate body of these interactions is then analyzed for a variety of linguistic features in order to assess the ability of each feature to reliably distinguish deceivers from truth-tellers.

While interesting results have come from these studies, the aforementioned experimental setups do have some disadvantages associated with them. The generalizability and ecological validity of their results are handicapped for several reasons: (1) the sample populations were comprised solely of undergraduates (so the population was not very diverse), (2) there were

artificial time and/or post quantity restrictions imposed as part of the design scheme, (3) subjects lacked motivation to succeed in their deception, and (4) subjects lacked experience with the problem or game at hand. These problems call into question the robustness of the linguistic cues identified in these studies. Put simply, in a more realistic environment, are these cues successful at identifying deception?

The current study attempts to address these issues by utilizing a different dataset that is likely to be a more realistic example of lying and truth-telling interactions. For our experiment, we selected a completed game from a popular online "Mafia" web forum, MafiaScum.net [7]. The forum allows individuals from all over the world to organize such games, in which a neutral moderator secretly assigns a small group of players at random to be "mafia members," while the rest of the players are assigned to be "innocent townspeople." Each group's goal is to eliminate, by vote, members of the other group (e.g. mafia members want to eliminate all the townspeople). This can only be achieved by accurately guessing which player belongs to which group and convincing the other players to concur in order to form a majority vote to remove the suspect player. Since the "mafia members" are in the minority and are being sought out for removal by the "innocent townspeople" majority, the "mafia's" goal is to convince the rest of the players that they, too, are innocent. Thus, they can be naturally classified as the deceiver group (D). The other players must convince the group of their genuine innocence, and so are classified as truth-tellers (T). At the end of the game, the group with the most members still alive is declared the winner, and each player's role is posted for all to see. There are different versions of Mafia that co-exist in the forum, but the rules of each game are publicly posted at the beginning of each group, allowing researchers to check that no major variations in task structure have occurred.

The use of a Mafia game to assess the deception detection power of certain LBCs confers several advantages over past experiments. First, analysis of the Mafia forum allows for more generalizable results, since players were not restricted by age or educational level and hailed from several different English-speaking countries – a more representative sample than simply using undergraduates from one university. Second, the selected Mafia game had a total of over 2300 posts that were fairly lengthy, written over the course of six-and-a-half months. This provides a much richer per-player data set than past experiments, which typically only allow a limited number of posts per person per day over the course of a few days. While Zhou, Burgoon, and Twitchell (2003) found that no set of LBCs reliably discriminated between D and T consistently across time, the greatly extended duration of Mafia games could perhaps eliminate data irregularities due to the limited time span examined in that study [12]. Third, the selected Mafia game was an invitational open only to experienced and historically successful players. Thus, most of the players already had experience interacting with each other from other games, were intimately familiar with the background and rules, and took the game very seriously, since ineptitude would tarnish their reputations in future games. Finally, the game was initiated independently by the players well before any experimental analysis was considered and was thus not altered by subjects second-guessing the purpose of the study or reacting to new, artificial conditions, giving the study an enormous advantage in terms of ecological validity.

The particular Mafia game examined here was selected for the aforementioned reason of its nature as an invitational, because there were relatively few additional roles to complicate analysis, and because it was one of the lengthier games available. While only one game was analyzed, thereby greatly limiting the results' statistical power, the improvements in terms of generalizability, quantity of available data, subjects' investment in their performance, and ecological validity for this particular data set merited a limited yet thorough pilot study. One other contemporaneous study has been conducted on a Chinese version of a similar Mafia forum, but the enormous linguistic differences, those players' apparently extreme lack of sophistication,

and the questionable statistical methods used therein (see discussion section for details) merited an alternate examination [13].

2. LBCs in the current study

The LBCs selected for analysis were based in large part upon the features examined by Twitchell and Zhou, et al. [10, 11, 13, 14], as we wished to examine their robustness across different data sets – particularly data sets that were more realistic, for reasons described in the previous section. Selecting particular LBCs required careful consideration for a variety of reasons: (1) there were substantial alterations in the definitions of the relevant LBCs between each of the previous studies, (2) a handful of previously-supported LBCs were abandoned in subsequent research for no clear reason, and (3) the categorization of features was apparently arbitrary and shifted frequently. As this was primarily an exploratory study, we opted to err on the side of inclusiveness and chose to analyze a feature as long as it had been proposed in at least one past experiment and its definition and operationalization made intuitive sense. In this way, some feature selection was admittedly subjective, but was judged to be a logical improvement upon past research.

It should be noted that prior experiments included umbrella categories containing a number of LBCs, but these categories have been discarded in the present study. For example, Zhou defined the category of "Uncertainty" to include the LBCs of modal verb quantity (e.g., "should," "can"), number of modifiers (defined as any adverb or adjective), number of other references (e.g., "they"), and the feature of "uncertainty" itself [11]. This category assignment, however, seems problematic. First, the number of modifiers and number of other references does not necessarily indicate uncertainty (as the use of descriptive language is not inherently a stalling tactic and references to others may merely reflect a communicative necessity of the task at hand). Second, the feature of "uncertainty" within the larger category of "Uncertainty" seems a nonsensical duplication. Due to the aforementioned category complications, we chose to abandon the larger LBC umbrella categories altogether. This seemed reasonable as they added nothing to the coherence of the features and seemed only to distract from the discriminatory power of individual features by confounding several - not necessarily inherently related - factors in an attempt to arbitrarily aggregate their results. For the LBCs ultimately selected in this study, we elected to adopt the definitions used in past studies wherever possible, making a few minor alterations when necessary in order to make the operationalization relevant to the Mafia scenario and to ensure the intuitive logic of the features. The list of selected features and definitions appears in Table A below.

Table A. LBCs used in the current study.

Feature	**Definition/Examples**
Number of words °	Word = a string of characters surrounded on either side by a space
Number of letters per word °	Letter = A single alphabetic character
Number of verbs °	Verb = word that expresses an act, occurrence, or mode of being; usually in grammatical center of the predicate*
Number of noun phrases °	Noun phrase = phrase formed by a noun, modifiers, and determiners *

Number of sentences °	Sentence = word, clause, or phrase (or group thereof) forming a syntactic unit that expresses an assertion, question, command, wish, exclamation, or performance of action; usually begins with a capital letter and concludes with appropriate end punctuation *
Number of passive voice constructions°	Passive voice construction = verb form used when the subject is being acted upon *
Number of generalizing terms °	*newbie*, *town*, *scum*
Number of self-reference terms °	*I, me, my, mine, myself*
Number of group reference terms °	*we, our, ours, us, ourselves*
Number of modal verbs°	*can, shall, will, must, may, dare, could, should, have to, might, ought to, might not, couldn't, wouldn't, won't, mustn't, shouldn't*
Number of other reference terms °	*he, she, it, they, them, those, him, her, them, his, hers, theirs*
Number of modifiers °	Modifier = adjective or adverb
Number of hesitation/ uncertain words °	Modal verbs, *maybe, possibly, I don't know, I'm not sure, …, um, uh*
Content words ratio	Total content words (e.g. *mafia*) divided by total words *
Function words per sentence	Total function words divided by total sentences *
Lexical diversity	Total unique words divided by total words *
Average number of clauses per sentence	Total clauses divided by total sentences *
Average sentence length	Total words divided by total sentences *
Average word length	Total characters divided by total words *
Pausality	Total punctuation marks divided by total sentences *
Number of quotes from other posts °	Instances of quotes from past posts
Number of hostility/ aggression terms °	Profanity, demands, name-calling
Number of interjections°	Interjection = a brief utterance that primarily conveys emotional context rather than concrete content; items tagged "UH" by the parser (those interjections that indicated hesitation as defined above were counted towards both features of hesitation and interjection)
Number of unique words °	Number of word types, or the first occurrence of each word over the entirety of a particular player's messages (i.e. the first occurrence of the word "the" counts as one unique word, all subsequent instances of "the" are ignored)

Notes

* Indicates a definition taken from Zhou 2004 [11]

° Indicates a feature whose number of occurrences is averaged per player's total number of posts. Each feature is initially analyzed in terms of its total number of occurrences per player. Since certain players are eliminated earlier in the game than others and thus do not have the same opportunity to continue posting, the number of feature occurrences is averaged per post to modulate the elimination effect.

3. Mafia corpus details

The particular Mafia game used contained 2,315 total posts and 198,611 total words from all the players combined (including administrative posts containing the rules of play and vote counts) [7]. Nine players were assigned the role of "innocent townspeople," three more to the role of "mafia members," two to "cops," one to "vigilante," and one served as the moderator, who assigned the roles, ensured fair play, and made announcements and vote counts about who was to be eliminated each day. Two of the original "innocent townspeople" were removed early in the game due to insufficient participation and were replaced by two new, outside players who assumed and carried on the roles of the two former players. In these cases, the four were treated as four separate truth-tellers (rather than combining each original and their replacement as one person since they shared the same function within the game), in order to avoid combining and confounding the individuals' unique writing characteristics in one aggregate file. It should also be noted that, for the final analysis, the roles of "cop," "vigilante," and "moderator" were all grouped into the T role, as their functions in the game did not involve deceit. Thus, there were thirteen players in the T group and three in the D group.

4. Data analysis

First, the LBCs of interest were assessed over the corpus. Several Perl scripts were written to organize and analyze the players' verbal output in the selected game. All 2,315 posts were compiled into one, large html-coded file. Then, the html code was removed to leave only the players' names, post numbers, the posts themselves, and any embedded quotations or icons (since they contained possibly important contextual and emotional references). Each player was assigned an anonymous identification tag and a file was created for each containing solely the content from that particular player. Then, each feature that was easily encompassed by only a small collection of terms (such as self reference, which only includes the words "I," "me," "my," "mine," and "myself") or a readily-identifiable structural feature (such as the number of words) was counted for each player using Perl. After counting the instances of embedded quotations from other players, the quotations themselves were removed from the posts to prevent double-counting of words and misattribution of the original quoted content to the player who merely referred to it. Emoticons and formatting (i.e. bolded text) were replaced by one-word labels describing the effect of the original, such as inserting "[happy]" in place of a smile emoticon, or "[bold]" to indicate emphasis. It was these labels, rather than their original counterparts, that were counted when necessary. For those features that encompassed large grammatical categories (like verbs), each player's file was analyzed using the Charniak parser to tag the part of speech for each word [1]. The files with the part-of-speech tags were then run through another Perl script, which counted the occurrences of the relevant tags.

Following this, we compared the LBCs for truth-tellers and deceivers. Because of the inconsistent selection of and results for LBCs in past studies, we simply chose to conduct a two-tailed test for each of the features listed in Table A, with the hypothesis that each of the features would discriminate between the D and T roles without surmising the direction of the relationship (i.e., whether Ds would use a feature more of less frequently than Ts). A student t-test was used, due to the small group sizes, to assess the probability that the differences in LBC usage between the two groups was due to chance alone.

5. Results

The feature counts for each player are shown in Table B, as well as the p-value of a student t-test comparing the LBC display frequencies for the D versus T groups.

Table B.

	Total words	**Unique words**	**Sentences**	**Letters**	**Verbs**	**Noun phrases**
Average usage frequency in **T** posts	83.901	24.338	8.108	346.263	11.258	26.458
Average usage frequency in **D** posts	82.417	19.742	6.851	343.161	10.838	26.893
P-value	**0.973**	**0.542**	**0.672**	**0.987**	**0.944**	**0.976**
	Modal verbs	**Passive voice**	**Modifiers**	**Pausality**	**Other reference**	**Generalizing term**
Average usage frequency in **T** posts	1.453	0.509	6.480	1.398	2.366	1.587
Average usage frequency in **D** posts	1.460	0.382	6.549	1.553	1.925	1.974
P-value	**0.990**	**0.633**	**0.984**	**0.250**	**0.748**	**0.711**
	Self reference	**Group reference**	**Quotations from past posts**	**Hesitation**	**Lexical diversity**	**Content words**
Average usage frequency in **T** posts	3.945	0.439	0.527	1.541	0.310	6.615
Average usage frequency in **D** posts	3.995	0.275	0.441	1.547	0.282	7.200
P-value	**0.971**	**0.371**	**0.821**	**0.991**	**0.762**	**0.557**
	Function words	**Interjections**	**Aggression/ hostility**	**Average word length**	**Average sentence length**	**Average # clauses per sentence**
Average usage frequency in **T** posts	3.672	0.273	0.099	4.103	10.288	1.921
Average usage frequency in **D** posts	4.026	0.260	0.060	4.101	11.227	1.669
P-value	**0.508**	**0.875**	**0.350**	**0.990**	**0.537**	**0.350**

6. Discussion

As is quite clearly illustrated in Table B above, none of the LBCs tested even remotely approached statistical significance (i.e., $p<0.05$). While some of this may be attributed to certain experimental design issues discussed below, a handful of the LBCs' p-values were so high as to cast grave doubt upon their discriminatory usefulness even given a much larger sample size. For example, average word length ($p=0.990$), self reference ($p=0.971$), modal verbs ($p=0.990$), hesitation ($p=0.991$), modifiers ($p=0.984$), letters ($p=0.987$), and noun phrases ($p=0.976$) so closely approach the level of 100% coincidence that their usefulness in any similar context seems unlikely. The most promising features were found to be pausality ($p=0.250$), average clauses per sentence ($p=0.350$), number of aggression/hostility terms ($p=0.350$) and number of group reference terms ($p=0.371$). Even these were nowhere close to a satisfactory level of significance, but they seem good candidates for further investigation with a larger pool of players.

Because this was designed simply to be a pilot study, the sample size was quite small and allowed for only minimal statistical power. The D and T averages in Table B above disguise the extreme variability of feature displays between players in the same groups. While the duration of the game and sheer volume of posts seems to suggest the feature counts for each player were legitimately representative of the individuals' characteristic writing patterns, the relatively small number of players sampled overall and their high within-group variability suggests that a much larger group of players is necessary to stabilize group averages and clarify which players' data should be treated as outliers. In other words, it is likely that different results could be obtained from a larger data sample.

There is a rather surprising discrepancy between our results and those of Zhou and Sung's 2008 study of Chinese mafia groups [13]. In that study, the researchers found that deceivers displayed fewer words and sentences, shorter words and sentences, and more third person references than did truth-tellers. This contrast with the results here may perhaps be explained by the different statistical tools applied to each data set. Interestingly, Zhou used a paired-sample t-test to obtain p-values, even though there is no evidence that the experimental design actually included a paired-sample setup. The use of this type of t-test automatically confers greater statistical power (and thus a lower p-value, suggesting a higher probability that the between-group differences were not due to mere chance) because it presupposes the comparison of a particular individual's performance in one experimental condition to that same individual's performance in the other experimental condition (which would eliminate potentially confounding individual differences) [4]. However, from our understanding, Zhou's mafia role assignment was nearly identical to that found in our mafia forum and thus is not an appropriate candidate for a paired-sample t-test; this may partially explain why our p-values for the same features were much higher. Other possible reasons for the discrepancies could include cultural and language differences, and differences in the players' familiarity and expertise within the context of the game.

With a few minor adjustments, the Perl scripts and parser program created for and used in this experiment could easily be applied to quickly gather very large amounts of data for analysis from multiple games in the MafiaScum forum. With some additional modifications, it would even be possible to extract data for individual players from games in which they were assigned to the D role and compare it to the same player's verbal output when assigned to the T role in other games. This would allow for a within-subject, paired-sample comparison, greatly increasing the statistical power of the results by eliminating likely-confounding between-subjects differences in vocabulary, personality, and writing style.

Given a more exhaustive study with a larger sample of players, we would expect most of the features approaching statistical significance to concur with those found in Zhou's Chinese mafia results. If such a study still yielded no significance for any of the features, it may imply that the LBCs under consideration are actually not useful deception detection tools in more real-world situations. Perhaps at that point it would be of use to examine more complex, higher-level LBCs, such as those that indicate certain emotional or contextual content. This deeper type of feature would be much more difficult to automate, but in the event the previously-tested features continue to fail, they may be another option for analysis.

7. Conclusions

The diverse and high-stakes need for deception detection continues to grow as communication becomes increasingly dependent upon computer-mediated text channels, and automation is key to making the process of deception detection affordable and widely accessible. The need to develop a set of linguistics-based cues that reliably predict deception by highly motivated, skilled individuals in long-term, realistic, reciprocal communication requires the analysis of the proposed LBCs set forth in prior research to be analyzed in a context more readily generalizable to that mentioned above. The MafiaScum.net forum allows for large-scale paired-sampling of TA-CMC data from a more realistic scenario that does not require and is not contaminated by artificial, experimental manipulation.

This experiment was intended as a pilot study to create a largely automated system for rapidly extracting and analyzing the Mafia data, which relies on LBCs. While our initial results do not support the high level of discriminatory robustness found for any of the LBCs in past research, a larger, more exhaustive study using the tools already developed is necessary before those LBCs can be conclusively dismissed. Still, the results cast doubt upon the diagnostic power of the previously-examined features in their current conception.

8. References

[1] Charniak, E. (2006). Charniak parser [computer software]. Brown University. Available at ftp://ftp.cs.brown.edu/pub/nlparser/

[2] Illinois. (1963). *Detection of deception examiner*. Springfield: [Dept. of Registration and Education].

[3] Library of Congress. *Literature and research support on deception, detection, and polygraph research*. Washington, D.C.: Federal Research Division.

[4] McDonald, J.H. (2008). *Handbook of biological statistics*. Sparky House Publishing, Baltimore, Maryland. 110-114.

[5] McKeever, L. (2006). Online plagiarism detection services--saviour or scourge? *Assessment & Evaluation in Higher Education.* 31 (2), 155-165.

[6] Moskalyuk, Alex. (2009, March 3). IT Facts 2005 [1.2 bln active e-mail accounts worldwide in 2005]. Message posted to http://blogs.zdnet.com/ITFacts/?p=9962

[7] MrStoofer, Moderator. (2008, October 23). Mafia60: Face to face. Messages posted to

http://mafiascum.net/forum/viewtopic.php?t=4642

[8] Riggio, R.E., & R.S. Feldman. Claremont Symposium on Applied Social Psychology, (2005). *Applications of nonverbal communication*. Mahwah, N.J.: L. Erlbaum Associates.

[9] Twitchell, D. P. (2005). *Automated analysis techniques for online conversations with application in deception detection*. Tucson, Arizona: University of Arizona. http://etd.library.arizona.edu/etd/GetFileServlet?file=file:///data1/pdf/etd/azu%5Fetd%5F1111%5F1%5Fm.pdf&type=application/pdf.

[10] Twitchell, D. P., Nunamaker, J. F., & Burgoon, J. K. (2004). Using speech act profiling for deception detection. *Lecture Notes in Computer Science.* 3073, 403-410.

[11] Zhou, L. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation.* 13, 81-106.

[12] Zhou, L., Burgoon, J. K., & Twitchell, D. P. (2003). A longitudinal analysis of language behavior of deception in e-mail. *Lecture Notes in Computer Science.* (2665), 102-110.

[13] Zhou, L., & 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. DOI=http://dx.doi.org/10.1109/HICSS2008.109.

[14] Zhou, L., Yongmei, S., & Dongsong, Z. (2008). A statistical language modeling approach to online deception detection. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.* 20 (8), 1077-1081.