A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication

LINA ZHOU, JUDEE K. BURGOON, DOUGLAS P. TWITCHELL,
TIANTIAN QIN, AND JAY F. NUNAMAKER JR.

LINA ZHOU is an Assistant Professor in the Department of Information Systems at the University of Maryland, Baltimore County. Dr. Zhou's research interest includes text mining, ontology, deception detection, machine learning, information retrieval, and machine translation. Her research has been published or will be appearing in journals such as the *Journal of the American Society for Information Science and Technology, Communications of the ACM, Group Decision and Negotiation, Journal of Computer Processing of Oriental Languages, IEEE Transactions on Systems, Man, and Cybernetics, Information Resource Management Journal,* among other journals.

JUDEE K. BURGOON is Professor of Communication, Professor of Family Studies and Human Development, and Director of Human Communication Research for the Center for the Management of Information at the University of Arizona. She has authored or coauthored seven books and monographs and nearly 200 articles, chapters, and reviews related to nonverbal, interpersonal, group, and computer-mediated communication. Her current research on communication via new technologies and on deception detection has been funded by the Department of Defense and Department of Homeland Security, among other agencies. An elected Fellow of the International Communication Association, she is a recipient of its Aubrey Fisher Mentorship Award and the National Communication Association's Distinguished Scholar Award.

DOUGLAS P. TWITCHELL is a doctoral student in Management Information Systems at the University of Arizona. His research interests include text mining, conversational analysis and profiling, machine learning, and natural language processing. Prior to his doctoral work, he received his B.S. in business administration with an emphasis in information systems and M.S. in Information Systems Management from Brigham Young University. He has also worked for two years in IT consulting with Arthur Andersen LLP and Information Technology International, Inc.

TIANTIAN QIN is a doctoral student in Management Information Systems at the University of Arizona. She is currently conducting research on automatic deception detection via data mining techniques. Her research interest includes general data mining techniques, deception detection, and Bayesian network pattern recognition.

JAY F. NUNAMAKER JR. is Regents and Soldwedel Professor of MIS, Computer Science, and Communication and Director of the Center for the Management of Information at the University of Arizona, Tucson. Dr. Nunamaker received the LEO Award from the Association of Information Systems at ICIS in Barcelona, Spain, December

Journal of Management Information Systems / Spring 2004, Vol. 20, No. 4, pp. 139–165.

0742–1222 / 2004 $9.50 + 0.00.

2002. This award is given for a lifetime of exceptional achievement in information systems. He was elected a fellow of the Association of Information Systems in 2000. He has over 40 years of experience in examining, analyzing, designing, testing, evaluating, and developing information systems. He has served as a test engineer at the Shippingport Atomic Power facility, as a member of the ISDOS team at the University of Michigan, and as a member of the faculty at Purdue University prior to joining the faculty at the University of Arizona in 1974. His research on group support systems addresses behavioral as well as engineering issues and focuses on theory as well as implementation. Dr. Nunamaker received his Ph.D. in systems engineering and operations research from Case Institute of Technology, an M.S. and B.S. in engineering from the University of Pittsburgh, and a B.S. from Carnegie Mellon University. He has been a licensed professional engineer since 1965.

ABSTRACT: The increased chance of deception in computer-mediated communication and the potential risk of taking action based on deceptive information calls for automatic detection of deception. To achieve the ultimate goal of automatic prediction of deception, we selected four common classification methods and empirically compared their performance in predicting deception. The deception and truth data were collected during two experimental studies. The results suggest that all of the four methods were promising for predicting deception with cues to deception. Among them, neural networks exhibited consistent performance and were robust across test settings. The comparisons also highlighted the importance of selecting important input variables and removing noise in an attempt to enhance the performance of classification methods. The selected cues offer both methodological and theoretical contributions to the body of deception and information systems research.

KEY WORDS AND PHRASES: classification methods, deception, deception detection, linguistic cues.

DECEPTION IN HUMAN COMMUNICATION occurs when information senders attempt to create a false impression in receivers. Most people have experienced deception of one form or another from outright lies and fabrications to little "white" lies [16, 22]. Deception may result in receivers taking actions unfavorable to themselves on behalf of the information senders. If the actions are critical to a person's life, an organization's survival, or even a nation's stability, neglecting deception may lead to immeasurable losses. Therefore, the need to improve deception detection is of longstanding concern and strong practical relevance to research communities, practitioners, and government agencies.

Extensive research has been conducted on cues that can be used to detect deception [17]. A close review of the literature reveals, however, that most studies mainly focus on deception in rich media channels (e.g., face-to-face interactions), with some exceptions in the area of textual and computer-mediated communication [45]. Increasing reliance on computers in support of human-to-human communication poses at least two challenges to deception research. The first is information overload: the pace of information growth has far exceeded our capability to process it. The second is humans' poor deception detection capability, which is seldom better than chance [1, 7, 17].

One obvious solution to address both issues is to automate the deception detection process. To achieve the automatic prediction of deception, we envision a three-step course of action: (1) identify significant cues to deception, (2) automatically derive the cues from various media, and (3) build classification models for predicting deception from new messages. Previous research has identified promising indicators of deceit based on language features [3, 45]. For example, deceptive messages have been found to include higher informality and expressivity, and lower lexical diversity and complexity. Focusing on language behaviors, rather than specific content, has the advantage that indicators derived from language behaviors may be relatively independent of context and are more amenable to simple parsing approaches (e.g., the very challenging task of semantic parsing can be bypassed). Moreover, deceivers may have control over the content of their messages, but deceptive intent may still be "betrayed" through one's language use. Some progress has been made in identifying and automatically deriving deception indicators from text by integrating findings and methods from multiple relevant disciplines, including natural language processing, linguistics [38], and stylistic research [5, 27]. The majority of deception studies have relied on human coders manually rating behavioral indicators of deception [1, 7, 37], and these have stopped short of prediction, instead simply reporting verbal and nonverbal cues associated with truth or deception.

The current investigation addresses the third objective of building classification models for predicting deceit by evaluating four popular classification approaches for their ability to discriminate truthful from deceptive text-based messages. Data from two empirical studies were subjected to classification with discriminant analysis, logit regression, neural networks, and decision tree analysis. Because a single message from a deceiver may present different predictive power than a series of messages from the same deceiver, the classification approaches were evaluated with both messages and individuals as the unit of analysis. The first consisted of individual messages. The second comprised all messages from individual subjects. These comparisons also shed light on the features of deceptive messages.

Discriminant analysis has been applied in detecting deception with some limited success (e.g., [15]). Compared with discriminant analysis, logistic regression places less stringent requirements on the underlying data features and may therefore be well suited to the task of predicting deception. In contrast with statistical methods, machine learning methods are not tied to the characteristics of the population distribution; however, they are usually more computationally intensive. Many common machine learning approaches, such as decision trees and neural networks, can automatically build classification models from the existing data and then predict the outcome for the new data. Neural networks have been found to provide better prediction than discriminant analysis in some applications [42, 43]. There has also been an initial attempt at applying decision trees in grouping messages into deceptive and truthful classes [3]. With the scattered reports using one classification method for deception detection, there is a great need to compare the performance of difference classification methods in the context of deception detection and identify the most appropriate ones. Hence, this paper extends prior work on cues to deception by investigating four classification methods—discriminant analysis, logistic regression, decision trees, and neural networks—for their predictive power in discriminating truth from deception. We begin by reviewing theories and prior empirical findings related to cues to deception, based on which cues were identified for inclusion into the classification models.

Related Work

Theories in Support of Detecting Deception from Lean Media

COMMUNICATION AND OTHER RELEVANT RESEARCH has long studied how and why individuals deceive. Within this broad area of investigation, there have been several theories proposed for rich media. Media richness is measured on a continuum and is determined by four criteria: feedback (asking questions or making corrections), multiple cues (transmitting voice inflection, body language, and so on), language variety (range of meaning that can be encoded in language symbols), and personal focus (transmitting feelings and emotions) [13]. According to these criteria, computer-mediated communication (CMC) is generally leaner than face-to-face communication, especially if it is text-based. Despite the lack of theories specifically directed toward text-based CMC, some theories that have been frequently used to guide deception research in media-rich channels may be extended to less rich channels.

Interpersonal deception theory (IDT) attempts to explain deception from an interpersonal and conversational perspective, rather than an individual and psychological perspective [2]. IDT posits that within the context and relationship of the sender and receivers of deception, deceivers will display strategic modifications of behavior in response to a receiver's suspicions, but may also display nonstrategic (inadvertent) behavior, or leakage cues, indicating that deception is occurring. The theory is not confined to any one modality, nor does it focus solely on physiological or nonverbal indicators. Therefore, it is also applicable to leaner mediated channels.

Channel expansion theory [8] expands media richness theory [13] by including the experience senders or receivers have with a channel or medium, the topic of the communication, the organizational context, and the other parties in the communication. The more experience the senders and the receivers have with each of these domains, the richer they find the media they are using. The heightened perceived richness may have one of two results: those more experienced with text-based CMC will find it richer and, in the process of using it, transmit more deception cues, or they may have a greater ability to strategically hide possible deception cues. People hold expectations about the discourse of others and assume that others will satisfy the interaction demands of exchanging messages that are sufficiently complete, truthful, clear, and relevant to the current topics [21]. To be successful in evading detection, deceivers have to make their messages appear complete, truthful, clear, and relevant. Therefore, some aspects of deceivers' linguistic patterns should be similar to those displayed in face-to-face situations. Others may take advantage of the unique properties of written, rather than spoken, language, and of the response delays in text communication that afford them more time to plan, monitor, and edit what they say.

Focusing on the dynamics in interpersonal interaction, interaction adaptation theory [6] describes and predicts patterns of reciprocity and compensation between communicators. The theory holds that there may be a fair amount of reciprocal behavior display (including language behavior) between deceivers and truth-tellers. This makes it easy for deceivers to get away with their deception because it is difficult to distinguish deceivers from truth-tellers within a given pair of communication. Conversely, it may be easier to differentiate deceivers from truth-tellers who are involved in different interactions. Interpersonal adaptation theory also implies that deception is likely to be a continuous event that unfolds over time [39]. Deceivers may manage to embed their deception intentions in other truthful messages to increase their chances of success. Therefore, aggregating all the messages from an information sender may provide a holistic view and, consequently, a better indication of deception than examining individual messages.

Linguistic Cues to Deception

The above-mentioned theories are concerned with how deception occurs, but do not supply guidelines on how to detect it. In addition, textual messages lack facial expressions, gestures, and conventions of body posture and distance, so the text itself is the only source for inferring personal opinions and attitudes and verifying message credibility. Two disciplines that have produced methods for detecting deceit from text are criminal justice and linguistics. In criminal investigations, the validity of statements made by suspects has been evaluated using criteria-based content analysis [35], reality monitoring [24], and scientific content analysis [34]. Each of these techniques includes a set of criteria against which susceptible statements is compared. The presence or absence of a criterion (e.g., contextual embedding) affects the judged truthfulness of an account. Such criteria are usually quite complex and mostly used by trained experts. Verbal immediacy research [29] offers detailed criteria for scoring the degree to which language creates psychological closeness or distance (non-immediacy). Non-immediate language is thought to signal both a distancing by deceivers from their messages, which reduces their accountability and responsibility for what they say, and an indication of negative feelings associated with the act of deceiving. Even though none of these systems for coding deceptive discourse was developed specifically for CMC, they provide the theoretical and evidentiary foundation for the investigation of deception in CMC.

Several prior investigations that have sought to determine the viability of using linguistics-based cues to distinguish truthful from deceptive messages have produced promising results [3, 45]. The majority of the classes of linguistic features studied received significant support. In particular, deceivers' messages were more expressive than their partners and they appeared more informal, as they had more typographical errors than truth-tellers. Deceivers' messages were less complex, which was manifested in less punctuation, fewer long sentences, and fewer syllables per word. Deceptive subjects displayed less diversity at both the lexical and content level than did truth-tellers. They also used more non-immediate and uncertain language in the form of fewer self-references and more modal verbs. However, the same cues could conceivably show opposite effects under deception in different modalities and task settings. A case in point: whereas deception has routinely been associated with shorter messages, in one of our experiments, the deceivers who performed a decision-making task with a partner via e-mail displayed higher quantity—of words, verbs, noun phrases, and sentences—and used more affective language [45]. Yet, in another experiment, during an interview conducted either via text-chat or audio-conferencing, deceivers tended to use briefer messages and less language referring to emotions and feelings than did truth-tellers [3]. This suggests that deception behavior is moderated by contextual factors. Therefore, the classification models built on top of cues identified in one type of deception context may not be suitable to other contexts. Consequently, the test data used in the current study to evaluate the classification models were collected under the same context as the training data.

Based on the aforementioned research results, we reevaluated the classification of linguistics features in our prior investigations (see [45]) and made several deletions, additions, and regroupings. Deletion was based on correlation analysis and the results of the prior studies. For cues that were highly correlated, all but one was removed. For example, words, sentences, and noun phrases from the quantity group were deleted because they were so highly correlated with verbs. Indicators that did not find support in any of the previous studies were also eliminated. As more resources supporting the automatic derivation of cues became available, several new cues were also added to the list. For example, Whissel et al.'s [38] affect dictionary made possible the inclusion of measures of pleasantness, activation, and imagery. If deception is indeed an unpleasant act, or if deceivers attempt to divert attention from themselves by becoming less expressive, their language should exhibit less pleasantness and be devoid of imagery or language conveying high arousal. The regrouping included merging two or more cues into one (e.g., individual references combined both first-person and second-person personal pronouns) and dividing one group into two subgroups (e.g., a new uncertainty grouping was created by extracting some cues from the original non-immediacy group). The resultant updated list of cues, as shown in Table 1, was used to create classification models.

Motivated by Whissel et al.'s affect dictionary [38] and the need to distinguish between positive and negative forms of affect, we generated six additional cues: positive pleasantness, positive activation, positive imagery, highly positive pleasantness, highly positive activation, and negative activation. Positive (or negative) affect cues are those with scalar values at least one standard deviation higher (or lower) than the mean of the corresponding affect, whereas those designated as highly positive (or negative) cues have values at least two standard deviations from the mean of the corresponding affect.

Table 1. Summary of Linguistic Constructs and Their Component Dependent Variables and Measures (Adapted from [45]).

Quantity

1. Verbs: words that are the grammatical center of a predicate and express act, occurrence, or mode of being.
2. Modifiers: adjectives and adverbs that describe words or makes the meaning of the words more specific.

Complexity

3. Average sentence length: $\frac{\text{total number of words}}{\text{total number of sentences}}$.
4. Average word length: $\frac{\text{total number of characters}}{\text{total number of words}}$.
5. Pausality: $\frac{\text{total number of punctuation marks}}{\text{total number of sentences}}$.

Uncertainty

6. Modal verbs: auxiliary verbs that are used with a verb of predication and express a modal modification.
7. Passive voice: a form of the verb used when the subject is being acted upon rather than doing something.

Non-immediacy

8. Individual references: singular first and second personal pronoun.
9. Group references: first personal plural pronoun.

Expressivity

10. Emotiveness: $\frac{\text{total number of adjectives} + \text{total number of adverbs}}{\text{total number of nouns} + \text{total number of verbs}}$.

Diversity

11. Content diversity: $\frac{\text{total number of different content words or terms}}{\text{total number of content words or terms}}$, where content words or terms primarily express lexical meaning.
12. Redundancy: $\frac{\text{total number of function words}}{\text{total number of sentences}}$, where function words express primarily grammatical relationships.

Informality

13. Typographical error ratio: $\frac{\text{total number of misspelled words}}{\text{total number of words}}$.

(continues)

Table 1. (Continued)

Specificity

14. Spatio-temporal information: information about locations of people or objects, or information about when the even happened or explicitly describes a sequence of events.
15. Perceptual information: indicates sensorial experiences, such as sounds, smells, physical sensations, and visual details

Affect

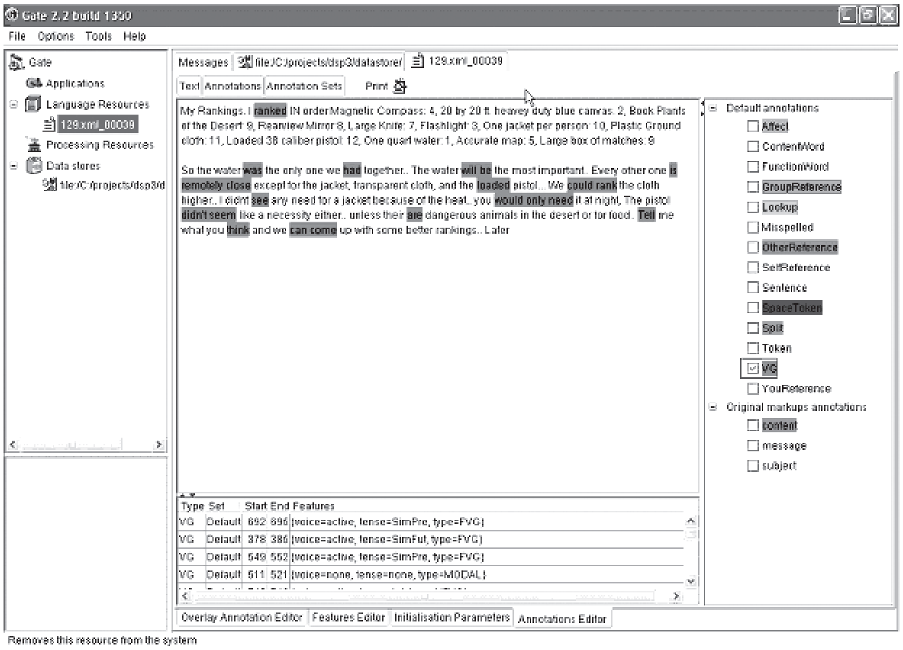
16. Affect: conscious subjective aspect of a emotion apart from bodily changes
17. Pleasantness: positive or negative feelings associated with the emotional state.
18. Activation: the dynamics of emotional state.
19. Imagery: words that provide a clear mental picture.

Automated Generation of Cues to Deception

In each of the studies reported in this paper, the messages were collected automatically via an online system. A Web-based e-mail messaging system automatically captured all of the textual data for the experimental task and stored it in a MS SQL 2000 database from which we retrieved the data.

To process the messages and extract each of the individual cues, we used GATE (general architecture for text engineering), created at the University of Sheffield. GATE 2.0 is a Java-based, object-oriented framework, architecture, and development environment for creating programs for analyzing, processing, or generating natural language [10]. It has been employed on many projects (see [28]) including, for example, creation of the American National Corpus (americannationalcorpus.org) and text summarization [26]. GATE is a component-based architecture based on two main components: *language resources* (LR) and *processing resources* (PR). LRs are data-only resources, such as single documents (or messages in our case), corpora, ontologies, and lexicons. PRs are programmatic or algorithmic resources that either use or process LRs, such as parsers, part-of-speech taggers, and cue derivation modules [11]. For example, to count all of the verbs in a document, one would create an LR that contains or represents the document. Next, two PRs, a part-of-speech tagger and a verb counter, are created. Last, an application, or *pipeline,* is created wherein the PRs are assigned to process the document LR and the number of verbs is counted. Figure 1 shows one of the messages from the experimental task in GATE with its verb phrases marked, which are counted to create the verb quantity cue.

To accomplish the goal of extracting and deriving cues from messages, we created a set of PRs, each of which extracted a cue or set of cues from the document. For example, we built a PR using the GATE-provided Java annotations processing engine (JAPE) [11] that recognized and counted group references such as *we, us,* and *ours.* A transformation-based part-of-speech tagger shipped with GATE was used to tag each word according to its part of speech. These tags were then used to extract cues, such as verb quantity. In addition, dictionaries were incorporated to enhance the abilities



My Rankings. I ranked IN orderMagnetic Compass: 4, 20 by 20 ft. heavey duty blue canvas: 2, Book Plants of the Desert: 9, Rearview Mirror 8, Large Knife: 7, Flashlight: 3, One jacket per person: 10, Plastic Ground cloth: 11, Loaded 38 caliber pistol: 12, One quart water: 1, Accurate map: 5, Large box of matches: 9

So the water was the only one we had together.. The water will be the most important.. Every other one is remotely close except for the jacket, transparent cloth, and the loaded pistol... We could rank the cloth higher.. I didnt see any need for a jacket because of the heat.. you would only need it at night, The pistol didn't seem like a necessity either.. unless their are dangerous animals in the desert or for food.. Tell me what you think and we can come up with some better rankings.. Later

Type	Set	Start	End	Features
VG	Default	692	696	{voice=active, tense=SimPre, type=FVG}
VG	Default	378	386	{voice=active, tense=SimFut, type=FVG}
VG	Default	549	552	{voice=active, tense=SimPre, type=FVG}
VG	Default	511	521	{voice=none, tense=none, type=MODAL}

Default annotations
☐ Affect
☐ ContentWord
☐ FunctionWord
☐ GroupReference
☐ Lookup
☐ Misspelled
☐ OtherReference
☐ SelfReference
☐ Sentence
☐ SpaceToken
☐ Split
☐ Token
☑ VG
☐ YouReference
Original markups annotations
☐ content
☐ message
☐ subject

Figure 1. Screen Shot of GATE with Verb Phrases Highlighted on a Message from the Experimental Task.

of GATE. Based on an affect dictionary [38], we were able derive the values of pleasantness and several other cues from messages.

All of the work described above lays the solid foundation for the automatic prediction of deception. It allows us to focus on investigating the performance of classification approaches in deception detection. The reviewed theories suggest observing deceptive behavior over the course of an interaction rather than from individual messages may yield the best predictive models. On the one hand, using individual messages as the unit of analysis has the potential advantage of providing greater precision with regard to deceivers' one-time behavior, but has the disadvantage that messages from the same individual are not independent of one another, thereby violating the assumptions of some statistical models. Moreover, people who produce numerous messages are overrepresented in the corpus and have greater influence on the resulting classification models. Furthermore, if some of the messages within the deception condition are truthful, this can introduce noise and lead to biased results. On the other hand, data that are aggregated over all the messages from a given person should benefit from being a more reliable summary of a given individual's linguistic behavior. But this approach yields smaller sample sizes with concomitant reductions in statistical power, and it averages estimates across variables in ways that may disguise the importance of a given cue. For example, if a sender uses high emotiveness in one message but not in another, the relationship between the predictor of emotiveness and the criterion measure of the message, even though not highly reliable in the given sample, may go undiscovered and not be investigated further. Therefore, as a secondary objective, we compared the classification methods on a new dimension: units of analysis.

Overview of Classification Approaches

In this section, we give a brief overview of the four classification methods under investigation: discriminant analysis, logistic regression, decision trees, and neural networks. Among many variants of each approach, we select one that is appropriate for the task under investigation. The mechanisms for selecting important cues (inputs or attributes) in each method are also discussed.

Discriminant Analysis

Linear discriminant analysis (simplified as discriminant analysis hereafter) [19] is a popular statistics method for the classification task. It constructs a linear function $\beta'X$ by maximizing the univariate between-groups variability relative to within-groups variability, which is stated below [42]:

$$\frac{\left(\beta'\mu_1 - \beta'\mu_2\right)^2}{\beta'\Sigma\beta}, \tag{1}$$

where Σ, μ_1, and μ_2 are the common covariance matrix and mean vectors of two groups π_1 and π_2, respectively.

Classification models derived from the above procedure are usually optimal in minimizing the expected cost of misclassification [36]. Fisher's linear discriminant function was developed under the assumption that different groups from the underlying populations have equal covariance structures [23]. In applied research, data are seldom compatible with the underlying assumption. Nonetheless, discriminant analysis has been found to be very robust to deviations from the above ideal condition when the data are substantially linear. Since it is still unclear whether deception data are "well behaved," discriminant analysis may work well in the current investigation.

When using discriminant analysis, it is common practice to remove independent variables that are not significant in a stepwise manner. We used both the forced entry and forward stepwise methods available in SPSS (statistical package for the social sciences) for the discriminant analysis.

Logistic Regression

In contrast to the linear relationships between the decision variable and the independent variables in linear regression models, logistic regression methods apply an additional logistic function and transform the linear probabilities into logit ones. The logit distribution constrains the estimated probabilities lying between zero and one, which can be represented as follows:

$$p = \frac{1}{1 + e^{-\alpha - \beta' x}}, \tag{2}$$

where x is a vector of input variables, β is a vector of coefficients, and α is a constant term. In our study, the logistic maximum likelihood procedure was utilized to estimate the coefficients of the classification model corresponding to each of the 19 variables listed in Table 1. The estimated coefficients β can be interpreted as the effect of the independent variable on the probability of the event (e.g., deception) divided by the probability of nonevent (e.g., truth) [9]. β values reflect the significance of input variables and can thereby help us identify important cues to deception from a candidate list. As with the discriminant analysis, we applied both the forced entry and forward stepwise methods in the logistic regression analysis using SPSS.

Decision Trees

The C4.5 algorithm [32] for inducing classification models of decision trees is an extension of the basic ID3 algorithm. It uses a greedy, top-down algorithm to produce the tree. If all data items belong to the same class, C4.5 keeps the decision tree as a leaf. Otherwise, it will recursively try to find the best attribute to split the data items into sub-leaves. Information gain after splitting relative to before splitting is the common criterion for selecting the best attribute. If no gains can be obtained by splitting, the set of mixed data items is made into a leaf, which is labeled as the most frequent class of data items in the set. From a decision tree, rules can be derived by following attributes and decision criteria on each path from the root to the leaf. An example of a decision tree created with data from the experimental task is shown in Figure 2. The values in boxes and associated arrows indicate the path taken by a single message to its classification.

Since the lower parts of a decision tree are often based on relatively few samples and can be inaccurate, it is desirable to have some way of pruning trees [30]. C4.5 can determine how deeply to grow a decision tree and reduce errors through pruning [40]. Pruning decision trees is a process of replacing a sub-tree with a branch or leaf node when the replacement can result in reduction of expected errors. A common problem with decision trees is overfitting the data, which may lower generalizability of the models. Pruning decision trees not only helps reduce errors but also avoid the overfitting problem [30]. In this study, we used the C4.5 revision 8 for decision trees (called J48) and the neural network implementation (discussed next) from WEKA [41].

Neural Networks

Neural networks are structured in layers, which generally consist of at least one input layer, one output layer, and a number of hidden layers existing in between. Each layer can have one or more nodes and there are weights to connect the nodes in different layers. Neural networks have a number of variations in terms of possible algorithms. We chose the most commonly and widely used back-propagation (BP) networks for

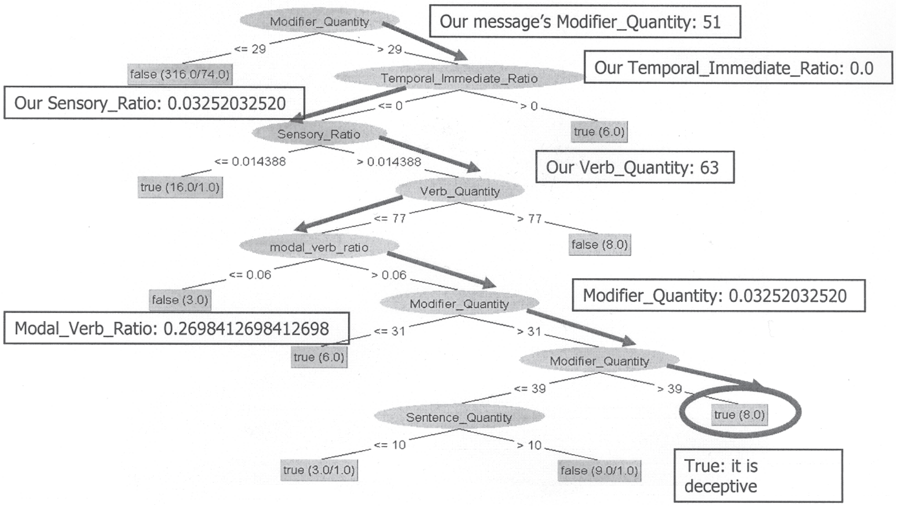


Figure 2. Example Decision Tree Created with Data from Experimental Task. Boxes and arrows show path taken by a single example message to its classification.

this study. The BP network is a supervised learning network, aiming to train the network to map input vectors to a desired output vector. During training, the training input variables are first fed to the input layer of the network and passed to the output layer gradually. Training is an iterative process of minimizing the differences between current actual output of the network and the desired output in the output layer. The desired output in the current study is the label of deceptive or truthful assigned to each data sample beforehand. The actual output is the output label automatically generated by the neural network after performing some kind of transformations such as sigmoid. The difference between the actual output and the desired output is calculated and back-propagated to the previous layer(s), which causes a series of adjustments of connection weights in such a way so as to reduce the observed output errors [33]. The above process is repeated for each sample in the training data set, and then repeated for the whole data set, until prespecified stopping conditions are met. The stopping condition could be either the maximum acceptable error rate or an arbitrary number of repetitions on the whole data set.

In the construction of neural networks, the number of hidden layers and the number of nodes in each layer are important decisions. Since increasing the number of hidden layers greatly increases the complexity of neural networks and requires larger data samples, the use of one hidden layer is very common. It is also a common practice to start the number of nodes in the hidden layer with $(I + O)/2$, where I and O are the numbers of input and output variables, respectively, and then adjust the number gradually to find the optimal one.

In the neural network, each input is indirectly connected to each output via all the units in the hidden layer. Due to the highly nonlinear structure of the neural network, there is no well-defined method to easily interpret the relative strength of each input to each output in the network. The sensitivity of each output to the small perturbation in each input may provide a good estimation of the relationship between the two. Based on the algorithm in Engelbrecht and Cloete [18], we developed an application to estimate the relative strength of each input variable to the output by calculating the partial derivatives of the output unit o with respect to each input unit $x_i (i = 1, \dots, I)$ while holding fixed the values of all other input variables x_k, $k \neq i$.

$$s_i = \frac{1}{N} \sum_{p=1}^{N_i} \frac{\partial o}{\partial x_i} = \frac{1}{N} \sum_{p=1}^{N_i} \sum_{j=1}^{L} \frac{\partial o}{\partial y_j} \frac{\partial y_j}{\partial x_i}, \tag{3}$$

where N is the number of input samples and L is the number of nodes in the hidden layer.

The partial derivative terms in formula (3) can be replaced with the multiplication of connection weights and derivatives of the activation function such as the sigmoid function. Thus, given the same perturbation in x_i, greater sensitivity is achieved when the change in derivatives of the activation function is greater.

Empirical Comparisons

Data Set Collection

EXPERIMENTAL DATA WERE COLLECTED from subjects performing a common decision-making task, the desert survival problem (DSP) [25], which we updated and modified for our program of research. The problem focused on being stranded in the Kuwaiti desert. The primary goal for participants was to achieve a consensus ranking of the items to be salvaged from their overturned jeep in order of their usefulness to survival. The participants were recruited from a management information systems course at a large southwestern university who received extra credit for their participation. The participants were randomly paired in two-person groups and communicated with each other about the ranking via an e-mail system developed by the research team. The e-mail messages exchanged between a pair of partners were the major source of data for analysis.

The groups were randomly assigned to one of the experimental conditions: deception or truth. In the deception condition, the participants who logged in first (randomly) were instructed to deceive the other partner; in the truth condition, no special instructions were given. In both conditions, the participants who sent messages first were called senders, and the other participants in dyads were called receivers. No participant was aware of the condition of his or her partner. Two different data sets were collected. The procedures for each were similar in terms of task design, communication tools, and duration of the experiment (three days in total), with a few exceptions. In the first experiment, DSP1, participants were given a half-day time slot to communicate with their partners. For example, if participants A and B were partners and A sent messages between midnight and midday on the first, second, and third days, B responded between midday and midnight on each of the three days. In other

Table 2. Summary of Data from the DSP Experiments.

Total number of	DSP1	DSP2
Messages	180	204
Subjects	60	52
Deceptive messages	48	55
Deceptive subjects	16	13
Truthful messages (senders)	48	39
Truthful subjects (senders)	16	13
Female:Male ratio	1.36:1	1:2.25

words, each participant was to compose at most three messages. In the second experiment, DSP2, the restriction on the number of message exchanges was relaxed. Thus, a participant was allowed to send any number of messages. Moreover, in DSP1, salvageable items were removed from consideration on the second and third days, so that the task was altered in succeeding days; with DSP2, the task remained constant.

Table 2 presents a descriptive summary of the data sets from DSP1 and DSP2. For purposes of the current analyses, only data from those members of each pair who were randomly designated as "senders" were included. The analyses thus constitute between-groups comparisons between senders in the respective deception and truth conditions. Data from receivers (partners) are not included.

Analysis Method

Twenty-four (eighteen original plus six additional) cues indicative of deception were used as input variables to train the classification models. Since the cue selection was mainly based on the analysis of DSP1 data (see [45]) and because of the similarity in the experimental design between DSP1 and DSP2, it was of interest to see whether the same cues remained effective for predicting deception in DSP2 data. When developing classification models, deceivers' data were used as positive samples, and truthful senders' data were selected as negative samples. Moreover, in order to determine which data unit (message or subject) could provide better predictive power, the classification methods were tested using both the collection of individual messages and the averages of each subject's messages. The first data set is called message data, and the second is labeled subject data. In the latter case, the estimate of each cue in individual messages was averaged by subject.

It is informative to train a classification method and measure its performance on the same data set. Testing classification models with the training data can reflect the ability to capture the underlying relationships between the input and output. It is more important, however, to evaluate the performance of classification methods on similar new data [14] because of the interest in the level of generalization and predictive power of the models. Therefore, each data set was randomly split into ten subsets. Nine of the subsets were selected to train the classification models, and the remaining subset was used to assess the performance of the models. In order to avoid the bias from a single split of data and get an objective measure of performance of classification methods, the split procedure is repeated ten times by choosing one of the subsets as the testing data and the other nine as the training data. Finally, the classification performance was measured by averaging the performances in the ten tests. The entire procedure is called tenfold cross-validation. This procedure was applied to testing all of the classification methods except discriminant analysis. For discriminant analysis, however, cross-validation was conducted by classifying each case with the functions derived from all cases other than that case in a process called leave-one-out cross-validation. Logistic regression models were evaluated using both tenfold and leave-one-out cross-validation approaches. On the one hand, logistic regression is as common a statistical approach as discriminant analysis. On the other hand, the logistic function and sigmoid function in BP neural networks are identical. In order to compare different methods on an equal footing, we applied both validation approaches to logistic regression methods.

The classification performance of the four methods was compared on three measures: (1) overall performance—overall percentage of accurate classification; (2) deception performance—percentage of deceptive messages accurately classified, and (3) truth performance—percentage of truthful messages accurately classified.

Results

Architecture of Network and Size of Decision Tree

CONSIDERING THE SMALL-SIZED DATA SET and relatively large number of cues, the neural network was configured with one hidden layer. As stated before, determining the number of nodes in the hidden layer is an important issue in configuring neural networks. A small number of nodes in the hidden layer not only reduces the computation complexity but also prevents overfitting the data. Based on the testing, three and one appeared to be a good choice for the number of nodes in the hidden layer for message and subject data of DSP2, respectively, and three for both message and subject data of DSP1.

The mean size of decision trees induced from message data of DSP2 included 18 leaves, 13 layers, and 35 nodes. After pruning, the tree size was reduced to 4 leaves, 3 layers, and 7 nodes. Nonetheless, the issue was greatly mitigated in the subject data of DSP2 because the size of the decision tree for subject data was reduced from 7 nodes to 3 nodes by pruning. A similar pattern occurred in the DSP1 data. The size of a decision tree suggests the difficulty of classification and the number of target classes. If the tree size is greatly reduced after pruning, it may indicate that the data are less statistically reliable [12]. Message data had a large initial tree and a much smaller pruned one, which revealed that there was a high degree of residual variation in these data sets. The small reduction in the size of trees for subject data suggested that subject data may be more statistically reliable.

Classification Performance of the Four Methods

The performances of the four classification methods are shown in Table 3. To reflect the level of degradation in classification performance with holdout samples, we list the results on the training data along with the cross-validation results. In order to gain a deeper understanding of the classification performance on deception data as opposed to truth data, Table 3 also shows the deception performance and truth performance with each classification method.

Overall, it can be observed that there were obvious reductions in the classification performance for the cross-validation compared with that for training. Three out of four methods could differentiate between deceptive subjects and truthful subjects from the training data nearly perfectly. However, when they were tested on the holdout data in cross-validations, the performances degraded substantially. It highlighted the great variability in the data set, pointing to the difficulty of predicting deception.

In comparing the training performance of classification models between subject and message data, it is notable that the former was superior to the latter. It implies that there may be more noise in the message data, which is difficult for classification models to capture. Despite the smaller size of subject data, the aggregate nature of subject data may have contributed to its lower variability.

The generalizability of classification models is reflected in the cross-validation performance. The results showed that the overall performance on subject data varied greatly for DSP2, ranging from a respectable 61.5 percent for neural network methods to 46.2 percent for decision trees, which is no better than chance, and ranging from a respectable 76.7 percent by neural network methods to 53.3 percent by both decision tree and discriminant analysis methods for DSP1. However, there were relatively smaller differences in the predictability power for message data, ranging from 66 percent by neural networks to 55.3 percent by decision trees for DSP2, and from 79.2 percent by neural networks to 66.7 percent by discriminant analysis for DSP1. The overall low performances across classification methods indicated that some noise existing in the data may have prevented the classification methods from achieving optimal performance. It underscored the need for screening the input variables by eliminating those variables that have little influence on the classification results.

The side-by-side comparison between the performance on deception data and truth data revealed that, in general, deception was more accurately identified than truth with few exceptions. It means that the percent of deception data predicted by the classification methods that were truly deceptive was higher than the percent of truth data that were actually truthful. This indicates a false-positive bias in these classification models toward declaring truthful messages or subjects as deceptive.

Important Cues Selected by the Four Approaches

Identifying important cues has theoretical, practical, and methodological implications. There are a host of theories on social interaction that may be applicable to the deception research. Conversely, significant cues identified in this study may provide

Table 3. Summary of the Performances of Classification Methods.

Classification methods	Discriminant analysis		Logistic regression		Decision trees		Neural networks	
Test methods	Training	Cross-validation	Training	Cross-validation	Training	Cross-validation	Training	Cross-validation
(a) DSP1								
Overall performance								
Subject	100	53.3	100	66.7/63.3	96.7	53.3	100	76.7
Message	85.4	66.7	93.8	70.8/70.8	95.8	72.9	98.9	79.2
Truth performance								
Subject	100	57.1	100	71.4/57.1	92.9	50	100	71.4
Message	83.7	60.5	90.7	67.4/69.8	90.1	67.4	97.7	72.1
Deception performance								
Subject	100	50	100	62.5/68.8	100	56.3	100	81.3
Message	86.8	71.1	96.2	73.6/71.7	100	77.4	100	84.9
(b) DSP2								
Overall performance								
Subject	100	42.3	100	53.8/57.7	96.2	42.3	100	61.5
Message	77.7	63.8	79.8	61.7/59.6	97.8	55.3	95.7	66
Truth performance								
Subject	100	38.5	100	53.3/57.1	100	41.7	100	60
Message	79.5	61.5	76.3	54.1/51.4	97.4	44.4	100	59
Deception performance								
Subject	100	46.2	100	54.5/58.3	92.9	42.9	100	63.6
Message	76.4	66.5	82.1	66.7/61.9	98.2	59.7	93.2	70.9

Notes: For cross-validation in logistic regression, for example, 53.8/57.7, the first number was based on the leave-one-out cross-validation, and the second number was based on tenfold cross-validation.

new evidence in support of existing theories. Moreover, the discovered important cues may be applied to detecting deception in practice. Furthermore, reducing the size of input variables may improve the performance of classification methods.

All of the four methods have some built-in mechanisms to assist selecting important cues from the original cue list. For example, the β coefficient and/or *p*-value from the discriminant analysis and logistic regression models are directly suggestive of the discriminatory power of a cue. The attributes included in the pruned decision trees allow for more information gains to make classification decisions. The input variables of neural network models to which the output is more sensitive are indicative of their importance to predicting the output. Based on the above cue-selection criteria, important cues were selected for each data set, as shown in Table 4.

It is notable from Table 4 that there is considerable overlap on the selected cues between discriminant analysis and logistic regression. For example, verbs, content diversity, and highly positive pleasantness were identified as significant cues for subject data by the two methods. Since both methods essentially use maximum likelihood estimation in selecting significant cues, it is not surprising that the selected cues were similar to each other. Decision trees and neural networks use different criteria in choosing important cues, consequently assigning higher weights to different sets of cues. Among the cues resulting from the four methods, those from decision trees were most distinct. For example, redundancy and perceptual information were considered as important for subject data only by the decision tree model. The most information gains required by decision trees resulted in the most reduction in uncertainty after splitting the data sets, which was divergent from the criteria employed in other approaches.

It is common in deception research to select cues to deception by examining their statistical significance. It is a good practice if we use statistical approaches for prediction. However, it should be cautioned that important cues identified by statistical methods might not be appropriate for machine learning models. On the other hand, cues that are important for machine learning models may not have statistical power. For example, spatiotemporal information and positive pleasantness selected by the two machine learning approaches for message data were not statistically significant. It highlights the difference between the highly nonlinear model of neural networks and the linear model of discriminant analysis.

There was a small overlap on the selected cues between the two data sets, as shown in Table 4. It is argued that deception is moderated by many factors, such as communication medium [20], interactivity [4], context of relationship and motivations [15], and so on. Due to the slight difference in experimental design between DSP2 and DSP1, the significance of a cue varied from one to the other. For example, content diversity was very important for DSP2 only, but pausality was important for DSP1 only. Nonetheless, some cues consistently emerged as important for both data sets, such as verbs, modal verbs, typo ratio, and highly positive pleasantness.

Classification Performance with the Important Cues Only

Reducing a large set of cues to a small set of important cues may bring multifold benefits to the classification models. First, it can directly lower the computation

Table 4. Important Cues Selected by the Classification Models.

	Discriminant analysis		Logistic regression		Decision trees		Neural networks	
Cues to deception	Subject	Message	Subject	Message	Subject	Message	Subject	Message
Verbs	1	2	1	2	1		1	
Modifiers		1		1|2			1	1|2
Average word length	2	2	2					2
Pausality	1	1	1	1		1		1
Modal verbs	2		1|2			1	2	2
Individual references		2		2		2		
Group references		1|2		1|2		1	1	
Emotiveness							2	
Content diversity	2	2	2	2		2	2	2
Redundancy					2			
Perceptual information					2	2		2
Spatiotemporal information					1	1|2	1	1|2
Typographical error ratio		2		1		2	2	1|2
Affect						1	1	2
Imagery						2		
Pleasantness			1|2			2		
Positive activation			1					
Positive pleasantness						1|2	1	1|2
Positive imagery						1|2	2	1
Negative activation				1			2	
Highly positive pleasantness	1|2	1	2				1|2	
Highly positive activation				1	1	2		1

Note: "1" and "2" indicate that the cue was selected for DSP1 and DSP2, respectively.

complexity. Second, it may reduce the error rate and improve the classification performance. Third, it avoids overfitting the data and increases generalizability of the resulting models. The performances of classification models using important cues are listed in Table 5. To gain a better understanding of the change in classification performance before and after adopting important cues, the cross-validation results for DSP1 data sets using all the original cues and using only important cues are displayed in Figure 3. The DSP2 data set shows a similar pattern.

As shown in Table 5 and Figure 3, cross-validation results were consistently improved with the reduced set of cues. They illustrate the advantage of selecting cues that increase the generalizability of the classification models. A smaller set of important cues was superior to the larger set of cues in predicting deception.

With important input variables only, the neural network outperformed the other methods in predicting deception for DSP2 subject data with 88.5 percent precision, while the decision tree with 65.4 percent was the worse. The classification performances for DSP1 subject data were similar across different methods, with logistic regression and neural network methods achieving 83.3 and 80 percent accuracy, respectively. The performances for message data were also similar across the classification methods despite the data set, with only one exception. Decision tree performed significantly poorer on the DSP2 data set with 57.4 percent precision. Comparison of the two statistical methods suggested that the logistic regression and discriminant analysis methods performed equally well on DSP2 data, but the former was relatively better on the DSP1 data. Between the two machine learning methods, neural networks achieved similar performances on the DSP1 data to decision trees, but the former was far better on the DSP2 data.

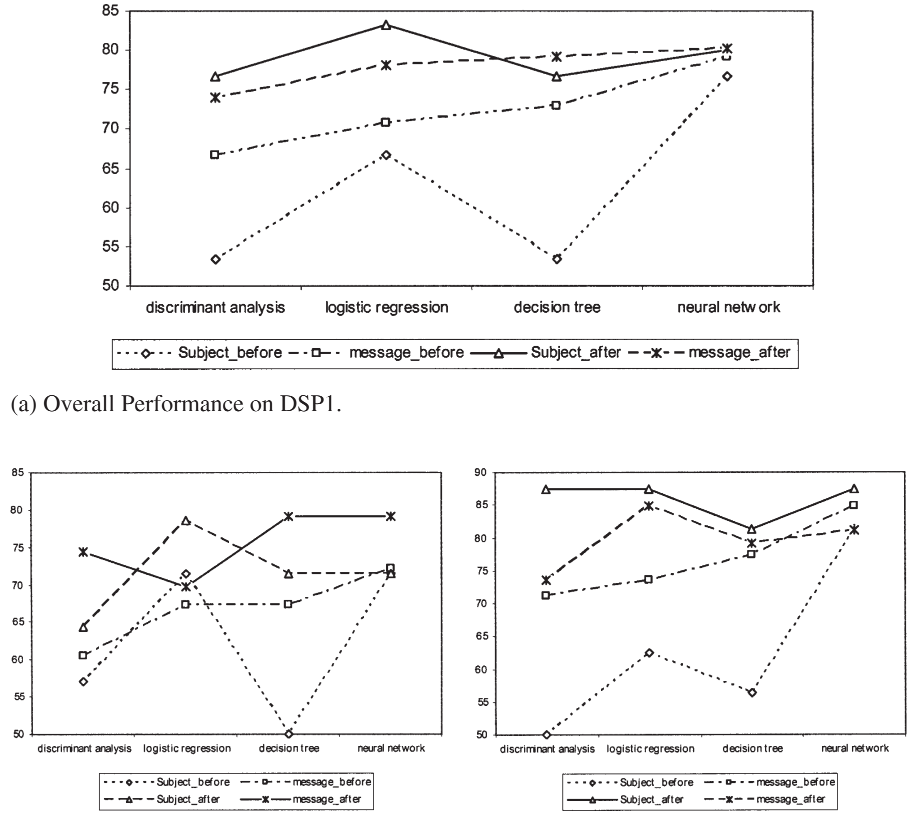
To sum up, all the classification methods appeared to be on par on predicting deception from DSP1 data. For DSP2 data, three methods showed similar performance, the exception being decision trees. In addition, the pattern that deception performance was higher than truth performance was consistently exhibited in the results of logistic regression and neural network models across data sets. However, the pattern received mixed support for discriminant analysis and decision tree approaches.

Discussion

NO SINGLE METHOD EMERGED AS SUPERIOR for predicting deception. All of the four methods under investigation are potentially good alternatives if only important cues are included in the model. The average performances on DSP1 subject and message data were 77.9 percent and 70.5 percent, respectively, and the average performances on DSP2 subject and message data were 79.2 and 78.1, respectively. The results indicate that the methods not only were able to capture the underlying relationships in the deception data but also showed good generalizability. The performances of discriminant analysis, logistic regression, and neural networks were consistent across data sets, while decision trees showed poorer performance on DSP2 data. Their low performance indicated that there was great variability within the data set. However, the performance of

Table 5. Summary of Classification Performance with Important Cues Only.

Classification methods	Discriminant analysis		Logistic regression		Decision trees		Neural networks	
Test methods	Training	Cross-validation	Training	Cross-validation	Training	Cross-validation	Training	Cross-validation
(a) DSP1								
Overall performance								
Subject	86.7	76.7	100	83.3/83.3	96.7	76.7	100	80
Message	76	74	82.3	78.1/79.2	95.8	79.1	92.7	80.2
Truth performance								
Subject	78.6	64.3	100	78.6/78.6	92.9	71.4	100	71.4
Message	76.7	74.4	74.4	69.8/72.1	90.7	79.1	95.3	79.1
Deception performance								
Subject	93.8	87.5	100	87.5/87.5	100	81.3	100	87.5
Message	75.5	73.6	88.7	84.9/84.9	100	79.2	90.6	81.1
(b) DSP2								
Overall performance								
Subject	84.6	80.8	84.6	76.9/76.9	96.1	65.4	100	88.5
Message	75.7	72.3	75.5	72.3/74.4	76.6	57.4	79.8	74.5
Truth performance								
Subject	84.6	76.9	84.6	76.9/76.9	100	66.7	100	85.7
Message	76.9	74.4	73.5	69.7/72.7	73	48.1	70.8	69.2
Deception performance								
Subject	84.6	84.6	84.6	76.9/76.9	92.9	64.3	100	91.7
Message	74.5	70.9	76.7	73.8/75.4	78.9	61.2	89.1	78.2



(a) Overall Performance on DSP1.

(b) Truth Performance on DSP1.

(c) Deception Performance on DSP1.

Figure 3. Cross-Validation Performance on DSP1 Data Sets. *Note:* (a) subject_before: performance on subject data before selecting important cues; message_after: performance on message data after selecting important cues leave-one-out cross-validation results for logistic regression models are included.

decision trees may be improved by improving tree pruning techniques [31]. Thus, we are not in a position to conclude that decision trees are inferior based on one case.

Neural networks are generally good at representing a high degree of nonlinear relationships between input and output. The close call between the neural network and other approaches implies that the level of nonlinearity in the experimental data were low. It is further confirmed by referring to the size of neural network models for both data sets, including either one or three nodes in the hidden layer. The size of neural networks indicates the level of interaction between the input variables [12]. Thus, it was not surprising that methods favoring linear or one-dimensional relationships were still effective. The near linearity might have resulted from our removing highly correlated data when refining the original list of cues.

With small margins, neural networks exhibited better performance than the other methods. The significance of the difference will be tested with future data sets. More importantly, neural networks were consistently reliable across all of test settings. As shown in Figure 3, the performances of neural networks on two data sets, either before or after selecting important cues were close to one another. Although both discriminant analysis and logistic regression performed equally well as neural networks on training data, their performances in various test settings were much more dispersed. For example, before selecting important cues, the overall performance of discriminant analysis on subject data of DSP1 was 53.3 percent, which jumped to 76.7 percent afterward. Furthermore, as statistical methods, both discriminant analysis and logistic regression methods assume that the underlying data possess certain features; otherwise, they may not fit the data. An advantage of neural networks is that they are quite robust across applications and data sets.

The improvement of the cross-validation results after pruning input variables underlines the importance of identifying important cues. For example, the overall performance of discriminant analysis and on DSP2 subject data climbed 38.5 percentage points with the pruned input. Selecting a smaller set of important cues as input not only directly reduced the computational complexity of classification models but also removed noise that is negatively associated with classification accuracy.

Compared with the classification performances on message data, those on subject data were generally better or as good, as shown in Table 5. It suggests that subjects (or an aggregated estimation of subjects' behavior over the entire course of an interaction) instead of messages may be a better unit for deception analysis. This argument is supported by the propositions in IDT that deceivers may strategically manage their behavior by occasionally taking a low-key stance and making their behavior look like truth tellers'. It is also supported by interaction adaption theory's claim that communication partners may reciprocate or compensate their behavior over the course of interaction. Thus, an individual message from a deceiver may not always exhibit deceptive behavior. On the contrary, if we combine a series of messages from a deceiver, it may bring deceptive behavior to light that is otherwise hidden in individual messages. Deception may not be reflected in one-point behavior, but is more likely to be unveiled from a series of behaviors [44]. One shortcoming of the current analysis is that the sample sizes of the subject data were much smaller than those of the message data. Both machine learning and statistical methods favor large data sets. With increases in sample size, the advantage of using subject data rather than message data should be more pronounced. Nonetheless, when there are only a few messages exchanged by a deceiver, individual messages may also be able to provide good indications of deception.

The mixed patterns of higher deception performance for some of the methods did not conform to the findings from a prior study showing that the detection accuracy was much higher on truth than on deception [7]. The test results even displayed a tendency toward an opposite pattern. Compared with the prior study that was conducted in face-to-face settings, the current investigation involved a different context and availability of cues to deception. To reach the end of predicting deception, deception performance may be more important than truth performance. Therefore, the linguistic cues accessible in CMC, as examined in this study, may be helpful for improving the accuracy of detecting deception. As a caution, however, methods that do well in detecting deception may also yield too many false alarms. Thus, there is potential risk in favoring such approaches. Of course, firm conclusions should not be drawn on the basis of a single study, but the results at least show that deception performance may be better than truth performance in some contexts.

In order to enrich and provide new evidence for existing cues to deception, we also identified the directions of important cues. As predicted in the prior research, deceivers used more modal verbs and fewer individual references, making their utterances more tentative and nonspecific [29]. Due to the negative experience associated with deception, deceivers tend to disassociate themselves from their messages and display higher non-immediacy in their language. Moreover, average word length, an indication of complexity, was lower for deceivers than for truth tellers. Deception is known as a cognitively taxing process. As a result of greater cognitive demands when fabricating plausible and consistent messages, deceivers tend to resort to less complex language in communication to save some effort. In contrast with the prior findings that the quantity of deceptive messages is less than that of truthful messages, we found the opposite. This is in line with the patterns from a prior study using a similar decision-making task [45]. Deceivers in the current study may have given more elaborate reasons for their rankings in order to gain credibility from communication partners and to achieve their deception goals. In this case, deceivers intent on being persuasive are likely to use more verbs and modifiers, and more expressive language. Most of the literature, however, focuses on statements of fact or recollections, such as criminal statement analysis, when deceivers do not have the same details to put into messages as truth tellers do. Under such circumstances, different linguistic patterns were expected. Moreover, in the current study, deceivers tended to include more sensory, spatial, and temporal details than truth tellers, which is opposite to the propositions in reality monitoring [24] and criteria-based content analysis [35]. Since the two criteria were based on the assumption that there is a real event or past memory for participants to resort to, the hypothetical context of desert survival may have caused a complete reversal of the above pattern. Therefore, when there is lack of support of real experience, more sensory and spatial details may indicate deception rather than truth. Imagery is a newly added cue in this study. On average, deceivers ($M = 1.58$) showed lower imagery ratio than truth tellers ($M = 1.69$) in DSP2. Compared with the previous finding on content diversity [45], DSP2 revealed an opposite direction. Instead of being lower, the content diversity of deceivers was slightly higher than truth tellers on DSP2. As stated before, we relaxed the restriction on the number of message exchanges in DSP2, which might have accounted for the discrepancy. When deceivers are given the opportunity to interact as many times as they want to, they tended to use different words to enhance their persuasiveness. The high content diversity may also be a result of deceivers' changing the subject to avoid being suspected, or it may be a result of creating shorter messages. Driving toward the goal of reaching consensus on the assigned task, truth tellers were likely to repeat some topic words in continuous communication.

The power of back-propagation neural networks lies in their ability to represent complicated, and highly nonlinear, relationships. They therefore have great potential. Whether or not this potential is realized depends, to a large extent, on the quality of data used to train them. It is premature to conclude the effectiveness of any classification methods at this point. Further investigation with larger data sets will give us deeper insight into the intra-relations of cues as well as the robustness of different classification methods.

Acknowledgments: The authors thank Fang Chen for her help in collecting the experimental data for DSP2. Portions of this research were supported by funding from the U.S. Air Force Office of Scientific Research under the U.S. Department of Defense University Research Initiative (Grant F49620–01–1-0394). The views, opinions, and findings in this report are those of the authors' and should not be construed as an official Department of Defense position, policy, or decision.

References

1. Bauchner, J.E.; Kaplan, E.A.; and Miller, G.R. Detecting deception: The relationship of available information to judgmental accuracy in initial encounter. *Human Communication Research, 6,* 3 (1980), 253–264.

2. Buller, D.B., and Burgoon, J.K. Interpersonal deception theory. *Communication Theory, 6,* 3 (1996), 203–242.

3. Burgoon, J.K.; Blair, J.P.; Qin, T.; and Nunamaker, J.F., Jr. Detecting deception through linguistic analysis. In H. Chen, R. Miranda, D. Zeng, C. Demchak, J. Schroeder, and T. Madjusudan (eds.), *First NSF/NIJ Symposium on Intelligence and Security Informatics.* New York: Springer, 2003, pp. 91–101.

4. Burgoon, J.K.; Buller, D.B.; and Floyd, K. Does participation affect deception success? A test of the interactivity principle. *Human Communication Research, 27,* 4 (2001), 503–534.

5. Burgoon, J.K.; Burgoon, M.; and Wilkinson, M. Writing style as predictor of newspaper readership, satisfaction and image. *Journalism Quarterly, 58,* 2 (1981), 225–231.

6. Burgoon, J.K.; Dillman, L.; and Stern, L.A. Adaptation in dyadic interaction: defining and operationalizing patterns of reciprocity and compensation. *Communication Theory, 3,* 4 (1993), 295–316.

7. Burgoon, J.K.; Buller, D.B.; Ebesu, A.S.; and Rockwell, P. Interpersonal deception V: Accuracy in deception detection. *Communication Monographs, 61,* 4 (1994), 303–325.

8. Carlson, J.R., and Zmud, R.W. Channel expansion theory and the experiential nature of media richness perceptions. *Academy of Management Journal, 42,* 2 (1999), 153–170.

9. Cleary, P.D., and Angel, R. The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior,* 25, (1984), 334–348.

10. Cunningham, H. A general architecture for text engineering. *Computers and the Humanities, 36,* 2 (2002), 223–254.

11. Cunningham, H.; Maynard, D.; Bontcheva, K.; and Tablan, V. A framework and graphical development environment for robust NLP tools and applications. Paper presented at the Proceedings of Proceedings of the Fortieth Anniversary Meeting of the Association for Computational Linguistics, Philadelphia, July 2002.

12. Curram, S.P., and Mingers, J. Neural networks, decision tree induction and discriminant analysis: An empirical comparison. *Journal of Operational Research Society, 45,* 4 (1994), 440–450.

13. Daft, R., and Lengel, R. Organizational information, message richness and structural design. *Management Science, 32,* 5 (1986), 554–571.

14. Denker, J.; Schwartz, D.; Wittner, B.; Solla, S.; Howard, R.; Jackel, L.; and Hopfield, J. Large automatic learning, rule extraction, and generalization. *Complex Systems, 1,* 5 (1987), 877–922.

15. DePaulo, B.M.; Stone, J.I.; and Lassiter, G.D. Deceiving and detecting deceit. In B.R. Schlenker (ed.), *The Self and Social Life.* New York: McGraw-Hill, 1985, pp. 323–370.

16. DePaulo, B.M.; Kashy, D.A.; Kirkendol, S.E.; Wyer, M.M.; and Epstein, J.A. Lying in everyday life. *Journal of Personality and Social Psychology, 70* (May 1996), 979–995.

17. DePaulo, B.M.; Lindsay, J.J.; Malone, B.E.; Muhlenbruck, L.; Charlton, K.; and Cooper, H. Cues to deception. *Psychological Bulletin, 129,* 1 (2003), 74–112.

18. Engelbrecht, A.E., and Cloete, I. A sensitivity analysis algorithm for pruning feedforward neural networks. In *Proceedings of IEEE International Conference on Neural Networks.* Los Alamitos, CA: IEEE Press, 1996, pp. 1274–1277.

19. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Annals Eugenics, 7,* 2 (1936), 179–188.

20. George, J.F., and Carlson, J.R. Electronic lies: Lying to others and detecting lies using electronic media. In *Proceedings of Fifth Americas Conference on Information Systems.* Atlanta, GA: AIS, 1999, pp. 612–614.

21. Grice, H.P. Logic and conversation. In P. Cole and J.L. Morgan (eds.), *Syntax and Semantics, Vol. 3, Speech Acts.* New York: Academic Press, 1975, pp. 41–58.

22. Hooper, R., and Bell, R.A. Broadening the deception construct. *Quarterly Journal of Speech, 70,* 2 (1984), 288–302.

23. Johnson, R.A., and Wichern, D.W. *Applied Multivariate Statistical Analysis,* 5th ed. Upper Saddle River, NJ: Prentice Hall, 2002.

24. Köhnken, G.; Schimossek, E.; Aschermann, E.; and Höfer, E. The cognitive interview and the assessment of the credibility of adults' statements. *Journal of Applied Psychology, 80,* 6 (1995), 671–684.

25. Lafferty, J., and Eady, P. *The Desert Survival Problem.* Plymouth, MI: Experimental Learning Methods, 1974.

26. Lal, P. Text summarisation. Masters dissertation, Department of Computer Science, Imperial College, London, 2002.

27. Lynch, M.D. Stylistic analysis. In P. Emmert and W.D. Brooks (eds.), *Methods of Research in Communication.* Boston: Houghton Mifflin, 1970, pp. 315–342.

28. Maynard, D.; Cummingham, H.; Bontcheva, K.; Catizone, R.; Demetriou, G.; Gaizauskas, R.; Hamza, O.; Hepple, M.; Herring, P.; Mitchell, B.; Oakes, M.; Peters, W.; Setzer, A.; Stevenson, M.; Tablan, V.; Ursu, C.; and Wilks, Y. A survey of uses of GATE. University of Sheffield, UK, 2000.

29. Mehrabian, A., and Wiener, M. Non-immediacy between communicator and object of communication in a verbal message: Application to the inference of attitudes. *Journal of Consulting Psychology, 30,* 5 (1966), 420–425.

30. Mingers, J. An empirical comparison of pruning methods for decision-tree induction. *Machine Learning, 4,* 2 (1989), 227–243.

31. Qin, T.; Burgoon, J.K.; and Nunamaker, J.F., Jr. An exploratory study on promising cues in deception detection and application of decision tree. In R.H. Sprague, Jr. (ed.), *Proceedings of the Thirty-Seventh Annual Hawaii International Conference of System Sciences.* Los Alamitos, CA: IEEE Computer Society Press, 2004 (available at www.hicss.hawaii.edu/HICSS37/apahome37.html).

32. Quinlan, J.R. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann Publishers, 1993.

33. Rumerlhart, D.E., and McClelland, J.L. (eds.) *Parallel and Distributed Processing: Exploration in the Microstructure of Cognition.* Cambridge, MA: MIT Press, 1986.

34. Sporer, S.L. The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology, 11,* 5 (1997), 373–397.

35. Steller, M., and Köhnken, G. Criteria-based content analysis. In D.C. Raskin (ed.), *Psychological Methods in Criminal Investigation and Evidence.* New York: Springer-Verlag, 1989, pp. 217–245.

36. Tam, K.Y., and Kiang, M.Y. Managerial applications of neural networks: The case of bank failure predictions. *Management Science, 38,* 7 (1992), 926–947.

37. Vrij, A.; Edward, K.; Robert, K.P.; and Bull, R. Detecting deceit via analysis of verbal and nonverbal behavior. *Journal of Nonverbal Behavior, 24,* 4 (2000), 239–263.

38. Whissell, C.; Fournier, M.; Pelland, R.; Weir, D.; and Makarec, K. A dictionary of affect in language: IV. Reliability, validity, and applications. *Perceptual and Motor Skills, 62* (June 1986), 875–888.

39. White, C.H., and Burgoon, J.K. Adaptation and communicative design: Patterns of interaction in truthful and deceptive conversation. *Human Communication Research, 27,* 1 (2001), 9–37.

40. Winston, P.H. *Artificial Intelligence,* 3d ed. Boston: Addison-Wesley Longman, 1992.

41. Witten, I.H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Francisco: Morgan Kaufmann, 1999.

42. Yoon, Y.; George Swales, J.; and Margavio, T.M. A comparison of discriminant analysis versus artificial neural networks. *Journal of Operational Research Society, 44,* 1 (1993), 51–60.

43. Zahedi, F. A meta-analysis of financial applications of neural networks. *International Journal of Computational Intelligence and Organizations, 1,* 3 (1996), 164–178.

44. Zhou, L.; Burgoon, J.K.; and Twitchell, D. A longitude analysis of language behavior of deception in e-mail. In H. Chen, R. Miranda, D. Zeng, C. Demchak, J. Schroeder, and T. Madhusudan (eds.), *First NSF/NIJ Symposium on Intelligence and Security Informatics.* New York: Springer, 2003, pp. 102–110.

45. Zhou, L.; Twitchell, D.; Qin, T.; Burgoon, J.; and Nunamaker, J.F., Jr. An exploratory study into deception detection in text-based computer-mediated communication. In R.H. Sprague, Jr. (ed.), *Proceedings of Thirty-Sixth Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press, 2003 (available at www.hicss.hawaii.edu/HICSS36/apahome36.htm).