2012

Discerning truth from deception: Human judgments and automation efforts

Victoria L. Rubin
Western University, vrubin@uwo.ca

Niall Conroy
The University of Western Ontario, nconroy1@uwo.ca

first monday
PEER-REVIEWED JOURNAL ON THE INTERNET

HOME ABOUT LOGIN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS SUBMISSIONS

Home > Volume 17, Number 3 - 5 March 2012 > **Rubin**

first monday
PEER-REVIEWED JOURNAL ON THE INTERNET

Discerning truth from deception: Human judgments and automation efforts

by Victoria L. Rubin and Niall Conroy

Abstract

Recent improvements in effectiveness and accuracy of the emerging field of automated deception detection and the associated potential of language technologies have triggered increased interest in mass media and general public. Computational tools capable of alerting users to potentially deceptive content in computer–mediated messages are invaluable for supporting undisrupted, computer–mediated communication and information practices, credibility assessment and decision–making. The goal of this ongoing research is to inform creation of such automated capabilities. In this study we elicit a sample of 90 computer–mediated personal stories with varying levels of deception. Each story has 10 associated human deception level judgments, confidence scores, and explanations. In total, 990 unique respondents participated in the study. Three approaches are taken to the data analysis of the sample: human judges, linguistic detection cues, and machine learning. Comparable to previous research results, human judgments achieve 50–63 percent success rates, depending on what is considered deceptive. Actual deception levels negatively correlate with their confident judgment as being deceptive (r = -0.35, df = 88, p = 0.008). The highest-performing machine learning algorithms reach 65 percent accuracy. Linguistic cues are extracted, calculated, and modeled with logistic regression, but are found not to be significant predictors of deception level, confidence score, or an authors' ability to fool a reader. We address the associated challenges with error analysis. The respondents' stories and explanations are manually content–analyzed and result in a faceted deception classification (theme, centrality, realism, essence, self–distancing) and a stated perceived cue typology. Deception detection remains novel, challenging, and important in natural language processing, machine learning, and the broader library information science and technology community.

Contents

OPEN JOURNAL SYSTEMS

Journal Help

USER

Username
Password
□ Remember me
Login

JOURNAL CONTENT

Search

All

Search

Browse

- By Issue
- By Author
- By Title
- Other Journals

FONT SIZE

CURRENT ISSUE

ATOM 1.0
RSS 2.0
RSS 1.0

ARTICLE TOOLS

Abstract
Print this article
Indexing metadata
How to cite item
Email this article (Login required)
Email the author (Login required)

ABOUT THE AUTHORS

Victoria L. Rubin
University of Western Ontario

Principal Investigator of the Language and

A half truth is a whole lie.
Yiddish proverb

The trouble with lying and deceiving is that their efficiency depends entirely upon a clear notion of the truth that the liar and deceiver wishes to hide.
Hannah Arendt (1906–1975)

Introduction

Recently mass media has shown acute interest to the topic of discerning truth from deception in computerized ways. This interest may be explained by recent advances in the research and development of automated deception detection with natural language processing (NLP) and machine learning techniques. In a recent CBC interview, a computer science student from MIT's Media Lab declared his intention to develop "a kind of spell checker for facts", called "Truth Goggles". They are imagined to be an automated version of Truth–O–Meter (politifact.com) which is a fact–checking tool powered by the research of many journalists cross–referencing statements made by high–profile politicians, statesmen, celebrities (CBC, 2011). A recent article in the *New York Times* Business Day section highlights research of several prominent figures in the computational speech community [1] (Julia Hirschberg, Columbia University; Dan Jurafsky and David Larcker, Stanford; Shrikanth Narayanan, University of Southern California; Eileen Fitzpatrick, Montclair State University) who are "pars[ing] people's speech for patterns that gauge whether they are being honest" and "training computers to recognize hallmarks of what they call emotional speech — talk that reflects deception, anger, friendliness and even flirtation" (Eisenberg, 2011). The envisioned potential outcomes of these cutting edge computational abilities are suggested:

> "Programs that succeed at spotting these submerged emotions may someday have many practical uses: software that suggests when chief executives at public conferences may be straying from the truth; programs at call centers that alert operators to irate customers on the line; or software at computerized matchmaking services that adds descriptives like 'friendly' to usual ones like 'single' and 'female'."

With the increasing use of computer–mediated communication (CMC) in all aspects of modern life, deception detection in CMC has emerged as an important issue in everyday communications, and is now of interest within the broad library and information science field. Deception is potentially disruptive in everyday communication, information seeking, and decision–making. Rubin (2010) positions the study of deception in library and information science and technology (LIS&T) alongside its positively–charged counterparts — trust, credibility, certainty, and authority, and affirms automated deception detection as a recently attainable contribution from natural language processing (NLP) and machine learning. Deception is defined in CMC as "a message knowingly and intentionally transmitted by a sender to foster a false belief or conclusion by the *perceiver*" (Rubin [2010]; synthesized from Buller and Burgoon, 1996; Zhou, *et al.*, 2004). Rubin (2010) argues that although largely socially condemned, deception is widespread and often undetected, especially in electronic environments where credibility assessments are difficult due to absence of many traditional cues such as verifiable credentials or face–to–face contact. The need arises for decision support tools capable of alerting users to potentially deceptive content:

> Tools to detect deceit from language use. pose a promising avenue for increasing the ability to distinguish truthful transmissions, transcripts, intercepted messages, informant reports and the like from deceptive ones. [2]

Human language technologies (used here synonymously with NLP) are said to be "dramatically improving in their effectiveness and accuracy, which is accompanied by a significant expansion of the HLT community itself" (Wyner and Branting, 2011). A new specialized workshop is put forward as a part of the 2012 Conference of the European chapter of the Association for Computational Linguistics calling, with great optimism, upon new methods in the area, and announcing "a relatively new area of

Information Technologies Research Lab and an Assistant Professor at the Faculty of Information and Media Studies at the University of Western Ontario, London, Canada. She received her Ph.D. in Information Science and Technology in 2006, her M.A. in Linguistics in 1997 from Syracuse University, NY, USA, and a B.A. English, French, and Interpretation from Kharkov State University, Ukraine in 1993. Dr. Rubin research interests are in information organization and information technology. She specializes in information retrieval and natural language processing techniques that enable analyses of texts to identify, extract, and organize structured knowledge.

Niall Conroy
University of Western Ontario

Doctoral student in LIS at the Faculty of Information and Media Studies at the University of Western Ontario, London, Canada. He has received BSc in Computer Science and an MLIS, and has a background in software engineering and information management. His current research interests include intelligent music information retrieval systems, collaborative filtering and multimedia data mining techniques, and online communities.

applied computational linguistics that has broad applications in business fraud and online misrepresentation, as well police and security work" (European chapter of the Association for Computational Linguistics [EACL], 2012).

To inform the creation of such emerging deception detection tools, our current study uses a training dataset and a variety of analytic techniques and metrics to investigate predictors of deception, as expressed in subjective participant explanations and, more importantly, in the objective content of the texts. We draw on our expertise in the related area of subjectivity and sentiment analysis or opinion mining (see Pang and Lee [2008] for an overview). In particular, the first author's previous experience is in emotion identification (Rubin, *et al.*, 2004), analysis of statement certainty levels (absolute, high, moderate, low or uncertain) (Rubin, 2007; 2010a), weblog credibility assessment (Rubin and Liddy, 2006), identification of trust and distrust (Rubin, 2009), and selective manual blog mining of serendipitous accounts (Rubin, *et al.*, 2011). Our previous studies model and acquire "reliable cues for recognizing complex phenomena which are, at least partly, expressed through language" in order to improve access to information by developing methods for applications that can approach human–like understanding of texts through NLP (Language & Information Technology Research Lab, 2012).

The remainder of this paper is structured as follows. First, we outline the study goals within text-based CMC environments and provide a literature overview focusing on both human and automated abilities to discern truths from lies [3]. Then, we describe our dataset elicitation procedures, introduce the experiments with the use of human judgments as a benchmark, and overview our methods, techniques, and measures. We conclude with error analysis that allows us to create the deception classification. We reflect upon encountered challenges, provide alternative paths towards solving the problem of automated deception detection. We start with a set of research objectives.

Study objectives

In order to be able to discern deceptive messages from truthful ones automatically, we require a) an elicitation technique for obtaining samples of both deceptive and truthful messages in CMC; b) "gold standard" results for comparisons of best human efforts to automated attempts for the sample dataset; c) analytical methods (including detection cues) capable of discerning one type of messages from another; and, d) although outside of the scope of this article, algorithms re–creating analytical methods and developed language models. Using these steps in the research process as a guideline, we aim to answer the following research questions that enable successful development of techniques for deception detection:

> 1. Does eliciting data in CMC impact the results of deception detection (by humans or algorithmically)?
>
> 2a. How well do human judges perform their task of lie–truth discrimination in CMC?
> 2b. Given two conditions for the task of discernment — a set of truthful and a set of deceptive messages, which ones are human judges better at identifying (with the least amount of false positive and false negatives)?
>
> 3. How do two automated types techniques (machine learning algorithms and linguistic feature extraction) compare among themselves, and to the human "gold standard"?
>
> 4. And, how feasible is the automation of the discovered differences?

In this study we reflect on each necessary step, as we create the sample dataset, and subject it to comparative analysis automatically and with human judges, highlighting practical and conceptual difficulties. The following section provides background on the topic. We review select relevant studies: summarizing the current theory on human abilities to discriminate truths from deception, the recent achievements in automated deception detection, and setting new conceptual points of interest for empirical explorations and automation efforts.

Background

Lie–truth discrimination by human judges

When deceptive behaviors (both verbal and non-verbal) are studies in the fields of interpersonal psychology and communication, respondents are typically asked to distinguish deceptive statements from truthful ones, in a so–called "lie–truth discrimination task". The outcome the research has repeatedly shown is that people are generally not very good at distinguishing between truthful and deceptive statements (Vrij, 2000). On average, when scored for accuracy, people succeed only about half of the time (Frank, *et al.*, 2004). In a meta‶analytical review of over 100 experiments with over 1,000 participants, DePaulo, *et al.* (1997) determined an unimpressive mean accuracy rate of 54 percent, slightly above chance. In this study, we replicate an overall accuracy since human judges serve as a "gold standard" or a benchmark for comparison in the overall task and comparing the rate of "false positive" answer: are respondents more likely to err when stories are truly truthful, or when they are really deceptive.

People may not be that successful in distinguishing lies, nonetheless, present studies that examine communicative behaviors suggest that liars may communicate in qualitatively different ways from truth–tellers. In other words, the current theory specifies that there may be stable differences in behaviors of liars versus truth tellers, and that the differences should be especially evident in the verbal aspects of behavior (Ali and Levine, 2008). Of the three broad categories of perceivable differences — verbal, auditory and visual cues — the visual cues are least reliable (potentially due to distractions and misinterpretations) (Wiseman, 1995). "[L]iars can be reliably identified by their words — not by what they say but by how they say it" (Newman, *et al.*, 2003). In text–based CMC environments, verbal cues might be the only kind of cues available to human judges, and after all, those who intend to deceive have to accomplish their task of through language.

The reasons for systematic differences between truthful messages and deceptive messages have been accounted by the widely accepted four–factor theory of deception (Zuckerman, *et al.*, 1981):

> ... relative to a truthful baseline, deception is characterized by greater arousal, increased emotionality (*e.g.*, guilt, fear of detection), increased cognitive effort, and increased effort at behavioral control. Because message veracity affects these internal psychological states, and because each of these states is behaviorally "leaked," observable behavioral differences are expected. Further, statement validity analysis (see Köhnken, 2004; Vrij, 2000) and reality monitoring (see Sporer, 2004; Vrij, 2000) approaches presume that truthful and deceptive accounts will systematically differ because of differences between true memories and fabricated stories. For example, the language used to describe an authentic memory should be higher in imagery, emotional connotation, and contextual information than that describing an imagined event. [4]

The efforts are focused on finding cues that are unconsciously revealed during communication, the kinds of cues that might "leak" the deceptive character of messages.

> The idea that "statements that are the product of experience will contain characteristics that are generally absent from statements that are the product of imagination" is historically known as Undeutsch Hypothesis ((Undeutsch, 1967) as cited in Fornaciari and Poesio (2011). [5]

Linguistic predictors of deceptive messages

There is a substantial body of research that seeks to compile, test, and cluster predictive cues for deceptive messages. There is no general agreement on an overall reliable set of predictors. Burgoon, *et al.* (2003) state:

> Decades of research have confirmed that there are few indicators of deceit that remain invariant across genres of communication, situations, communicators, cultures, and other features of communication contexts. [5]

Ali and Levin (2008) echo this concern:

> Thus, the published research on verbal clues to deception reports numerous statistically significant differences, but the findings do not seem to replicate across studies. To the extent that systematic linguistic–based deception cues exist, evidence for their existence seems clouded by the presence of situational moderators. [6]

Situational contexts for baseline truthful texts are often drastically different, complicating direct comparisons. For instance, in an analysis of synchronous text–based communication, deceivers produced more total words, more sense–based words (*e.g.*, seeing, touching), and used fewer self–oriented but more other–oriented pronouns when lying than when telling the truth (Hancock, *et al.*, 2008). Compared to truth–tellers, liars showed lower cognitive complexity, used fewer self–references and other references, and used more negative emotion words (Newman, *et al.*, 2003). In the analysis of conference calls in a financial area, Larcker and Zakolyukina (2010) found deceptive statements to have more general knowledge references and extreme positive emotions, and also fewer self–references, extreme negative emotions, as well as certainty and hesitation words. In an interrogation context, Porter and Yuille (1996) found three significantly reliable, verbal indicators of deception (out of the 18 verbal cues derived from Statement Validity Analysis techniques used in law enforcement for credibility assessments): amount of detail reported, coherence, and admissions of lack of memory. In their mock theft experiment study, Burgoon and colleagues (2003) did not show statistically significant differences between deceptive and truthful texts, but they were able to identify a trend from profile plots:

> ... deceivers' messages were briefer (*i.e.*, lower on quantity of language), were less complex in their choice of vocabulary and sentence structure, and lack specificity or expressiveness in their text–based chats. This is consistent with profiles found in non–verbal deception research showing deceivers tend to adopt, at least initially, a fairly inexpressive, rigid communication style with "flat" affect. It appears that their linguistic behavior follows suit and also demonstrates their inability to create messages rich with the details and complexities that characterize truthful discourse. Over time, deceivers may alter these patterns, more closely approximating normal speech in many respects. But it is possible that language choice and complexity may fail to show changes because deceivers are not accessing real memories and real details, and thus will not have the same resources in memory upon which to draw. [7]

With more evidence for statistically reliable combinations of individual verbal cues as indicators under appropriate conditions, human abilities to flag deceptive messages can at least be complemented, if not enhanced, by automated tools based on natural language processing and probabilistic techniques. The most valued cues are those that are the least context–sensitive, and can be subjected to statistical analyses for probabilistic estimates of deception, and hence, most amenable to automation.

Limitations of current automated deception detection

In the last decade that several studies have undertaken the task of automatically identifying linguistic cues, other studies utilize pre–existing predictors of deception for a binary truth–lie text categorization task as another way of automated deception detection. The task is considered to be challenging (DePaulo, *et al.*, 1997). It turns out that combining cues may have good predictive abilities in specified contexts (Vrij, 2000). Zhou, *et al.* (2004) used nine clusters of 27 linguistic cues such as diversity, complexity, specificity, and non-immediacy. When implemented in decision support tools, three standard classification algorithms (neural nets, decision trees, and logistic regression) achieved 74 percent accuracy (Fuller, *et al.*, 2009). Another approach adapted an existing psycholinguistic lexicon, reaching an average 70 percent classifier accuracy (Mihalcea and Strapparava, 2009).

Achieving decent predictive ability comes at a price of non–generalizability. When creating a test dataset by eliciting truthful and deceptive messages from study participants, researchers need to specify the context, the topic, the message format and length, restricting experiments to fairly rigid ‘incubator environments’. For instance, Mihalcea and Strapparava (2009) who specified the context of a debate in which opinions are argued, chose three popular topics, such as the pro–life versus pro–abortion argument, and asked the participants to argue for their true point of view as well as the opposing one. The latter would inevitably be deceptive since it would not match their true beliefs. In our current study, we leave the context and topic of the elicited personal stories intentionally open–ended, simply asking participants for a story rich in details and specifying a suggested length.

Previously discussed studies flag the presence of general deceptiveness within a message. One study is unique in the sense of adding a finer level of granularity — which propositions may be deceptive. Using a corpus of criminal statements, police interrogations and legal testimonies, Bachenko, *et al.* (2008) manually annotated each proposition in text for its truth or falsity. Their classification and regression tree–based automatic tagger performed well on test data (average 68.6 percent recall and 85.3 percent precision) when compared to the performance of human taggers on the same subset. We see a great potential at distinguishing deception within finer discourse levels, such as individual sentences, partial statements, or phrases, and calls for further research in this direction. For instance, an e–mail solicitation from a company representative may be truthful about the company name and location as well as the fact of a promotional sale, but may conceal or obscure the sale conditions or return policies. Our study re–defines the task of deception detection at a within–message level with the goal of identifying what message segments might exactly be deceptive, and in what respect.

Another finer level of granularity in deception detection analysis may come from the gradations of truths and lies. In the context of credibility assessments in law enforcement, participants were asked to use a non–binary distinction by providing either a truthful alibi, a partially deceptive account, a completely false alibi, or a truthful confession regarding theft (Porter and Yuille, 1996). There are currently no automatic deception detection studies, to the best of our knowledge, addressing the possibility of distinguishing discrete, non–binary degrees of truth. Neither do the above discussed automatic detection studies attempt to decipher the nature of deceptive content in texts.

To summarize, the above automated approaches show promise in distinguishing deception, even at somewhat better rates than humans. In this study we further explore the phenomenon of qualitative differences between deceptive and truthful messages, their perception by human judges and linguistic cue predictive abilities under an unspecified context or message topic in CMC environments. Acknowledging the extreme difficulty in the binary prediction task, our study raises the issues of non–binary deception categorization and scales, as well as the categorization of types of deception content. Such questions pose further challenges in the field of automated deception detection, and we attempt to explore various degrees of deception on the truth–deception continuum, and increase our understanding of the phenomenon by categorizing dimensions of context–based variability within messages.

Data collection methods

Deceptive story elicitations (Phase 1)

In Phase 1 of this study, we elicited personal stories using Amazon’s online survey service, Mechanical Turk (www.mturk.com). The data collection instrument requested each respondent to write a rich unique short story which may be completely truthful, or contain some degree of deception. Each contributing respondents rated his or her message with a deception level using a Likhert scale of 1 (completely truthful) through 7 (completely deceptive). Examples below illustrate the two extremes:

> **Example 1 (self–rank 1 = completely truthful):** *"Once I accidentally ate my own tooth. I had a leftover baby tooth that hadn't come out when the adult one grew in — it just got pushed to the inside — and it had been wearing away in my mouth for a while. One day when I was eating some chicken nuggets, I bit down on something hard. I thought it was a crunchy bit of chicken breading, so I chewed it up. A minute later I tasted blood, and realized my gum was bleeding. I gasped and said, "I think I just ate my tooth!" My family, who I was sitting with, were aghast and disgusted. But I was glad that I wouldn't have to have the tooth pulled out by a dentist."*
>
> The contributor comments: *"It was all true, but it is the most bizarre story I can think of from my life, so I thought it would be a fun one to share." (1NU1Y5QJR ...)*
>
> **Example 2 (self–rank 7 = completely deceptive):** *"Today I almost got in a wreck on the highway. A guy in a big jacked up pickup truck was driving stupidly fast, weaving in and out of traffic and being totally. I could see in my mirror that was waving his hands around and flipping people off, so I tried to get out of his way as fast as I could. But someone else changed lanes right at that moment so I couldn't get over, so the truck guy zoomed up and tailgated me, then roared past me on the right. Then he cut back over right in front of me, but there wasn't actually enough room for him to do it, so I had to slam on brakes to keep him from tearing my front bumper off! I'm still kind of shaking a little just thinking about it."*
>
> The contributor explains: *"The whole thing was made up, though plausible based on my highway driving experiences." (16KB1K2WOC ...)* [8]

The seven–point scale establishes gradations within the truth–deception continuum, without imposing any specific values to participants' self–ratings for the categories in between the two extremes. What is a half truth to some is a whole lie to others, as the Yiddish proverb suggests.

Deception detection task (Phase 2)

In Phase 2, another Mechanical Turk task was set up for a different group of participants to read one of the stories (produced in Phase 1) and decide whether the story was truthful or deceptive. Participating perceivers (*i.e.*, judges) were requested to elaborate on their judgments and explain what they believed to be deceptive in the story. This explanatory content serves as the basis for a perceived cue typology. Regardless of perceiver's accuracy in discriminating lies from truths, each sender's message is associated with perceivers' confidence levels in their assessments, which is thought to reflect cognitive elements of deception cues. Respondents provided a confidence score each for their own judgment: from uncertainty to certainty. In total, 900 unique respondents participated in Phase 2, 10 perceivers per each story.

Data analysis methods

Scales and thresholds

We used three different scales: 7–point, 5–point, and binary. Respondents writing the stories (the senders) used a 7–point scale to describe where each story fits on the linear truth–deception continuum. There was no definitive mapping of the shades of truth, deception, or half–truth in the open–ended task: we allowed for respondents to provide their best fit to the scale with minimal labeling on the extremes. This fits with our explorative goals of highlighting practical and conceptual difficulties in non–binary distinctions.

The seven–point linear truth–deception continuum was open to interpretation by the sender: how much truth is there in a lie? To consider possible interpretations we created four different threshold settings. The 7–point scale was converted to a pair of binary truthful vs. deceptive classes for each of the four thresholds (T1–T4), as shown in Table 1. Respondents' self–ranking and mapping to the suggested truth–deception continuum were interpreted in four ways. The T1 setting is the most conservative separation of the stories, and defines as deceptive those stories with a self–rank other than one. It excludes from the truthful group even the slightest deviation from reality (consistent with the Yiddish proverb in the first epigraph to this article). The T2 setting is more inclusive, and also considers stories with self–rank 2 as being truthful, since deceptive content in these stories is still minimal. T3 is the most liberal setting, yet a rank of 3 still reflects the fact that the majority of content is truthful. In the 4th setting (T4), 18 border–line cases (self–rank 3, 4, and 5) were removed as borderline half–truth cases, reducing the total to 72. T1–T4 settings were used to experiment with interpretations of truth and deception since, as the German historian in the second epigraph points out, the notion of 'the truth' is unclear and depends on what the deceiver wishes to hide.

Table 1: Thresholds on the truth–deception continuum.

Truth–deception level	**1 Truthful self–rank**	**2**	**3**	**4**	**5**	**6**	**7 Deceptive self–rank**	**Total**
Number of stories	*39*	*11*	*3*	*9*	*6*	*4*	*18*	***90***
T1 threshold	*truthful (39)*	*deceptive (51)*						***90***
T2 threshold	*truthful (50)*		*deceptive (40)*					***90***
T3 threshold	*truthful (53)*			*deceptive (37)*				***90***
T4 threshold	*truthful (50)*		*removed (18)*			*deceptive (22)*		***72***

When judging a story, perceivers were asked to use a 5–point scale for providing a confidence score for their own judgments (1 = totally uncertain, 2 = somewhat uncertain, 3 = guess, 4 = somewhat certain, 5 = totally certain). The scale is not meant to map to the 7–point scale of degree of deception. In fact, when making judgments about the content of the story, perceivers' task was a simple binary answer to the question: "is the story truthful or deceptive?"

Normalized perception measure: Deception and confidence

Perceivers' binary judgments (truthful vs. deceptive) and confidence ratings (a value of 1=uncertain through 5=certain) collected in Phase 2 were combined to be represented in a single measure where certainty represented the range of 1–5, and judgment represented the valence. Positive valence was assigned to messages perceived to be truthful and negative valence was assigned to messages perceived to be deceptive. For instance, if a responder was somewhat uncertain that the message was deceptive, the data point was converted to a value of -4 If on the other hand, the story were judged truthful, the normalized score would be +4.

Each story was associated with a list of 10 normalized independent judgments in the range of -5 to +5. To obtain a reflection of the overall judgment of responders, the average (mean) and most frequently occurring value (mode) of these normalized score were obtained. These two statistics were used in a Pearson's correlation evaluation to determine whether the normalized response value was indicative of the deceptive level.

Composite 'fooling' success measure

Each story generated 10 judgments that each either correctly or incorrectly identified the deceptive value of story. A composite measure of success 'at fooling' was assigned to each story which is based on the percentage of incorrect judgments from the total judgments. In this way, the success rating reflects the sender's ability to 'fool' the reader into believing a plausible deceptive story, or to write a truthful story in such a way that it would be interpreted as deceptive. A group of stories with a high 'fooling' success rate (80–100 percent of readers guessing incorrectly) was formed and analyzed separately for the prevalence of linguistic cues, since these stories were believed to contain unique properties given their ability to mislead the majority.

Linguistic cues

The content of the 90 stories was analyzed linguistically with the motivation of determining whether different linguistic cues are present between deceptive and truthful story groups. Linguistic cues were identified, calculated, and compared within and across deceptive and truthful story groups, following the work of Zhou, *et al.* (2004). In their study, the occurrence of these linguistic cues was correlated to positive or negative indicators of deception:

Figure 1: Verbal cues and calculation explanations.
a. Average sentence length
b. Average word length
c. Generalizing terms (words indicating a sense of generality)
d. Emotiveness ((total adjectives + total adverbs)/(total nouns + total verbs))
e. Pronoun count (first person singular, first person plural, third person singular, third person plural)
f. Lexical diversity (total unique words/total words)
g. Negative effect (words alluding to negative feelings)
h. Modifiers (total adjectives + total adverbs)
i. Pausality (total punctuation/total sentences)
j. Place terms (words alluding to physical spaces)
k. Sense terms (words alluding to physical senses and feelings)
l. Time terms (words alluding to points in time)
m. Uncertain terms (words alluding to feelings of uncertainty)

Figure 1 contains the inventory of the considered cues and their explanations. The Linguistic Inquiry and Word Count (LIWC) dictionaries were used as standardized word lists, such as adjectives or generalizing terms, and a Python script was created to automatically count and calculate linguistic figures by comparing the text to these lists. From this, a set of values representing each story's linguistic cues served as individual, independent quantitative variables for backward entry logistic regression modeling in SPSS. The aim is to see whether a linear function of variables could predict group membership.

Machine learning

In preparation for machine learning experiments with binary text classification (truthful/deceptive), the data were pre–processed: all tokens were decapitalized, digits and punctuation excluded. Three datasets were created. The first by removing more than 500 of the stop wards prescribed by Lewis, *et al.* (2004) [9]; the second by removing none; and the third by removing only some stop words but retaining point of view words (*e.g., believe, sure, wonder; I, you, us*). Four thresholds (T1–T4, Table 1) were applied to each preprocessed dataset. Two Weka implementations of machine learning algorithms were used — J48 for decision trees (DT) and SMO for support vector machines (SVM) (Witten and Frank, 2005). Each used 'leave–one–out' cross–validations to compensate for small data sizes.

Results

Our dataset contained 90 stories (distributed over seven levels from truth to deception, Table 1), and 900 judgments (each story rated by 10 perceivers).

Human judgments

Each perceiver–s binary truth–deception judgment was combined with the perceiver's confidence value associated with that judgment. Each story was assigned a median of 10 such normalized perception scores.

Overall average accuracy

When messages with only a self–rank of 1 were considered truthful (T1), human raters achieved 50 percent accuracy (454 correct judgments out of 900). In the T2 setting, 57 percent were correct (509 out of 900) and with T3: 58 percent (519 out of 900). When half–truth stories were removed (T4) from the judgment task, performance showed 63 percent accuracy. Adjusting the threshold towards more inclusive truth categories increased accuracy in human judgments. Removal of half–truths improved accuracy further, yet overall, these measures were not significantly above chance performance. This does not preclude that there are distinguishing features (linguistic or other) between deceptive and truthful stories; however, whatever factors may exist were not consciously perceived in the judgment task.

Comparison of accuracies by type of messages being judged

The overall success rate with 50–63 percent accuracy can be further separated using the traditional contingency table in which two access (the actual self–ranked values and perceived values of truth and deception) create true and false positives and negatives. Such separation allows us to see that judges did better with identifying truthful messages correctly than with detecting deception correctly. For instance (see Table 2), out of 39 truthful stories which were judged 10 times each (n=390), only 27 percent were misjudged as deceptive; while with the 51 remaining deceptive stories (n=510) the false negative rate was 67 percent, potentially indicating a truth bias during judgments.

Table 2: Accuracy of human judgments by type of actual self–rank.

Actual truthful		**Accuracy of human judgments based on truthful stories only**		
		Perceived judgments		**Total number of stories**
		Correctly perceived as truthful	**Incorrectly perceived as deceptive (false positive)**	
	Threshold 1 (self–rank 1)	n=286 (73%)	n=104 (27%)	n=390
	Threshold 2 (self–rank 1–2)	n=368 (74%)	n=132 (26%)	n=500
	Threshold 3 (self–rank 1–3)	n=388 (73%)	n=142 (27%)	n=530
Actual deceptive		**Accuracy of human judgments based on deceptive stories only**		
		Perceived judgments		**Total number of stories**
		Correctly perceived as deceptive	**Incorrectly perceived as truthful (false negative)**	
	Threshold 1 (self–rank 2–7)	n=169 (33%)	n=341 (67%)	n=510
	Threshold	n=142	n=258	

	2 (self–rank 3–7)	(36%)	(65%)	n=400
	Threshold 3 (self–rank 4–7)	n=132 (36%)	n=238 (64%)	n=370
	Threshold 4 (self–rank 6–7)	n=86 (39%)	n=134 (61%)	n=220

Writing time

The average authorship time was also investigated as to whether it is also correlated with deceptive level. The analysis is based on the notion that deceptive stories require greater time due to the need to manipulate reality, and invent new information rather than report a factual account. Using the most conservative threshold T1, the 51 deceptive stories were found to be completed in an average of 7.9 minutes, and 39 truthful stories took an average of 10.8 minutes — the reverse of the anticipated outcome.

Table 3: Distribution of mean writing time per level.

Truth–deception level	**1**	**2**	**3**	**4**	**5**	**6**	**7**
T1 Threshold	truthful (39)	deceptive (51)					
Mean writing time	***10.8***	*7.3*	*7.8*	*9.6*	*5*	*8.1*	*8.4*
		7.9					

Comparing these two groups using a means comparison t–test in SPSS (T=1.617, ρ=0.109) showed that no statistically significant difference exists between the assumed time to compose deceptive stories versus purely non–deceptive stories.

Normalized perception measure: Deception and confidence

Perceivers' truth–deception binary judgments and ordinal confidence ratings were combined across all responses. A correlation was sought between the actual and perceived deception based on the deceptive value of the story (1 to 7) and the normalized perception (-5 to +5).

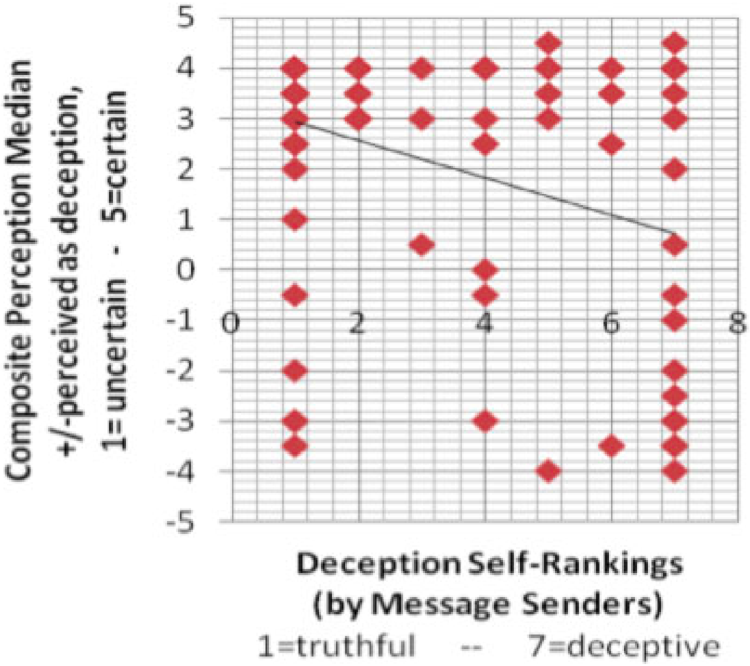

Figure 2: Plot of negative correlation between perceived deception and

confidence level.

Using the Pearson's correlation statistic in SPSS, the test produced a co-efficient value of -3.49 with a confidence ρ=0.008 <0.05, indicating that a statistically significant, weak, negative correlation exists between the two factors. This shows that on average, the higher the actual deception level of the story, the more likely a story would be confidently assigned as deceptive.

Beyond the factors described here, no further independent dimensions were investigated as to whether they correlate to actual or perceived deceptive value. For example, no subject–specific dependency was determined, largely because subject and theme analysis revealed a high dispersion among categories, meaning that categories were not sufficient in size to indicate conclusive findings. Further data collection could indicate differences between actual and perceived deceptive value which are based on other story characteristics, such as subject.

Linguistic cue analyses

Truthful vs. deceptive

Per previous research, the proposed linguistic cues used in stories were expected to differ from truthful to deceptive groups. When truth is assigned a threshold of 1 (T1), no statistically significant difference was found. Further tests were performed for settings (T2–T4), but logistic regression tests remained inconclusive when analyzed according to the 13 cue values (a–m, Figure 1). This indicates that our dataset is linguistically homogenous. Confounding factors such as sampling methods or task definition may account for the difficulty in establishing this differential between groups, since previous work as shown some degrees of correlation between deception and one or a combination of the 13 linguistic cues.

Success 'at fooling'

The sender's ability to fool the perceiver into believing a deceptive story, or to write unbelievable truthful story was analyzed in two groups. The first is those 'high–success' stories (80–100 percent perceivers judging incorrectly) and the second is the remaining stories, which had less than 80 percent guessing incorrectly. Contrary to our expectations, the results of this analysis show that there were no identifiable linguistic distinctions. Logistic regression failed to determine statistically significant verbal predictors with the composite measure of the success 'at fooling'.

Confidence rate

Of the 900 responses, 617 produced confidence ratings of either four or five, indicating that perceivers tend to place high certainty in their own judgments. In fact, all 90 stories generated at least one high confidence response regardless of whether the stories were deceptive or truthful. To test whether there is a relationship between the linguistic cues used in the story, and the prevalence of high–confidence responses, a group of 26 stories was formed consisting of those with eight or more responses with a confidence rating of four or five. As with 'fooling' success rate, logistic regression models were unsuccessful at identifying linguistic predictors for high confidence response stories when compared to the remaining 64 stories.

Machine learning experiments

In Table 4, percent correct values represent the average over all cross–validation runs. Weka's ZeroR classifies all cases into the most abundant class and was used as the baseline. The Matthews correlation coefficient (MCC) was used to verify the results. MCC is a balanced measure of the confusion matrix for uneven sizes (Almeida and Yamakami, 2010), with +1 = perfect prediction, 0 = average random, and –1 =inverse prediction. Four algorithms achieved accuracy higher than the baseline with a moderate, positive MCC value (Table 4, bold), comparable to human accuracies of 50–63 percent in the task.

Table 4: Machine learning experiment results.

Classifier	Stop word removal	Threshold	Zero R %	Correct %	MCC
DT	All	T1	56.67	37.33	-0.02
DT	All	T2	55.56	51.78	0.04

DT	All	T3	58.89	**64.89**	**0.25**
DT	All	T4	69.44	46.11	-0.36
DT	None	T1	56.67	50.67	0
DT	None	T2	55.56	48.89	0
DT	None	T3	58.89	58	-0.11
DT	None	T4	69.44	54.72	-0.17
DT	Partial	T1	56.67	35.33	-0.03
DT	Partial	T2	55.56	50.67	0.03
DT	Partial	T3	58.89	**64.44**	**0.25**
DT	Partial	T4	69.44	44.72	-0.37
SVM	All	T1	56.67	52.22	0.28
SVM	All	T2	55.56	46.67	-0.12
SVM	All	T3	58.89	51.11	-0.08
SVM	All	T4	69.44	69.44	0.03
SVM	None	T1	56.67	**56.67**	**0.36**
SVM	None	T2	55.56	57.78	0.13
SVM	None	T3	58.89	57.78	0.07
SVM	None	T4	69.44	70.83	0.1
SVM	Partial	T1	56.67	**58.89**	**0.35**
SVM	Partial	T2	55.56	51.11	0
SVM	Partial	T3	58.89	57.78	0
SVM	Partial	T4	69.44	68.06	0.06

Content analysis results

We manually content–analyzed the dataset (per Krippendorff, 2004) focusing on what stories have in common, and in what respects they differ. This systematic qualitative account for the variability within the dataset served as an error analysis and resulted in an empirically–derived faceted classification of potentially deceptive messages which varied along five facets (Figure 3).

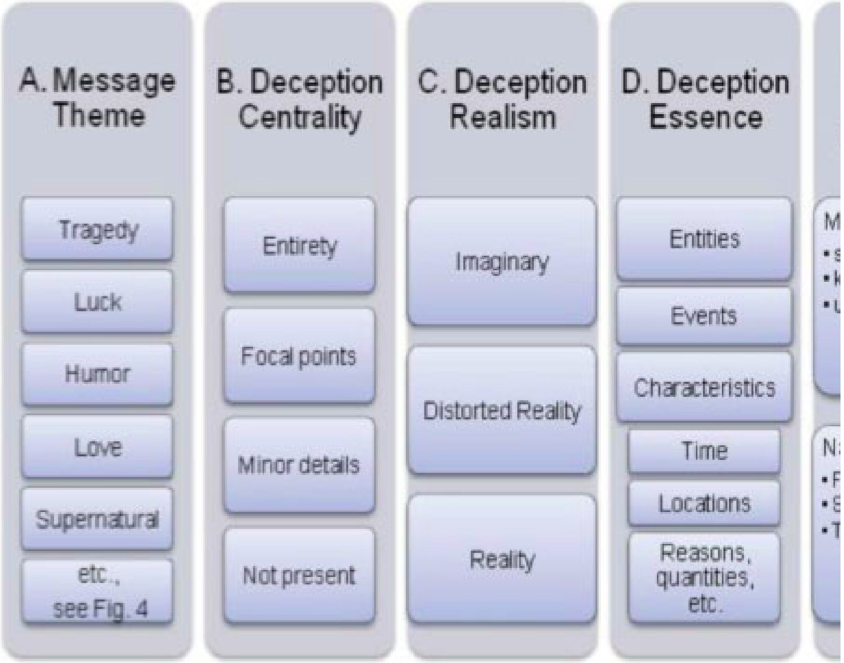

Figure 3: Faceted classification of potentially deceptive message

A. Message themes

Message 'theme' refers to what the message is generally about, or its subject matter. The analysis showed that messages are distributed over 12 thematic categories. For instance, 31 percent of the stories described a tragedy of some kind, 17 percent referred to unexpected luck (Figure 4).

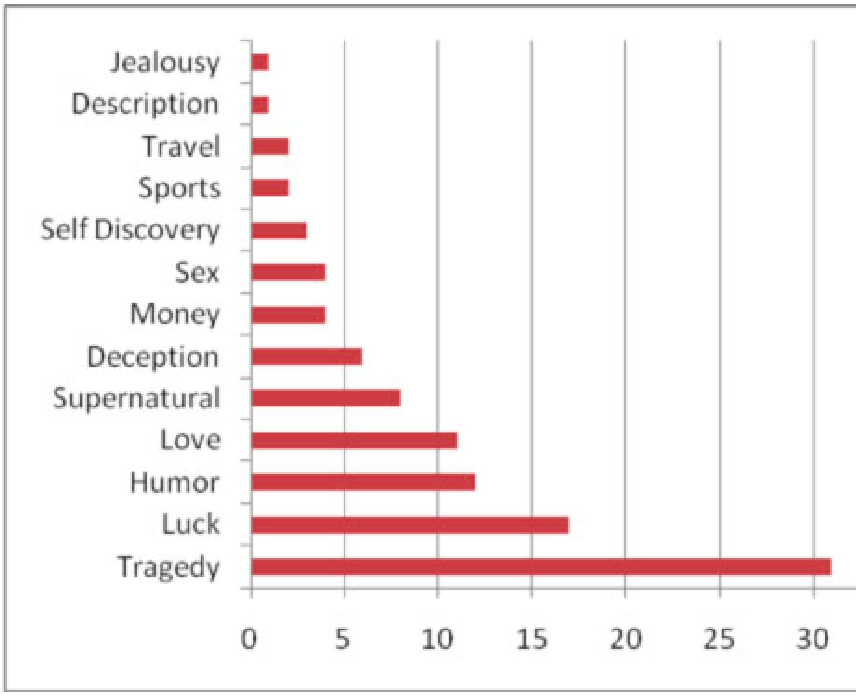

Figure 4: Distribution of message theme categories, percentage.

The most prevalent theme, tragedy, described rare incidents which led to unfortunate outcomes of some kind, as exemplified by a story of a death after an argument:

> **Example 3 (self–rank 2):** *"An African American woman came to me and said, 'I've been watching your tutors working with their students. I have two daughters, older than most of your students. Can they join your program?' I agreed and assigned individual tutors to each of her daughters. Every week or two weeks, this mother would drop by. We would talk of her daughters' progress. ... Three months later, as I was talking to another parent who knew this woman, I mentioned that the latter's daughters were absent from the program that day. The parent exclaimed, 'Haven't you heard? She's dead. Her son came around for some money. They argued, and he shot her.' Still incredulous, I saw the sad news on TV that evening. I could not believe that a woman I have been talking regularly is no longer here. ..." (1CFJUOUJ5N ...)*

Luck was typically described as an unexpected fortune such as winning or finding something of value:

> **Example 4 (self–rank 1):** *"Once, when i (sic.) [10] was walking down the street i came across a $100 bill. It was just lying there, on the side of the road. I looked around, and saw no one, so i picked it up. I was so excited to have found some money. I hurried home to tell my husband what i had found. He didn't believe me, until i showed it to him." (1C8TZBLO1T ...)*

The identified categories are not necessarily mutually exclusive: 134 annotations were made showing that stories often consist of more than one theme. For instance, in some examples evidence of luck is in the avoidance of tragedy. Money–related stories were not always about fortunes:

> **Example 5 (self–rank 7):** *"I am sick of being no body. I work hard, be a good son, good husband, great father but I have not done anything great other than making the ends meet. At the age of 58, I just want to rob a bank or steal loads of cash from the big shots. Sometime, I dream that god will give me a chance of finding 50 million dollars on the road or in my back yard just for me. ... So, I am planning to land on one of the millionaire's building from a chopper and steal the money. No traces are behind and this will become a challenge to the police. At the end, I feel sorry for what I have done and use the money for a good cause and give charity." (1Z4GT47GVP ...)*

Supernatural themes included incidents of hearing voices or of alien abductions. The deception theme described those stories where the author is said to experience lying or being lied to. The rest of the less populous theme categories are named according to their subject matter: humor, sex, sports, or jealousy. Description was a distinct category since it did not offer a plot but rather described the characteristics of a setting or situation:

> **Example 6 (self–rank 1):** *"... The place is ancient, has weathered many a hurricane, and has settled so much that some of the doors are now cut into trapezoids rather than rectangles. The room I like the most is a long skinny one upstairs that I think used to be a closet. ... The common kitchen is full of roaches ... and the common bathrooms are not air conditioned at all. ..." (1MFHWXB9ED)*

There were six cases of deception message themes (Figure 4) within truthful stories. Speaking truthfully about lying seems to create some confusion among human perceivers and add complexity to detection automation. Judges resort to guessing, as in the following example, which was judged on average as "I'm guessing it is truthful" 10 human judges (*i.e.*, normalized perception score of "+3").

> **Example 7:** *"So the other day I'm sitting at home, and I get a call from a carpet cleaner service that's offering deals on carpet cleaning. I don't need my carpets cleaned, but I decide to yank their chain a little bit. I ask the guy 'Can you get out blood?' He says 'Is it just a little bit?' and I tell him 'Oh no, its quite a bit. Like someone bleed to death. Oh, and if you clean it up, I'll pay you double, but you have to promise not to call the police.' He hung up. Bastard."*
>
> The contributor explains: *"It's a true story. I tried to make a carpet cleaner company believe I killed someone and wanted them to clean up the blood and they probably didn't buy it." (1X5WSQPFM8 ...)*

This is a complex case in which the judges task may have an increased cognitive load due to the nested question — is he lying about lying? Linguistic cues may be misleading in analysis since the direct quotes are used to in the story to render truthfully how the participant tried to convince create a false belief in the carpet cleaner's mind.

In sum, the identified message themes orient the reader to a topic and provide a thematic landscape for our dataset. Compared to Rubin's (2010b) domains identified in blogs, themes are narrower and mostly fall within Personal Relations or Finances & Insurance domains. They span across everyday and serious lies, as in DePaulo (1994). Message theme variability poses distinct problem for both automation and human discernment.

The next four Facets (B–E, Figure 3) address properties of deception within messages, taking into account the variable interpretation of truth (T1–T4, Table 1). Facets B–E arise from the content analysis of the messages, senders' accompanying explanations, and perceivers' reflections.

B. Deception centrality

Deception centrality refers to what proportion of the story is deceptive, and how significant the deceptive part is. Messages vary from being entirely deceptive, to being deceptive in its focal point or in a minor detail (Figure 3, B). Of the 90 messages, only 18 messages were confirmed by senders as being deceptive in their entirety. In the following example the sender claims all but minor details to be true:

> **Example 8:** *"I've only bought dark lipstick, two tubes of it, and not from the beauty store." (12NG4N9H36 ...)*

C. Deception realism

Deception realism refers to how much reality is mixed in with the lie. A message can be based predominantly on reality with a minor deviation, or can be based on a completely imaginary world Figure 3, C). Out of 90, 41 senders claimed in their verbal explanations that their messages were nothing but the truth. The entirely deceptive stories (self–ranked 7) were often fiction–like:

> **Example 9 (self–rank 7):** *"about a week ago i ran into this guy — he was good looking. He also was a fast talker. I told him to leave me alone because I was married but, he said he did not care. Well the man ended up following me home. I ran in the house and told my husband I was being stalked. He then proceeded to call 911. 911 then told him an officer would respond. The guy was so nuts he came up and knocked on my door. My husband answered and then there was an altercation they began fist fighting. Needless to say the police showed up and took them both to jail."* The contributor explains: *"... just something i made up off the top of my head." (ID=1LFVZ5N9IL ...)*

The remaining categories were based on some degree of distortion from the reality.

D. Deception essence

Deception essence refers specifically to what the deception is about, its nature (Figure 3, D), not to be confused with message theme or topicality. When explaining his truthful message, one of the senders felt compelled to clarify that his story was true in many respects, verbalizing how he could have lied, in principle, about the person or events:

> **Example 10:** *"Nothing was deceptive. The roommate's name was John. All the other events did happen." (1ZO1SPGVAZ ...)*

Similar testimonies lead us to believe that message senders are obviously aware of the underlying possibilities. We subdivided these types of deception essence into events, entities (a collective term for people, organizations, and objects), and their characteristics (referring to qualifiers of both, events and entities). The rest of the essence categories (time, location, reason, degree, amount, etc., Figure 3, D.) conceptually align with content questions such as when, where, why, how much, etc.

We hypothesize that certain combinations of centrality (focal point), realism (reality–based), and essence (events), are more re–current than isolated uses of deception topics, and thus deserve special consideration in deception detection efforts. This is subject to future testing and targeted elicitations. The deceptive piece below describes how distortions of reality are details of events which are, nevertheless, focal to the message:

> **Example 11 (self–rank 5):** *"i moved to canada when i was 23 years old. i went to study at a local canadian university for my bachelor's degree. on this trip i met a guy i fell in love with. after i had to return to my home country in europe our contact broke. i*

> *still consider him the biggest love of my life. i am 30 yrs old now."* The contributor explains: *"i went to canada when i was 24, i went to work, not to study. i didn't fall in love with anyone and i didn't break contact with anyone after return." (1EYU7H6RY6 ...)*

Similarly, the sender who wrote Example 2 reveals the reality–based distortion (Facet C) of the event as the focal point (Facet B):

> **Example 12:** *"I didn't have an accident on the way home." (1UWFAQB2PS ...)*

Thus, each deception essence (event, entity, characteristics, etc.) may vary by the orthogonal deception facets, either in its centrality to the message (such as focal point or minor detail) or in its degree of realism. In the case below, focal events are true but the entity (the man) is imaginary:

> **Example 13:** *"This type of event has actually happened, but the specific story of the man from the midwest was something I made up." (1XIQHS0LU6 ...)*

E. Self–distancing degree

The distance between the message sender and plot characters transpired as variable dimension across stories, created by misattributions (revealed by liars afterwards) and by narrator's perspective (revealed by writing stories from the first or third person). Misattributions were three–fold. Authors' self–involvement was often misattributed (34 percent), where the events did happen but not to the sender. Events are also often misattributed to known entities (*e.g.*, friends, co–workers; 36 percent) or to unknown ones (30 percent) as follows:

> **Example 14:** *"The story behind the song is true (according to Rod), but it was not MY aunt Maggie. I don't have an aunt Maggie." (1LVQDBTE57 ...)*

Narrator's voice was also distinguished into three categories in our dataset: first person (88 percent), second (eight percent), and third (one percent), the latter two signifying distancing from the message.

Stated perceived deception detection cues

Out of 900 judgements, 275 generated verbal explanations describing cues by which perceivers said they judged deception. The cues fell into four categories of this data–driven, perceived cue typology: world knowledge, logical contradiction, linguistic evidence and intuitive sense.

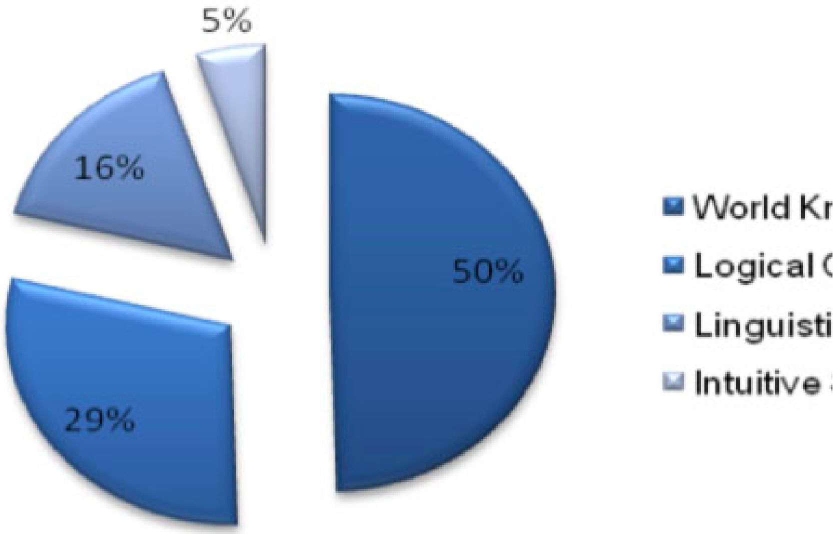

Figure 5: Stated perceived deception detection cues.

Of the 275 stated perceived cues, half (Figure 5) belonged to perceivers' world knowledge, including personal experiences:

> **Example 15:** *"plastic surgery does not involve covering some one from head to toe with a plastic sheet, with a hole cut out around my eye." (15I8RA20KV ...)*

Twenty–nine percent of explanations pointed to a logical contradiction:

> **Example 16:** *"Why would someone steal a cellphone and the elderly woman had a cane and was stooping so presumably it would be hard for her to snag a phone unless the cane and the stoop were an act as well." (1CXLIK31TX ...)*

Sixteen percent vaguely referenced linguistic evidence, for instance, in regards to a story about cricket:

> **Example 17:** *"Some of the grammar and terminology does not make sense." (1KRHRTLZ9R ...)*

Five percent openly stated a decision based on intuition, relying on hunches, or impressions outside of empirical evidence. The perceivers tended not to be very descriptive in unpacking their reasons around their sense of deception:

> **Example 18:** *"It may be a joke. Doesn't seem to be a true story" (1BC2H6DL3L ...);* or **Example 19:** *"It just doesn't seem very believable." (12LMYHR57T ...)*

Discussion

With the elicitation methods and experiment set–up detailed above, human judges achieved on average overall 50–63 percent success rates, depending on what is considered deceptive on the truth–deception continuum. Judgments of deceptive stories produced more false negative results that judgments of truthful stories, potentially indicating the judges' truth bias in performing the task.

One significant finding is that actual self–ranked deception levels negatively correlate with their confident judgment as being deceptive (r = -0.35, df = 88, ρ = 0.008). In other words, the higher the actual deception level of the story, the more likely a story would be confidently assigned as deceptive. This finding is consistent with the Undeutsch Hypothesis and current theory on qualitative differences in deceptive texts. In addition, it emphasizes the value of human judges' confidence level in relation to the degree of deceptiveness as self–declared by liars. Such combination makes the trend obvious, implying that even though humans are not as good at detecting overall deception; its extreme degrees are more transparent and thus obvious to judges. It would then make sense to focus further research on defining each category on the truth–deception continuum, and consequently eliciting more refined data. What would most people consider a half–truth as opposed to an absolute lie?

The highest–performing machine learning algorithms in our experiments reached 65 percent accuracy, but linguistic cues as tested with degrees of deceptiveness, were not found not to be significant predictors of deception level, confidence score, or an authors' ability to fool a reader. This supports the idea that further methods are needed to combine isolated predictors into more complex constructs.

Areas for improvement

We further reflect on conceptual and methodological challenges in elicitation and analytical methods, and identify potential areas of improvements.

The first challenge arises from interpretations of what constitutes deception. Half–truths and half–lies inevitably interfere with binary classification by humans. In general, humans performed better when there were fewer ambiguous cases (*e.g.*, T4). It would be beneficial if people could defer to a machine–learning prediction in cases where they lack confidence, thus future research is needed on finding areas in which computational techniques supersede human judgments that lack confidence.

Second, our data elicitation task was open-ended, allowing writers to make choices and us to charter their preferences and existing case scenarios. This approach served an excellent purpose for surveying potential context and classifying deception facets (such as theme, centrality, realism, essence, self–distancing). The heterogeneous data, however, lessened linguistic predictive power. Previous computational attempts succeeded with well–defined tasks for eliciting deception about, for example, an opinion entirely contrary to one's belief (Mihalcea and Strapparava, 2009) or the most significant person in one's life (Hancock, *et al.*, 2008). In addition, since qualitative types of deception do not easily translate to linear deception level scales, we propose further research focuses on the following categories: events, entities or characteristics as a focal point of the story.

Third, respondents' awareness of the study goals may have prompted them to apply extra effort in 'fooling' perceivers with truthful yet unbelievable or bizarre stories. This may have morphed the original intent of elicitations to an even broader, yet realistic, task.

Two other broader challenges are associated with the experimental design. The relative unimportance of linguistic cues (16 percent) is evident from our proposed cue typology. Contrary to the majority of current computational efforts at the lexico–semantic level (*e.g.*, LWIC-based), this finding resonates with concerns previously expressed in psychology in regards to how people detect deception in real life (Park, *et al.*, 2002). Based on their survey, Park and colleagues found that only two percent of the lies are caught in real–time and are never purely on the basis of verbal and non–verbal behavior of the sender.

> *"Most lies were detected well after the fact, using information from third parties and physical evidence to catch liars. Some liars later confessed or let the truth slip out. Sometimes the lie was simply inconsistent with prior knowledge. None of these common discovery methods are available to judges in typical deception detection experiments. Perhaps people are not very accurate in deception experiments because deception detection experiments incorrectly presume that deception is detected based on real–time leakage, since experiments fail to accurately capture the ecology and process of deception detection."* (Levine, *et al.*, 2011)

On the other hand, if verbal cues are still detected objectively — based on data at hand and in real–time with some predictive power — automatic technique would have an upper hand over human efforts, ultimately proving invaluable, especially in CMC environments.

Finally, we offer an insight from the language pragmatics perspective. Human communication, including CMC, involves complex social phenomenon consisting of an interplay of language, shared frames of reference, and culturally specific contextual knowledge in order to convey meanings between the sender and receiver. A thorough understanding of deception, therefore, must account for the pragmatic use of language. Deception involves communicative action oriented toward reaching understanding, and in such action, the sender and perceiver must share particular conventions, expectations or presuppositions about the communicative exchange, namely that the sender's "utterances are justifiable in relation to interpersonally accepted roles and norms" (Mitchell, 1996). As a result, deception paradoxically requires "a shared system of interpretation and meaning". Mahon (2008) identifies the "addressee condition", which is that when lying it is not merely the case that the person who makes the untruthful statement intends that some other person believe that untruthful statement to be true, but that the person intends that the *perceiver* believe that untruthful statement to be true. Without this relationship, the author is not subject to the cognitive demands associated with deception in more intimate scenarios. In our experiments, each sender wrote an unverifiable story for an anonymous recipient, with possibly inaccurate expectations about how the message would be received; each responder relied on an imaginary author and a subjective schema of the events being described. Such elicitation conditions reduce the possibility of shared linguistic conventions and contextual knowledge that deception demands, and may undermine the ability of the data to meet the requirements of deception, or at least separate truth in a significant sense from deception as it occurs in a real–world communication scenario.

Conclusions and future work

Deception detection is a challenging task. With pervasive use of text–based computer–mediated communication, automated deception detection continues to increase in importance in natural language processing, machine–learning communities and broader library and information science and technology. Tools with detection capability can support undisrupted communication and information practices, credibility assessments, and decision–making. Our study shows typical success rates by humans (50–63 percent) and machine–learning algorithms (65 percent). We established that the higher the actual deception level of the story, the more likely a story would be confidently assigned as deceptive, and have seen a distinct truth bias in human judgments. However, we were unable to find significant linguistic predictors of deception in our sample, explained, in part, by the challenges encountered. We reflect upon reasons, and offer a systematic content analysis of elicited stories and surrounding senders' and perceivers' explanations. Such error analysis resulted in a proposed facetted classification of variance within deception by five facets: theme, centrality, realism, essence and self–distancing. Our next step is to devise ways of exerting greater control over variability, increase and even out sample sizes, and improve measurement scales. We are considering changes to our experimental design and the elicitation task in view of encountered challenges, discussed at length in this paper.

About the authors

Victoria L. Rubin is the Principal Investigator of the Language and Information Technologies Research Lab and an Assistant Professor at the Faculty of Information and Media Studies at the University of Western Ontario, London, Canada. She received her Ph.D. in Information Science and Technology in 2006, her M.A. in Linguistics in 1997 from Syracuse University, N.Y., and a B.A. English, French, and Interpretation from Kharkiv National University, Ukraine in 1993. Dr. Rubin research interests are in information organization and information technology. She specializes in information retrieval and natural language processing techniques that enable analyses of texts to identify, extract, and organize structured knowledge. Victoria Rubin is the corresponding author and can be contacted at: vrubin [at] uwo [dot] ca

Niall Conroy is a doctoral student in LIS at the Faculty of Information and Media Studies at the University of Western Ontario, London, Canada. He has received B.Sc. in Computer Science and an M.LIS, and has a background in software engineering and information management. His current research interests include intelligent music information retrieval systems, collaborative filtering and multimedia data mining techniques, and online communities.

Acknowledgments

This research is funded by the New Research and Scholarly Initiative Award (10–303), entitled *Towards Automated Deception Detection: An Ontology of Verbal Deception Cues for Computer–Mediated Communication* from the Academic Development Fund at the University of Western Ontario. Special thanks to Jacqueline Burkell for her methodological insights, to Martha Stortz for her efforts in data coding and classification, and to Jeff Taylor for running the WEKA experiments.

Notes

[1] The work of our colleagues in speech community is not reviewed here due to our primary interest in texts and verbal (not auditory) cue found in CMC messages.

[2] Burgoon, *et al.*, 2003, p. 91.

[3] The terms "lying" and "deceiving" are used interchangeable in this article, though we are aware of the fact that deception can be accomplished in various ways (such as equivocating, misrepresenting and evading), and "lying" as "prevaricating" is a distinctly narrower term (see Rubin [2010] for the discussion of various kinds of deception).

[4] Ali and Levine, 2008, p. 83.

[5] Burgoon, *et al.*, 2003, p. 91.

[6] Ali and Levin, 2008, p. 84.

[7] Burgoon, *et al.*, 2003, p. 96.

[8] Identification numbers are listed for all examples matching database entries for further inquiries.

[9] The list is available at http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop.

[10] The participants' spelling and capitalization are left uncorrected to preserve their authentic CMC writing style and habits.

References

M. Ali and T. Levine, 2008. "The language of truthful and deceptive denials and confessions," Communication Reports, volume 21, number 2, pp. 82–91.http://dx.doi.org/10.1080/08934210802381862

T. Almeida and A. Yamakami, 2010. "Content–based spam filtering," *Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7.

J. Bachenko, E. Fitzpatrick and M. Schonwetter, 2008. "Verification and implementation of language–based deception indicators in civil and criminal narratives," *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 41–48, and at http://www.aclweb.org/anthology/C08-1006, accessed 3 March 2012.

D. Buller and J. Burgoon, 1996. "Interpersonal deception theory," *Communication Theory*, volume 6, number 3, pp. 203–242.http://dx.doi.org/10.1111/j.1468-2885.1996.tb00127.x

J. Burgoon, J. Blair, T. Qin and J. Nunamaker, 2003. "Detecting deception through linguistic analysis," *ISI '03: Proceedings of the 1st NSF/NIJ conference on Intelligence and Security Informatics*, *Lecture Notes in Computer Science*, volume 2665, pp. 91–101.

CBC, 2011. "Truth Goggles interview," *CBC Radio* (25 November), at http://www.cbc.ca/day6/blog/2011/11/25/interview-truth-goggles/, accessed 3 March 2012.

B. DePaulo, 1994. "Spotting lies: Can humans learn to do better?" *Current Directions in Psychological Science*, volume 3, number, pp. 83–86.

B. DePaulo, K. Charlton, H. Cooper, J. Lindsay and L. Muhlenbruck, 1997. "The accuracy–confidence correlation in the detection of deception," *Personality and Social Psychology Review*, volume 1, number 4, pp. 346–357.http://dx.doi.org/10.1207/s15327957pspr0104_5

A. Eisenberg, 2011. "Software that listens for lies," *New York Times* (3 December), at http://www.nytimes.com/2011/12/04/business/lie-detection-software-parses-the-human-voice.html, accessed 3 March 2012.

European chapter of the Association for Computational Linguistics [EACL], 2012. "Call for papers: EACL 2012 Workshop on Computational Approaches to Deception Detection," at http://www.chss.montclair.edu/linguistics/DeceptionDetection.html, accessed 20 January 2012.

T. Fornaciari and M. Poesio, 2011. "Lexical vs. surface features in deceptive language analysis," *Proceedings of the ICAIL 2011 Workshop: Applying Human Language Technology to the Law*, pp. 2–8, at http://wyner.info/research/Papers/AHLTL2011Papers.pdf, accessed 3 March 2012.

M. Frank, T. Feeley, N. Paolantinio and T. Servoss, 2004. "Individual and small group accuracy in judging truthful and deceptive communication," *Group Decision and Negotiation*, volume 13, number 1, pp. 45–59.http://dx.doi.org/10.1023/B:GRUP.0000011945.85141.af

C. Fuller, D. Biros and R. Wilson, 2009. "Decision support for determining veracity via linguistic–based cues," *Decision Support Systems*, volume 46, number 3, pp. 695–703.http://dx.doi.org/10.1016/j.dss.2008.11.001

J. Hancock, L. Curry, S. Goorha and M. Woodworth, 2008. "On lying and being lied to: A linguistic analysis of deception in computer–mediated communication," *Discourse Processes*, volume 45, number 1, pp. 1–23.http://dx.doi.org/10.1080/01638530701739181

G. Köhnken, 2004. "Statement validity analysis and the detection of the truth," In: P. Granhag and L. Strömwall (editors). *The detection of deception in forensic contexts*. Cambridge: Cambridge University Press, pp. 41–63.

K. Krippendorff, 2004. *Content analysis: An introduction to its methodology*. Second edition. Beverly Hills, Calif.: Sage.

Language & Information Technology Research Lab, 2012. "Mission statement," at http://publish.uwo.ca/~vrubin/lab/index.html, accessed 21 January 2012.

D. Larcker and A. Zakolyukina, 2010. "Detecting deceptive discussions in conference calls," *Stanford University Rock Center for Corporate Governance Working Paper Series*, number 83 and *Stanford GSB Research Paper*, number 2060.

T. Levine, S. Mccornack and H. Park, 2011. "Deception research at Michigan State University," at https://www.msu.edu/~levinet/deception.htm, accessed 28 May 2011.

D. Lewis, Y. Yang, T. Rose and F. Li, 2004. "RCV1: A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, volume 5, pp. 361–397.

J. Mahon, 2008. "The definition of lying and deception," *Stanford Encyclopedia of Philosophy*, at http://plato.stanford.edu/entries/lying-definition/, accessed 3 March 2012.

R. Mihalcea and C. Strapparava, 2009. "The lie detector: Explorations in the automatic recognition of deceptive language," *Proceedings of the ACL–IJCNLP 2009 Conference Short Papers*, pp. 309–312.

R. Mitchell, 1996. "The psychology of human deception," *Social Research*, volume 63, number 3, pp. 819–861.

M. Newman, J. Pennebaker, D. Berry and J. Richards, 2003. "Lying words: Predicting deception from linguistic styles," *Personality and Social Psychology Bulletin*, volume 29, number 5, pp. 665–675.http://dx.doi.org/10.1177/0146167203029005010

B. Pang and L. Lee, 2008. "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, volume 2, numbers 1–2, pp. 1–135.

H. Park, T. Levine, S. Mccornack, K. Morrison and M. Ferrara, 2002. "How people really detect lies," *Communication Monographs*, volume 69, number 2, pp. 144–157.http://dx.doi.org/10.1080/714041710

S. Porter and J. Yuille, 1996. "The language of deceit: An investigation of the verbal clues to deception in the interrogation context," *Law and Human Behavior*, volume 20, number 4, pp. 443–458.http://dx.doi.org/10.1007/BF01498980

V. Rubin, 2010a. "Epistemic modality: From uncertainty to certainty in the context of information seeking as interactions with texts," *Information Processing & Management,* volume 46, number 5, pp. 533–540.http://dx.doi.org/10.1016/j.ipm.2010.02.006

V. Rubin, 2010b. "On deception and deception detection: Content analysis of computer–mediated stated beliefs," *Proceedings of the American Society for Information Science and Technology*, volume 47, number 1, pp. 1–10.http://dx.doi.org/10.1002/meet.14504701124

V. Rubin, 2009. "Trust incident account model: Preliminary indicators for trust rhetoric and trust or distrust in blogs," *Proceedings of the Third International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, pp. 300–303, and at http://www.icwsm.org/2009/, accessed 3 March 2012.

V. Rubin, 2007. "Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements," *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 141–144.

V. Rubin and E. Liddy, 2006. "Assessing credibility of Weblogs," *AAAI Symposium on Computational Approaches to Analyzing Weblogs*, at http://aaaipress.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-038.pdf, accessed 3 March 2012.

V. Rubin, J. Burkell and A. Quan–Haase, 2011. "Facets of serendipity in everyday chance encounters: A grounded theory approach to blog analysis," *Information Research*, volume 16, number 3, at http://informationr.net/ir/16-3/paper488.html, accessed 3 March 2012.

V. Rubin, J. Stanton and E. Liddy, 2004. "Discerning emotions in texts," *AAAI Symposium on Exploring Attitude and Affect in Text*, at http://www.aaai.org/Library/Symposia/Spring/ss04-07.php, accessed 3 March 2012.

S. Sporer, 2004. "Reality monitoring and detection of deception," In: P. Granhag and L. Strömwall (editors). *The detection of deception in forensic contexts*. Cambridge: Cambridge University Press, pp. 64–102.

U. Undeutsch, 1967. "Beurteilung der glaubhaftigkeit von aussagen [Veracity assessment of statements]," In: In U. Undeutsch (editor). *Handbuch der Psychologie*, volume 11: *Forensische Psychologie*. Göttingen: Hogrefe.

A. Vrij, 2000. *Detecting lies and deceit: The psychology of lying and the implications for professional practice*. New York: Wiley.

R. Wiseman, 1995. "The megalab truth test," *Nature*, volume 373, number 391, p. 391, and at http://www.nature.com/nature/journal/v373/n6513/abs/373391a0.html, accessed 3 March 2012.

I. Witten and E. Frank, 2005. *Data mining: Practical machine learning tools and techniques*. Second edition. Boston: Morgan Kaufman.

A. Wyner and K. Branting, "Preface," *Proceedings of the ICAIL 2011 Workshop: Applying Human Language Technology to the Law*, p. ii, and at http://wyner.info/research/Papers/AHLTL2011Papers.pdf, accessed 20 January 2012.

L. Zhou, J. Burgoon, J. Nunamaker and D. Twitchell, 2004. "Automating linguistics–based cues for detecting deception in text–based asynchronous computer–mediated communications," *Group Decision and Negotiation*, volume 13, number 1, pp. 81–106.http://dx.doi.org/10.1023/B:GRUP.0000011944.62889.6f

p>M. Zuckerman, B. DePaulo and R. Rosenthal, 1981. "Verbal and nonverbal communication of deception," *Advances in Experimental Social Psychology*, volume 14, pp. 1–59.http://dx.doi.org/10.1016/S0065-2601(08)60369-X

Editorial history

Received 25 January 2012; accepted 2 March 2012.

Discerning truth from deception: Human judgments and automation efforts
by Victoria L. Rubin and Niall Conroy
First Monday, Volume 17, Number 3 - 5 March 2012
http://firstmonday.org/ojs/index.php/fm/article/view/3933/3170
doi:10.5210/fm.v17i3.3933