Using Natural Language Processing to Identify the Rhetoric of Deception in Business and Competitive Intelligence Email Communications

by

Jeanna E. Cooper

M.S., Competitive Intelligence Systems, Robert Morris University, 2008
M.S., Computer Information Systems, University of Phoenix, 2003
B.S., Science, The Pennsylvania State University, 1999

Submitted to the Graduate Faculty of the
School of Communications and Information Systems
in partial fulfillment of the requirement for the degree of

Doctor of Science in
Information Systems and Communications

Advisor, Daniel R. Rota, Ph.D.

Committee Member, A.J. Grant, Ph.D.

Committee Member, Robert J. Skovira, Ph.D.

External Reader, Michelle G. Hough, D.Sc.

Student Reader, Darryl M.Husenits, M.B.A.

Student Reader, Xiaolu Li, M.A.

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

This study extended previous work that identified linguistic-based cues (e.g. an increased use of modifiers), through the use of natural language processing techniques, as effective discriminators between truthful and deceptive messages in negotiative email communications. The extension of the previous research included linguistic-based cues as they related to management informative tasks comprised of business and competitive intelligence information, rather than negotiation-based information sharing. A follow-up survey captured demographic and socioeconomic and sociocultural information about participants. Such types of information have previously been shown as having the potential to influence participants' language use. The results of the study indicated that many of the linguistic-based cues that effectively discriminate between deceptive and truthful message senders differ in both type, and in many cases, direction, by genre discourse (negotiative versus informative), though some commonalities were found. Participants in this study also showed differences in affective characteristics associated with their levels of comfort in composing either truthful or deceptive messages. The confidence levels of participants also varied by management information types. Finally, when considering demographic (sex) and socioeconomic (income level) variables across informing conditions (truthful, deceptive), differences were found within several of the linguistic-based cues both independently and as a result of the moderating effects of sociocultural and/or demographic variables. These findings provide evidence of 1) the need to consider the genre of the writing when trying to detect deception; 2) the potential affective influences of deceptive writing tasks; 3) an area in higher education, forecasting, which may need to be remediated; and 4) the need to consider

demographic and sociocultural information in the selection of cues to apply in detecting deception in business and competitive intelligence emailed communications.

ACKNOWLEDGEMENTS

My program places its emphasis upon the journey rather than some final destination. I strongly agree with this notion, and see my ability to make this journey as a result of the influence and support of others. It is without such influence and support that can cause the journey to end too early or to never to begin at all. It is for these "others" this section is written, to acknowledge that without their support and influence, even my first steps down this road may never have been taken. For this (and so much more) my sincerest and heartfelt gratitude goes out: To my parents, Joseph & Judith Cooper, I remain forever grateful for their unconditional support during my time in this program as well as their patience and the encouragement they afforded my curiosity and independent thinking from the time I was young, their love of learning, as well as their strong work ethic and the example they set for the use of dogged determination through difficult times, for without their influence, I may not have had the tools to forge ahead. To my partner, Cheryl Tallent, who, when told that I wanted to take this journey said, "wither thou goest, I will go" and matched her step with mine down this road, hers was the one whose steadfast support and daily practice of acta non verba which made it possible for me to focus on my studies. To my mentors, Dr. Margaret Signorella and Dr. Jeanne Amlund, who passed on to me their strong love of research, strongly encouraged me to take this path, and gave me the time I needed to do it and their time when answering every panicked call. To my advisor, mentor, and friend, Dr. Daniel R. Rota, who always made himself available to guide me and provided sound direction and support on both my project, and life in general. To my committee members, Dr. Robert J. Skovira, Dr. A. J. Grant, and Dr. Michelle Hough, each of which who afforded me their own special insight, support, and assistance on the various aspects of a broad project. To my readers, Darryl Husenits

and Xiaolu Li, who spent hours reading and commenting on this work, this work would never have been done without you. To my cohort in the program who questioned, prodded, argued, and supported, and made this journey so much fun. To Lee Steen who spent a significant amount of time building ParseLee out of the goodness of his heart so I could actually analyze my data. And, to the whole team at Connexor, who not only supplied me with the software for data analysis, but who also provided support for the software and guidance on selecting the right modules.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION/RATIONALE

Framework

In 2008 we find ourselves inundated with information. These informational resources afford us the opportunity to make better, and more informed, decisions in both our individual and business lives. However, this vast influx of information also increases the likelihood of our making decisions based on inaccurate information. Inaccurate information that is integrated into decision-making processes has the potential to yield dire results (Biros, George, & Zmud, 2002) in nearly all facets of society. This problem is one that is both pervasive, in terms of both individual's interactions and organization's processes, and expansive, with respect to its reach throughout the world. From the early years of our nation's history through 2008 there have been numerous reports of false information and its ill effects across a variety of fields. Examples of such reports include:

- The effects of disinformation on military tactics and strategies, of which General William Techumseh Sherman was one, of many, to use over the years in his fight to take Atlanta during the Civil War. In the late summer of 1864, Sherman had some 70,000 troops outside the northern tip of Atlanta, but after many months of strikes he had been unable to force a surrender of the Confederacy, under the command of General John Bell Hood. Finally, on Thursday, August 25th, General Sherman decided to try using a deceptive ploy and he moved all of his troops in the night out of the sight of the Confederacy. Several days passed and the people of Atlanta, as well as many of the soldiers, became increasingly convinced that Sherman and his troops had retreated. This was supported by

articles in newspapers and reports of sightings of battalions of soldiers marching away from Atlanta. As the new week began, more people began to believe that Sherman was gone, until Sunday, August 28th, when Sherman's troops arrived southwest of Atlanta and took out the railways. From there they moved into Atlanta taking General Hood and the city by surprise. By Friday, September 2nd, Atlanta was under control of the Union Army, as a result of their fighting capabilities and a deceptive strategy devised by General Sherman (Army Times Editors, 1963, pp.15-22);

- The release of false propaganda by various governments across the globe, which, when left unaddressed by the countries involved, results in the inaccurate/incomplete documentation of history. This problem is described by Stokes (1994) in a discussion of the longstanding and tight linkages between the intelligence community, specifically M15- the Security Service and the SIS- Secret Intelligence Service, in the United Kingdom and the media. The author contrasts the well-kept, secret propaganda actions taken by government-sponsored organizations, which have close collaboration with the media, with those that have been well-documented and declassified in the United States. Stokes proposes that there is a potential for other similarly-typed activities to occur when the elite control or insinuate themselves into relationships with communications media;
- The reporting of inaccurate health information online. Eysenbach (2002) reports that online nutritional information has been found to have only a 10% accuracy rating;
- Fraudulent accounting information, released by the company Enron (Schlike, 2003), which caused billions of dollars in losses to shareholders as a result of hiding losses in a shell-type company and reporting false information;

- A loss of $40 billion annually in the manufacturing industry, which according to Ventana Research (2006) is the result of companies' "…inability to organize and analyze spend data" appropriately (Ventana Research, 2006), where spend data refers to expenditure information at various levels of granularity (e.g. product, category, supplier) within the supply chain (Spend analysis, 2007). Such a failure can ultimately cause companies to make decisions based upon inaccurate information.

From these examples, one overarching question resonates, how do we distinguish accurate information from deceptive information?

Problem

When inaccurate information finds its way into the decision making process, whether maliciously intended or not, it can be very problematic (Biros, George, & Zmud, 2002). However, according to Zmud (1990), when data is entered erroneously into a system to the benefit of someone other than the organization, the situation is called "strategic information manipulation." Rockman (2005) believes that it has become even more of a necessity for researchers to evaluate information critically. Efforts have been made to develop assessment methods for discriminating between inaccurate and accurate information across a variety of fields, in both the off-line and online contexts. One example comes from the field of competitive intelligence where a staff writer for *ONLINE* magazine indicates that

> [t]here is no great difference between evaluating methods developed by library scientists and those of intelligence agencies. Both reflect on a source's accuracy, timeliness, accessibility, and content. However, intelligence agencies pay

particular attention to a source's bias, veracity, and timeliness in order to evaluate it for analysis. Other factors to consider include a source's uniqueness, credibility, scope, depth, tip factor [explicit, rather than implicit, information], and other costs of getting useful information for intelligence (Anonymous, 2000).

However, instead of having a method for evaluating the authenticity of information internal to an organization, the current types of assessment methods were designed for use previous to its release. Finding a better technique to detect deceptive information in internal business communications may be a first step in identifying inaccurate organizational information and could help in the prevention of its public release.

Audience

This research project was directed at both decision-makers and the individuals that advise them. This project could be important to these audiences as since it will provide them with additional criteria to consider in the evaluation of information gained through text-based information systems used in communicative exchanges (i.e. email). Koops (2004) has shown that even in the animal kingdom, both risk and reliability are inextricably linked with respect to value of information and the decision of whether or not to use the information received. Koops, relative to the animal kingdom, (2004) proposes that

> …free information should not necessarily be used. Due to the costs associated with acting on misinformation, free information can be detrimental to the consumer. However, even if the cost of misinformation is high, information can still be used if the misinformation is relatively rare (i.e. the reliability is relatively

high). If the cost of misinformation is small relative to the benefit of correct information, then reliability can be low and misinformation can be more common than correct information and the consumer should still respond. In fact, the more beneficial the information is when correct, the lower its reliability can be degraded. Finally, if misinformation makes a consumer better off than no information, (but worse than correct information), then the occurrence of these white lies cannot make free information detrimental.

The cost of misinformation to a species can be incurred in two ways, by either a loss of their physical resources like energy that is expended in an effort to acquire food, water, shelter, or copulation or, by physical harm or death as a result of these same efforts (Koops, 2004). Koops (2004) goes on to note that simple changes related to the acquisition of the information have only minor consequences. For example, if one pays for the information, Koops (2004) indicates that it only slightly raises the level of minimum reliability needed to use the information. Also, when one pays to get reliable information, there will be a level at which the reliability of the information is maximized. Koops calls this optimal reliability and states that it "…will be higher than the minimum reliability, but less than perfect reliability..." and that, "[u]nlike the minimum reliability, optimal reliability decreases as acquisition costs increase and as the benefit of correct information decreases." (Koops, 2004, p. 109)

These relationships between information value, reliability, and risk are significant to decision-makers since frequently they must make high-risk decisions based on information from a variety of sources and some of the information that can help to guide their decisions may come from questionable sources. Evidence of the interrelated nature of these factors can also be seen

from an organizational perspective. When organizations purchase information from information brokers, like marketing research firms, they do so now with the expectation that the information broker has assessed the reliability information source prior to passing the associated information on to the client (Bates, 1997). Companies can then rely upon their information broker's source evaluation as a measure by which to ascertain the reliability of the purchased information and as justification to use the information in decision making when free information is not considered to be reliable enough. Further, as the cost of maximizing the optimal reliability increases, and as the benefit of the information decreases, it becomes less likely that an organization will purchase the information. Nonetheless, any increase in the probable reliability, loosely analogized to veracity in this example, of the purchased information, will afford decision-makers with the opportunity to pursue riskier options, in situations where normally they would not (in this case, free information acquisitions), or may provide a solid reason to consider other options or information sources.

A second aspect that is critical to consider from the perspective of the decision maker is the purpose behind the false information. In some cases, it may be simple error, but in other cases it may be subversive in its intent, for example The Office of National Counterintelligence Executive (NCIX) reports that 108 different countries were responsible for targeting technologies in the United States during 2005 (Annual report to congress, 2006). The NCIX also reports that many of the people involved included foreign businessmen, scientists, engineers, students, and academics, most of whom were not initially connected with any foreign governments or intelligence agencies when they arrived in the United States, but rather, "…instead, after finding that they had access to information that was in great demand abroad,

most engaged in illegal collection to satisfy their desire for profits, for academic or scientific acclaim, or out of a sense of patriotism to their home countries." (Annual report to Congress, 2006).

The financial loss of the companies victimized by economic espionage (or industrial espionage) are, according to the NCIX, very difficult to quantify because of the multiple costs (marketing, long-term market share, sales, replacement costs) associated with the loss, as well as the inability to actually predict future losses and companies are hesitant to report the attack and corresponding losses. Further, the report goes on to state that

> [m]ost of the foreign students and academics working in US research institutes are not involved with U.S. technology theft. In fact, many significantly contribute to the advancement of research at their respective universities and institutes. However, the sheer size of the population and the access that some have to key R&D projects make it inevitable that this group will serve as an important funnel abroad for technologies. (Annual report to Congress, 2006)

While these acts of economic and industrial espionage are theft-related in nature the individuals responsible may use various forms of deception to accomplish their activities in the academic, corporate, or government environments in order to remain undetected. Further, as reported by NCIX, the vast majority of individuals responsible for economic and industrial espionage are not "trained agents" or "government recruits," (Annual report to congress, 2006) and thus may be more easily identified when using deceptive practices.

With the changes in the global economy introducing increased competition into the marketplace and highlighting the need for reliable information sources, and the threat of