Information Manipulation Classification Theory for LIS and NLP

Victoria L. Rubin and Yimin Chen
Language and Information Technology Research Lab (LIT.RL)
Faculty of Information and Media Studies
University of Western Ontario
North Campus Building, Room 260,
London, Ontario, Canada N6A 5B7Affiliation
vrobin@uwo.ca, ychen582@uwo.ca

ABSTRACT
The Information Manipulation Classification Theory offers a systematic approach to understanding the differences and similarities among various types of information manipulation (such as falsification, exaggeration, concealment, misinformation or hoax). We distinguish twelve salient factors that manipulation varieties differ by (such as intentionality to deceive, accuracy, and social acceptability) to provide an abstract framework and conceptualize various permutations. Each variety then is represented as a set of features in the twelve-dimensional space. Our contributions are two-fold. In Library and Information Science (LIS) literature, a nuanced understanding of information manipulation varieties and their inter-relation lends greater awareness and sophistication to the ways we think about information and information literacy. For Natural Language Processing (NLP), the model identifies salient features for each manipulation variety, creates a potential for automated recognition and adaptability from deception detection technology to identification of other information manipulation varieties based on similarities.

Keywords
Deception varieties, lie-truth discrimination, information manipulation, disinformation, distortion; information literacy, faceted classification, feature-based approach, faceted classification.

INTRODUCTION
Information literacy, in spirit if not in its precise present form, has been among the foundational tenets of the Library and Information Science (LIS) since the dawn of our discipline. In his inaugural address at the thirteenth annual meeting of the American Library Association, President Arthur Bostwick called upon librarians to ensure that their collections were good, true and beautiful – and conversely, that they reject books that are bad, false, and ugly (Bostwick, 1908). While the court of public and professional opinion has not been particularly kind to Bostwick’s vision of The Librarian as a Censor in the intervening century, his warning of “books whose authors desire to deceive the public” and “glaring misstatements” as “objectionable feature[s]” that warrant exclusions from library collections of the time (Bostwick, 1908, p. 13-14) deserve to be revisited a century later, in light of resisting issues of misinformation and deception and associated research.

It seems no exaggeration to say that an overabundance of information is a universal condition today. In fact, Paul Zurkowski said exactly that in his report to the National Commission of Libraries and Information Science in 1974 – well before the popularization of the World Wide Web. These days, we routinely face a veritable “tsunami of hogwash… we are inundated… by online noise pollution” (Rheingold, 2009). Against this tide of shams, scams, and spam, the critical evaluation skills at the core of information literacy are indispensable in navigating through the flotsam and jetsam of inaccuracies and disinformation. And libraries have acted eagerly to meet these challenges: public, academic, special librarians and information professionals work tirelessly in classrooms, chatrooms, and boardrooms, arming their charges against the onslaught of undesired content. Conceptual tools provide sets of questions to evaluate the Currency, Reliability, Authority, and Purpose/Point of View of any information and its source (“The CRAP Test. LOEX Wiki (Library Orientation Exchange)," 2008). Google provides training webinars and lessons plans to help better understand search results and evaluate credibility. Other conceptual tools include McManus’ “BS detector” (2009) and Sagan’s “baloney detection kit” (1996). (See also Rieh (2010) who overviews recent credibility typologies in LIS and HCI, and (Rubin & Vashchilko, Forthcoming) who review several state-of-the-art automated deception detection tools with implications for information quality assessments).
Although there is no shortage of teaching tools for information literacy, the need to reduce content into neat, digestible bullet points and best practices gives the impression that all forms of deception are equivalent, and that following the guidelines should lead to a binary accept/reject decision. Yet, clearly, there is a difference between outright lies and lies of omission, between exaggeration and fabrication. A biased source may still be a useful and informative, provided it can be recognized as such. It seems logical that a more nuanced understanding of the various types of deceptions and their relation to each other could lend greater sophistication the ways we think about information. However, thus far in the LIS literature, there has been no abstract framework to conceptualize these permutations of deception and varieties information manipulation.

OBJECTIVES
In LIS terms, we set out to unify and develop a conceptual holistic system that covers all possible types of information manipulation in the information transmission channel from sender to receiver, accounting for personal, interpersonal, societal and cultural affordances in the information production, presentation, and perception cycle.

Current methods for deception detection, though state-of-the-art in their development, can only handle the general sense of deceptiveness in any given message. Lumping together all varieties of information distortions as deception, at best, produces an answer to a binary question: is there a general sense of deceptiveness? Considering all varieties of deception for detection indiscriminately is unrealistic. Some fuzziness around concepts prevents a clear task definition for potential automation techniques, so in this work we apply systematic sorting with NLP application in mind.

LITERATURE REVIEW
Information Manipulation
What we mean by Information Manipulation is a process in which information (the artifact in some shape or form) is being transmitted between human agents, yet certain types of distortions occur in the process. It is an extension of the classical Shannon-Weaver’s model of information transmission with a crucially different outcome in the fidelity of what is received, perceived and concluded based on what or how the information was presented. In Shannon’s view, the goal is to transmit the message from the source to the decoder in “exactly or approximately the same” way as a reproduction. But what if the sender has alternative goals? Either the source intentionally attempts to create a false impression or conclusion in the receiver’s mind, and such phenomenon is typically termed deception (Buller & Burgoon, 1996; Rubin, 2010; Zhou, Burgoon, Nunamaker, & Twitchell, 2004). Alternatively, other distortions occur in the interpretation or perception of the message by the receivers, in which case the communication is equally unsuccessful since the resulting impression is erroneous. Ideally, to support their decision making, information users should rely on accurate, truthful, and complete information from credible confident expert sources. Their information literacy is precisely that ability to identify and weed out manipulations, but distortions may occur at a vast number of points in the information communication channel, and the paper aims to model this conceptual space with a classificatory approach by identifying salient factors and variations within each.

McCornack’s (1992) Information Manipulation Theory refers to information manipulation as the management of information by the sender to provide the receiver with a perception of that same information believed to be false by the sender. He emphasizes how information is manipulated, “assuming that interactants possess assumptions regarding the quantity, quality, manner, and relevance of information that should be presented, it is possible for speakers to exploit any or all of these assumptions by manipulating the information that they possess so as to mislead listeners” (p.1). McCornack suggests that deceptive messages function deceptively because they covertly violate the principles that govern conversational exchanges. This principle is known as Cooperative Principle (Grice, 1975) with four Gricean Maxims (of Quality, Quantity, Relation and Manner) following which, in principle, should lead to successful communication and allow sense making of the conversational acts, in spite of blatant violations such as sarcasm. McCornack (1992) characterizes verbal deception as “a particular sub-class of uncooperative acts, a sub-class of acts in which the principles guiding cooperative exchanges are covertly violated” (p. 13).

Deception Varieties
Taxonomies of deception have often been proposed in contrasting categories, ranging in number of categories from 2 to 46. For instance, Chisholm and Feehan (1977) distinguish two broad categories by passive or active role of the deceiver: commission (purposefully and consciously communication) and omission (allowing a person to believe something untrue).

In their Interpersonal Deception Theory, Burgoon and Buller (1994) distinguish three deception varieties based on seven differentiating features: amount and sufficiency of information, degree of truthfulness, clarity, relevance, ownership, and intent. The types are falsification (lying or describing “preferred reality”), concealment (omitting material facts) and equivocation (dodging, skirting issues by changing the subject or offering indirect responses) (Burgoon & Buller, 1994).

Metts (1989) also names three basic "lie types", slightly diverging from Burgoon and Buller: falsification (asserting information contradictory to the true information or explicitly denying the validity of the true information), distortion (manipulation of the true information through exaggeration, minimization, and equivocation, such that a listener would not know all relevant aspects of the truth or
Given the number of existing schemes and taxonomies, how can they be reconciled and unified meta-analytically?

**METHODS FOR DERIVING THE THEORY**

What we would like to offer in this theoretical work is a set of distinguishing factors or dimensions by which most, if not all, of the above various types of deception and information manipulation vary. We employ binary distinctions within each dimension that are often used in computing and linguistics: the positive and negative valence on each particular continuum. In classification theory, such dimensions are also referred to as facets with their internal categories on the continuum as foci. For instance, distortions or inaccuracies vary in terms of the facet of the source’s intentionality to deceive: ± intentionality. A combination of two dimensions creates a two-dimensional distortion space for information manipulation. For instance, taking the facets of intentionality and accuracy together, we obtain a two-by-two matrix with exemplifies intersections of the properties on each dimension (see Table 1).

If we then invert this example and look at individual manipulation varieties, each instance that fills the matrix box then acquires a set of features with positive and negative valence from each of the two dimensions (see Table 2).

We build our classification theory incrementally, by adding dimensions first and identifying their valence in each case (based on prior research, e.g., (Hopper & Bell, 1984) and creating a multi-dimensional information manipulation space. Simultaneously, each variety of information manipulation, has its own coordinates within that space, where the coordinate values represent the variation on the continuum for each dimension.

**Table 1. Two-dimensional Information Manipulation Space Example: Intentionality to Deceive x Information Accuracy.**

<table>
<thead>
<tr>
<th>INFORMATION ACCURACY</th>
<th>+ accurate</th>
<th>- accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTENTIONALITY TO DECEIVE</strong></td>
<td>+ intended</td>
<td>- intended</td>
</tr>
<tr>
<td>truth</td>
<td>i.e., statements matching speaker’s beliefs</td>
<td>deception varieties</td>
</tr>
<tr>
<td>untruth</td>
<td>e.g., slip of the tongue</td>
<td>misinformation</td>
</tr>
<tr>
<td>CRIMES</td>
<td>e.g., erroneous statements</td>
<td></td>
</tr>
<tr>
<td>LIES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FICTIONS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLAYINGS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MISLIES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UNLIES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Inverted Feature Set for Each Information Manipulation Variety in Table 1.**

- truth -> [+ intended, + accurate]
- falsification -> [+ intended, - accurate]
- slip -> [- intended, + accurate]
- misinformation -> [- intended, - accurate]

The caveat in this process is that binary distinctions on each continuum are simplistic. While it is convenient to think about the extremities (such as ± socially acceptable lie), in reality each dimension implies a continuum, from for instance highly socially [- acceptable] (like fraud), to...
mildly [- acceptable] (like spam), to perfectly [+ acceptable] (like a concealed surprise birthday party). Thus, while Figure 2 shows the two extremities, we emphasize the gradient nature of each suggested dimension and offer an intermediate “somewhat” category on the continuum (with a ~ valence). In addition, in some cases the valence may be unknown or questionable (“?”) or not applicable (“n/a”). Each facet receives five possible features (or foci) in sum: “-”, “~”, “+”, “?”, “n/a”. Figure 2 exemplifies them with the information accuracy dimension.

**Figure 2. Three Gradients on Information Accuracy Continuum, and Two Other Possible Values: unknown and not applicable.**

### INFORMATION MANIPULATION CLASSIFICATION

The Information Manipulation Classification Theory is comprised of 12 distinguishing factors. It creates a twelve-dimensional space that account for various manipulations that occur to information when it is being transmitted from sender to receiver (Table 3). We suggest feature values (Figure 2) for each of the reviewed manipulation varieties (some in unification). As the work progresses, we are considering appropriateness and exact feature values for other factors that may play a role in perception of information such as inter-personal trust, source’s credibility and expertise, as well as infidelity due to the information channel distortions (e.g., in OCR).

<table>
<thead>
<tr>
<th>Existing Taxonomies of Deception Varieties</th>
<th>Distinguishing Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Information Properties</td>
<td>2. Sender’s Intentions</td>
</tr>
<tr>
<td>Quality</td>
<td>Quantity</td>
</tr>
</tbody>
</table>

**Table 3. The Information Manipulation Classification.**

**CONCLUSIONS AND IMPLICATIONS**

Information Manipulation Classification Theory is a holistic classification of information manipulation multi-dimensional conceptual space. This multi-dimensional approach offers orthogonal facets that describe how information manipulation types vary, covering the whole continuum and exemptions within each facet. The feature-based system works as a checklist for conceptual LIS assessments and can be adapted for computation with Natural Language Processing techniques, as we offer pre-defined sets of dimensions (facets) and exhaustive and
mutually exclusive features on each continuum (foci). While the Information Manipulation Classification Theory presents a synthesis of empirically validated dimensions that distinguish varieties of information manipulation, it requires further validation and testing for interactions, exhaustively, and mutual exclusivity, subject to further research. By identifying and qualifying different dimensions of distortion and manipulation, we can move away from a basic accept/reject binary model to more sophisticated ones which recognize degrees of "untruth" and present content in more authoritative ways.

Our broader applied research agenda involves creating and developing methods for automatically discerning deceptive messages from truthful ones in computer-mediated communication contexts including personal stories, resume falsifications, and online citizen journalist news (Rubin, 2010; Rubin & Conroy, 2011, 2012; Rubin & Vashchilko, 2012). Having a clear abstract conceptual LIS model of inter-connections and difference between varieties of deception strategies is beneficial for practical applications such as various text analytics and language technologies (typically based on NLP and machine learning techniques). Being able to adapt existing prior technologies and customize them (for instance, form spam detection to identity theft domains) would offer portable automated solutions to identifying information manipulation forms.

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 (Academic Development Fund at the University of Western Ontario).

REFERENCES


