Trust Incident Account Model: Preliminary Indicators for Trust Rhetoric and Trust or Distrust in Blogs

Victoria L. Rubin

Faculty of Information and Media Studies
University of Western Ontario
London, Ontario, Canada N6A 5B7
vrubin@uwo.ca

Abstract

This paper defines a concept of *trust incident accounts* as verbal reports of empirical episodes in which *a trustor* has reached a state of positive or negative expectations of *a trustee*'s behavior under associated risks. Such expectations are equated to *trust* and *distrust*, correspondingly, and present a sharp contrast with hypocritical use of *trust rhetoric* with ulterior motives such as an attempt to manipulate readers or gain trustworthiness. Distinguishing the three: trust, distrust, and trust rhetoric, is formulated as a new challenge in sentiment analysis and opinion-mining. Based on a preliminary exploration of trust narratives in blogs, 14 categories of textual indicators were identified manually. The finer-grain analytical model of trust incident accounts is proposed to include 12 information extraction frame components: trustor, trustee, source, textual clue, trust valence, risks, reasons, actions, trustor-trustee relationship, narrow context, broad domain, and complements. The study draws a cross-disciplinary theoretical bridge from social science and information technology trust literature to opinion-mining, and emphasizes the value of understanding trust in longer-term social relations.

1. Introduction

Trust permeates "the whole mechanism of society… like the air we breathe: and its services are apt to be taken for granted and ignored, like those of fresh air, until attention is forcibly attracted by their failure" (Marshall 1919). Almost any successful social interaction or relation requires trust on behalf of its participants. Distrust can have a paralyzing effect. Trust rhetoric, on the other hand, is a way to manipulate trust feelings. It raises suspicion, carries a negative connotation, and may be perceived as weakness (Möllering 2006).

The social blogging medium abounds in subjective testimonials of concrete personal episodes of having placed trust in a particular individual, organization, or an artifact of human activity, and having enjoyed or suffered the consequences. This vast resource of volunteered and publically available empirical evidence is largely uncharted waters due to its being dispersed over millions of blog posts. Aggregation and a certain structure (e.g., by trust salient features such as trustor, trustee, context, risks, domain) can provide a powerful empirical tool for an interdisciplinary research on trust. Content-analyzing existing texts is an alternative way of reaching trustworthiness and obtaining information for a recommendation system without user feedback. The more trust incident accounts for an entity or content, the more trustworthy it can be considered. Automating the acquisition and analysis of trust narratives, at least in part, is the ultimate goal.

1.1. Trust as a Challenge for Sentiment Analysis

While trust is not an inherently linguistic concept, trust incident accounts are naturally expressed through language, and have distinct linguistic regularities amenable to automated identification with information extraction (IE) techniques in Natural Language Processing (NLP). Opinion-mining, or sentiment analysis, entails computational treatment of opinion, sentiment, and subjectivity in text with NLP techniques (Pang and Lee 2008). Even though from a rationalist's point of view, a decision to trust or not is an estimation of maximizing trustor's potential benefits, trust is private and subjective. Trust has an emotional component and requires "a leap of faith" (Möllering 2006), and willingness to tolerate uncertainty and accept vulnerability (Rousseau et al. 1998). Trust can serve as a gap-filler for explicit knowledge (Marsh and Dibben 2003) in the absence of adequate information for rational decision-making.

Recent most relevant efforts in opinion-mining have been concentrated on identifying opinion-holders (Kim and Hovy 2004), targets towards which sources hold opinions (Ruppenhofer, Somasundaran, and Wiebe 2008), semantic orientation: positive, negative, or neutral (Esuli and Sebastiani 2005); levels of certainty in modalized statements (Rubin 2007); and sources of happiness and sadness in everyday life (Mihalcea and Liu 2006). Typical subject

Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

domains include product reviews or customer feedback and focus on short-lived commercial interactions. Trust incident accounts extend opinion-mining into a new domain of longer-term sustained social relations of both professional and personal nature (e.g., doctor – patient relations and friendships).

1.2. Trust in Blogs

Blogs offer a wealth of subjective information about meaningful social relationships and can be analyzed in at least three different ways. First, which bloggers trust which bloggers? Influential blogs and like-mindedness within blogger networks are predicted via blog link structure and link polarity propagation techniques (Kale et al. 2007). Second, how much do blog readers trust a particular blogger? Authority ranking or recommendation mechanisms rely on blog-readers' approval ratings by a particular post and by a cumulative blogger score, e.g., USAToday.com. Similar to e-auctioning or e-commerce reputation systems, strangers can decide whom to trust based on aggregate ratings. However, eliciting users' feedback has proven to be problematic (Resnick et al. 2000). And third, the focus of this work: who trusts whom (or what) to do what in which context? Trust incident accounts include non-blogging entities; and no direct data elicitation is required. Trustworthiness of blogs requires a different inquiry approach, e.g. credibility assessment (Rubin and Liddy 2006), and is beyond the scope of the present study.

2. Trust Incident Account Model

2.1. Trust and Distrust

Trust (or **distrust**) is defined here as a positive (or, correspondingly, negative) expectation of *a trustor* regarding the behavior of *the trustee*, in a context that entails risk to the trustor (synthesized from Marsh and Dibben (2003); Rousseau et al (1998)). Trust is manifested through **trust incidents:** empirical episodes in which this state of trust is reached "irrespective of whether the trustor is conscious of this or whether it is directly observable by others in any way" (Möllering 2006). **Trust incident accounts** in texts are, thus, verbal descriptions[1] of such incidents. In example 1, *trust* is never explicitly mentioned but advice results in action:

> *(1) …I **sought out advice of**[2] a financial **guru** friend who told me to consolidate my loans for a better rate… I **followed that advice** and consolidated… If I had not done that I would have been stuck paying 6.8% interest* (120: engineeradebtfreelife.com)[3].

In this work, **distrust** is assumed to be a direct opposite to trust. In the sample **distrust incident account** 2, the blogger is distrustful of the doctors in her cautionary tale:

> *(2) Just be careful when it comes to your gallbladder. Get a second opinion if you don't feel comfortable with what you are being told. I trusted what my family doctor and the general surgeon told me and **should have gotten a second opinion.** I was told by them that the problems I was having would be all better after the surgery. Well, they weren't* (101:copycatchat.com)

2.2. Information Extraction Frame Components

Various trust classifications emphasize importance of particular components. Artz and Gil (2007) distinguish entity and content as objects of trust. Zucker (1986) sees three reasons for trusting: process (i.e., previous experiences), characteristics, and institutions (as guarantors). Lewicki and Bunker (1996) emphasize closeness of the trustor-trustee interpersonal relationship: calculus, knowledge, and identification.

Representing trust conceptually becomes a multi-piece puzzle with each component having its own typology. Hardin's (1993) four-part formulation *"A trusts B to do X in matters Y"* has an advantage of making trust contextual and implies a list of specifics that A may trust B with (e.g, with house keys but not child care). Hardin's formulation is extended here to serves as an IE frame, a data-structure for representing a stereotypical situation (Minsky 1974), with letters indicating variables in this formulation:

> In a trust or distrust incident account **N** identified based on a textual clue **M**, the trustor **A** manifested a positive or negative expectation **T** in the trustee **B**'s behavior **X** in the matter of **Y** at the risk of **Z** for reasons **S** in a broader domain **D,** according to the narrator **C.**

Core model components are directly observable and necessarily stated in texts (and inscribed into solid line blocks in Fig. 1): trustor **A** (the trusting party), trustee **B** (the object of trust, the trusted party), source **C** whose perspective is reported (same parties as **A, B,** or a third party), trust valence **T** [trust, distrust], and a textual clue **M** that anchors the reader **R**'s perception of trust valence.

Peripheral model components (inscribed into cursory line blocks in Fig. 1) are elaborations that may (not) be explicitly stated: expected actions or behavior **X**, specific contexts **Y**, potential risks or outcomes **Z**, reasons or justifications **S,** and trust complements **Q**, closely associated coordinate concepts, e.g., trust and guidance. With some heuristics and inference, further analytical types could be extracted (**A**'s type, α = [person, organization]; **B**'s type, β = [person, organization, artifact]; **A-B** relationship type, ρ = [professional capacity; personal relationship; mixed; unknown]. In addition, domain **D** situates the account incident in a broader context, e.g., health care, law, sports, finances.

[1] We assume incident accounts are typically truthful.
[2] Anchoring textual clues are in bold in examples.
[3] Database record numbers are followed by root blog URLs.

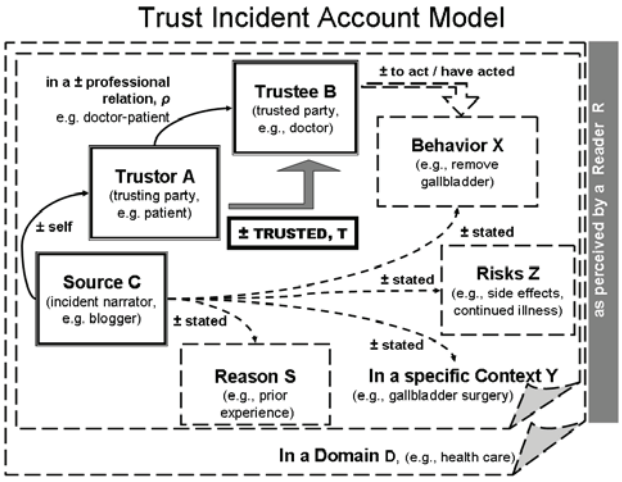

Figure 1. **Trust Incident Account Model.** The scheme visualizes the model components and their relations for a verbalized trust incident account **N**, as seen by the reader **R**.

2.3. Trust Rhetoric

Trust rhetoric, in contrast with trust and distrust incident accounts, is a linguistic façade. It is often used hypocritically in a conscious attempt to invoke a socially desirable notion of trust, control the reader, or exploit the situation. Consider:

> *(3)* ***You have to trust me as his personal attorney that*** *this is a risk free transaction and does not have any criminality with you. You will have to provide a bank account where the fund will be transferred into*...(127: 419-scams.the-world-in-focus.com).

Other trust rhetoric may be subtler. With ulterior motives, managers may try to impose their trust on employees:

> *(4) In return for my efforts in creating such a work climate,* ***they give me their best efforts in*** *building software. I* ***trust them to do this because*** *I know they appreciate what I do* (302: redtape.msnbc. com).

Promises, reassurances, and pleas may be offered in an (un)conscious attempt to boost or gain trustworthiness:

> *(5) I'm* ***fully committed to*** *performing* ***at my highest level at all times****...*(128: finddreamjobs.blogspot.com)
>
> *(6) I* ***promise you that*** *while I'm there I'm going to be working for you guys*(125:wealthyaffiliatereviews.net)

3. Methodology

The analytical model of trust incident accounts is empirically developed with the analytical induction technique (Punch 1998). A pilot corpus of 302 trust-related narratives is collected based on hypothesized keywords using GoogleBlog Search Engine. Retrieved blog posts and comments are content-analyzed (Krippendorff 1980) and textual indicators are harvested. Trust narratives are classified by 2 coders three-way: trust, distrust, trust rhetoric, and then annotated with finer-grained model components. Collected textual indicators can serve as seed words to bootstrap other trust expressions from relevant unannotated texts, as in Riloff and Wiebe (2003). The feasibility of extracting more complex fine-grain components based on semantic role analysis is being assessed.

4. Preliminary Findings and Discussion: Emerging Indicator Categories

The preliminary observations are, firstly, that genuine trust tends to be found in positive semantic orientations contexts. Secondly, possibly apparent, genuine trust incidents are incompatible with hateful, violent, or antagonistic rhetoric. And, thirdly, somewhat counter-intuitively, at the lexical level *trust* and its synonyms are likely to lead to either distrust or trust rhetoric rather than trust reports. The paradox is that the mere fact of uttering *"I trusted John to get here on time"*, triggers as association in the reader's mind that John failed to do whatever he was trusted with, as if uttering *trust* diminishes its powers. Some obvious search keywords such as *trusted*, *gained trust*, or *believed in*, are misleading. Examination of indirect vocabulary that is likely to retrieve trust narratives revealed 14 emerging themes.

4.1. Preliminary Trust Indicators

Trust incident accounts are found with indicators of:

- recommendation or referral (e.g., "*he was the man to go to for*", "*who was known to be great at*");
- praise, expressed admiration, thankfulness (e.g., *"helped me and he was amazing", " was the real deal"*);
- stated actions upon advice or advice seeking (e.g., *"took his advice and", "followed their advice"*);
- leap of faith; reliance; acting upon intuition (e.g., *"had a good feeling about him", "had faith in him", "had confidence in her", "relied on their judgment to"*);
- decisions or actions under lack of information (e.g., an ambivalence typically expressed with "*despite", "although", "as if", "nevertheless"*).

4.2. Preliminary Distrust Indicators

Distrust incident accounts were easier to retrieve with an expanded WordNet synset, e.g., *mistrust, suspect, doubt, disbelieve, suspicion,* and with uncertainty modifiers such as *claimed, alleged,* and *supposedly*. Also, distrust was palpable in 3 other negative connotations:

- anger strongly expressed with profanities and hostilities *(e.g., "a bunch of phonies", "absolutely stinks", "goddamn", and variations of "f*** idiots"),* or mildly expressed as an intuition, suspicion, apprehension *(e.g., "had a terrible gut feeling about");*
- strong disapproval (e.g., *"don't think you should"*);
- blame, criticism, direct accusations (e.g., *"darn liars"*).

4.3. Preliminary Trust Rhetoric Indicators

Trust rhetoric currently appears to be in six forms:

• clichés, meaningless use of words:

(7) …we have lost friends, gained friends, ***gained trust, lost trust...*** *it comes with a longer list, but* ***you get the point*** (13: jraedavis.blogspot.com).

• appeals for trust (e.g., *"Trust me!", "believe me I'm an expert",* also see example 6);

• ad-like overstatements, over-eager or over-generalized:

(8) "… ***has so much to offer*** *for your construction needs… They* ***already gained trust*** *from the customers…(49: lirastafford.com);*

• promises, reassurances, guarantees (*"I swear I'll never");*

• loyalty and devotion pledges (example 5); and

• stated bets.

5. Concluding Remarks

A new task in opinion-mining and sentiment analysis is formulated as a trust incident account identification problem and a three-fold classification problem (trust, distrust, rhetoric) in IE and NLP. Preliminary observations from the pilot data include entailment of trust in positive semantic contexts and an incompatibility of trust with hateful rhetoric. Trust rhetoric is often found accompanied by clichés and ad-like over-statements, as well as in appeals for trust, promises, and devotion pledges. Interdisciplinary links with social science and information technology trust literature are made on a conceptual level, and a finer-grain classification for IE and corpus construction is proposed. Future work will include scaling up and bootstrapping pattern learning.

Acknowledgements

Special thanks to Jacquie Burkell and three anonymous reviewers for their suggestions, and to Dex Gittens and Olga Buchel for data collection. The work is funded in part by the University of Western Ontario Grant no. R3995A02.

References

Artz, D., and Gil, Y. 2007. A Survey of Trust in Computer Science & the Semantic Web. *Web Semantics,* 5(2): 58-71.

Esuli, A., and Sebastiani, F. 2005. Determining the Semantic Orientation of Terms through Gloss Classification. In *Proceedings of the 14th Conference on Information and Knowledge Management*, 617-624, Bremen, Germany: ACM.

Hardin, R. 1993. The Street-Level Epistemology of Trust. *Politics & Society,* 21(3): 505-529.

Kale, A., Karandikar, A., Kolari, P., Java, A., Joshi, A., and Finin, T. 2007. Modeling Trust and Influence in the Blogosphere Using Link Polarity. In *Proceedings of the International Conference on Weblogs and Social Media.*

Kim, S.-M., and Hovy, E. 2004. Determining the Sentiment of Opinions. In Proceedings of the 20th Intenational Conference on Computational Linguistics, 1367-1373, Switzerland: ACL.

Krippendorff, K. 1980. Content Analysis: An Introduction to Its Methodology. Beverly Hills, CA: Sage Publications.

Lewicki, R. J., and Bunker, B. B. 1996. Developing and Maintaining Trust in Work Relationships. In R. M. Kramer and T. R. Tyler. eds. *Trust in Organizations*: 114-139. Thousand Oaks: Sage.

Marsh, S., and Dibben, M. R. 2003. The Role of Trust in Information Science and Technology. *Annual Review of Information Science and Technology,* 37(1): 465-498.

Marshall, A. 1919. *Industry & Trade*. London: Macmillan.

Mihalcea, R., and Liu, H. 2006. A Corpus-Based Approach to Finding Happiness. In *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*: AAAI Press.

Minsky, M. 1974. A Framework for Representing Knowledge. In P. Winston. ed. *The Psychology of Computer Vision*: 211–277. New York: McGraw-Hill.

Möllering, G. 2006. *Trust: Reason, Routine, Reflexivity*. Amsterdam: Elsevier.

Pang, B., and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.,* 2(1-2): 1-135.

Punch, K. F. 1998. *Introduction to Social Research: Quantitative and Qualitative Approaches*. London: Sage.

Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. 2000. Reputation Systems. *Communications of the ACM,* 43(12): 45-48.

Riloff, E., and Wiebe, J. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the Empirical Methods in NLP*, 105-112: ACL.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. 1998. Not So Different after All: A Cross-Discipline View of Trust. *The Acad. of Manag. Rev.,* 23(3): 393-404.

Rubin, V. L. 2007. Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of the Human Language Technologies Conference*, 141-144, Buffalo, NY.

Rubin, V. L., and Liddy, E. 2006. Assessing Credibility of Weblogs. In *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*: AAAI.

Ruppenhofer, J., Somasundaran, S., and Wiebe, J. 2008. Finding the Sources and Targets of Subjective Expressions. In *Proceedings of the 6th International Language Resources and Evaluation*, Marrakech, Morocco.

Zucker, L. G. 1986. Production of Trust: Institutional Sources of Economic Structure, 1840 -1920. In B. M. Staw and L. L. Cummings. eds. *Research in Organizational Behavior*, 8: 53-111. Greenwich, CT: JAI Press.