MIS Quarterly

RESEARCH ARTICLE

TEXT ANALYTICS TO SUPPORT SENSE-MAKING IN SOCIAL MEDIA: A LANGUAGE-ACTION PERSPECTIVE[1]

Ahmed Abbasi
McIntire School of Commerce, University of Virginia,
Charlottesville, VA 22908 U.S.A. {abbasi@comm.virginia.edu}

Yilu Zhou
Gabelli School of Business, Fordham University,
New York, NY 10023 U.S.A. {yzhou62@fordham.edu}

Shasha Deng
School of Business and Management, Shanghai International Studies University,
Shanghai, CHINA {shasha.deng@shisu.edu.cn}

Pengzhu Zhang
Antai College of Management and Economics, Shanghai Jiaotong University,
Shanghai, CHINA {pzzhang@sjtu.edu.cn}

Social media and online communities provide organizations with new opportunities to support their business-related functions. Despite their various benefits, social media technologies present two important challenges for sense-making. First, online discourse is plagued by incoherent, intertwined conversations that are often difficult to comprehend. Moreover, organizations are increasingly interested in understanding social media participants' actions and intentions; however, existing text analytics tools mostly focus on the semantic dimension of language. The language-action perspective (LAP) emphasizes pragmatics; not what people say but, rather, what they do with language. Adopting the design science paradigm, we propose a LAP-based text analytics framework to support sense-making in online discourse. The proposed framework is specifically intended to address the two aforementioned challenges associated with sense-making in online discourse: the need for greater coherence and better understanding of actions. We rigorously evaluate a system that is developed based on the framework in a series of experiments using a test bed encompassing social media data from multiple channels and industries. The results demonstrate the utility of each individual component of the system, and its underlying framework, in comparison with existing benchmark methods. Furthermore, the results of a user experiment involving hundreds of practitioners, and a four-month field experiment in a large organization, underscore the enhanced sense-making capabilities afforded by text analytics grounded in LAP principles. The results have important implications for online sense-making and social media analytics.

Keywords: Design science, text analytics, social media, natural language processing, language-action perspective, conversation disentanglement, coherence analysis

[1] H. R. Rao was the accepting senior editor for this paper. Paul Hu served as the associate editor.

The appendices for this paper are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

DOI: 10.25300/MISQ/2018/13239

Introduction

The rapid growth of social media and online communities has dramatically changed the manner in which communication takes place. Organizations are increasingly utilizing general-purpose social media technologies to support their business-related functions (Mann 2011). According to a *McKinsey Quarterly* report, 50% of the more than 1,700 organizations surveyed are using social networking, 41% are using blogs, 25% are using wikis and 23% are using microblogs (Bughin and Chui 2010). Moreover, these numbers have more than doubled over a four-year period (Bughin and Chui 2010). Web 2.0 technologies are being leveraged for internal purposes, customer-related purposes, and to work with external suppliers and partners. Organizations are deriving considerable benefits from their use, including increased speed of access to knowledge, enhanced identification of experts, increased number of successful innovations, and reduced communication and operational costs (Bughin and Chui 2010; Chau and Xu 2012).

Sense-making is an information-processing task that serves as a critical prerequisite for decision-making (Russell et al. 1993; Weick et al. 1995). Despite their various benefits, existing social media technologies suffer from two important limitations which inhibit sense-making:

- Communication modes such as chat rooms, newsgroups, forums, blogs, social networking discussions, and microblogs are highly susceptible to intertwined conversations and incoherence (Honeycutt and Herring 2009). In group discussion, these issues make it difficult for analysts and supporting technologies to determine the correct message-conversation affiliations and reply-to relations among messages (Aumayr et al. 2011; Fu et al. 2008; Herring 1999).

- Existing text and social media analytics tools tend to focus on the semantic dimension of language: what people are saying. However, while using such technologies, organizations have difficulty understanding discussants' actions, interactions, and intentions (Mann 2011).

These limitations have significant implications. Ineffective sense-making can impact quality of decisions and actions (Klein et al. 2006; Russell et al. 1993). Furthermore, information sources and/or technologies deemed by users to not adequately support sense-making see diminished usage in future decision-making processes (Pirolli and Card 2005; Russell et al. 1993). In the context of social media analytics tools, based on industry surveys of key value-driving use cases, and multiple independent assessments of existing social media technologies that support these use cases, Table 1 summarizes challenges stemming from the two aforementioned limitations (Mann 2013; Zabin et al. 2011). According to industry surveys, three of the most important use cases for social media analytics are (1) identifying issues described in user-generated content; (2) identifying ideas and opportunities; and (3) identifying important discussion participants (Zabin et al. 2011). Multiple independent assessments of the functionalities of nearly 40 major existing social media analysis technologies highlight their exclusive reliance on keyword, topic, and sentiment analysis, underscoring their limitations for key use cases (Mann 2013; Zabin et al. 2011). Consequently, the inability of state-of-the-art text and social media analytics tools to provide sufficient sense-making has diminished their perceived return on investment (Zeng et al. 2010). Supplementing the pervasive semantic view with a pragmatic perspective is critical for comprehending communicative context and intentions surrounding issues and ideas (Te'eni 2006), and for understanding participant roles and importance (Fu et al. 2008). Over 80% of organizational data is represented in the form of unstructured data (Kuechler 2007), with email and social media accounting for a growing proportion (Chau and Xu 2012; Halper et al. 2013; Kuechler 2007). There is thus a need for advanced text analytics tools capable of supporting sense-making in online discourse.

In addressing the aforementioned challenges, there are two major research gaps. First, existing text analytics research has adopted a semantic view (Abbasi and Chen 2008; Lau et al. 2012), with thousands of studies looking at topic and sentiment analysis. The body of literature on the pragmatic view emphasizing communication context, actions, and interactions, has received less attention. Second, text analytics studies that have adopted the pragmatic perspective are fragmented. No overarching framework exists to guide the design and development of these artifacts. In order to address these gaps, in this study, we adopt the design science paradigm to guide the development of the proposed IT artifacts (Hevner et al. 2004): a language aspect perspective (LAP) based text analytics framework and system. By emphasizing the pragmatic aspect of language, LAP provides insights for the design of information systems that consider communicative context and actions (Schoop 2001; Winograd and Flores 1986). In particular, LAP emphasizes the interplay between conversations, communication interactions between users and messages, and the speech act composition of messages. Guided by LAP, the proposed framework encompasses three components designed to collectively alleviate the current challenges and facilitate enhanced sense-making from online discourse.

We rigorously evaluated a system that was developed based on the framework in a series of experiments that demonstrate the utility of each individual component of the system in com-

Table 1. Summary of Key Social Media Analysis Use Cases and Challenges

Use Case	Challenges
Identifying Issues	Most state-of-the-art social media analysis tools only include keyword, topic, or sentiment analysis for messages or threads. These tools make it very difficult to identify questions, suggestions, desires, assertions, declarations, etc. Furthermore, by focusing at the message or discussion thread level, these tools fail to consider communication within its conversation context. Collectively, these challenges can impact capabilities for identifying issues or opportunities such as customer churn, brand devaluation issues, popular suggestions, etc.
Identifying Ideas and Opportunities	
Identifying Important Participants	Key participants, including brand advocates, influencers, experts, connectors, and leaders, are typically identified using interaction metrics based on social network centrality measures. Existing tools' reliance on system-based interaction cues dramatically diminishes the accuracy and quality of insights pertaining to participant roles and rankings in social media.

parison with existing methods. Furthermore, the results of a user experiment involving practitioners from multiple industries illustrate the enhanced sense-making capabilities afforded by LAP-based text analytics systems. Additionally, a four-month field experiment revealed that social media team members at a telecommunications company perceived the additional LAP-based (pragmatic) information to improve system usefulness and ease-of-use for monitoring tasks, relative to those members relying on (solely semantic) information from an existing social media analytics system.

The study makes two sets of research contributions. Our primary contributions are from a design science perspective. We present a robust *framework* and *system instantiation* grounded in LAP principles, which emphasizes the interplay between conversations, coherence relations, and message speech acts. We also propose *novel text analytics methods* for conversation disentanglement, coherence analysis, and speech act classification, thereby enhancing the state-of-the-art for IT artifacts that analyze social media. We also present several *empirical insights*, such as the impact of incoherent reply-to relations on error rates for social network centrality metrics across various social media channels. By demonstrating the efficacy of the proposed system in user and field studies, the results have important implications for researchers analyzing social media, as well as various organizational functions that leverage internal and/or external sources of social media to support communication and decision-making, including customer relationship management, workforce analytics, risk management, and market research.

The remainder of the paper is organized as follows. The next section presents a motivating industry example highlighting the need for sense-making. The subsequent section describes our LAP-based framework, reviews work related to key components of the framework, and presents research questions. Based on this framework, we then describe a text analytics system for online sense-making that incorporates important concepts from prior LAP studies. This is followed by the presentation of robust evaluation of various facets of the proposed system, including experiments that evaluate each component, user experiments, and a field study that provides an in-depth assessment of the system's overall sense-making capabilities. The final section offers our conclusions.

The Need for Sense-Making: The TelCorp Example

In this section, we present a motivating industry example highlighting the need for enhanced sense-making from social media. It is important to note that the example presented is not nuanced or niche, but rather, represents the type of situation encountered by organizations in various industry verticals on a routine basis. We mention a few other high-profile examples at the end of this section, and later incorporate data from organizations in different industries as part of the test bed.

In the fall of 2012, TelCorp (fictious name), one of the ten largest telecommunications and data service providers in the United States, increased the maximum upload speed for customers subscribed to their highly profitable premium Internet plan. A press release was placed on the company's website and messages describing the move were posted on several social media channels, including TelCorp's Facebook fan page, Twitter, and various web forums. Like most large telecommunications service providers, TelCorp's customer relationship management (CRM) division included a team that monitored their social media presence through dashboards that provided real-time data on key topics, sentiments, and users. During the first 24 hours, the team monitored sentiments and key users in over 2,000 threads related to the increase, across various channels, noting that discussions were positive. However, during the same time frame, TelCorp's call centers observed a marked increase in customer complaints. Over the next 24 hours, various CRM teams carefully combed through all customer communications across channels and surmised that the problem was as follows. The majority of TelCorp's customers were subscribed to non-premium plans and either thought this offer applied to them and didn't notice improved performance, and/or were upset that it didn't apply to their plans. In hindsight, publicizing something that only applied to 20% of the customer base, and then poorly describing it in some of the social media channels, created a feeling of exclusion and/or confusion, leading to anger (i.e., a perfect storm of customer discontent). Exactly 54 hours after the initial announcement, the company made amends by introducing similar maximum upload speed increases for customers on non-premium plans, providing promotional offers on additional services and upgrades, and apologizing for the confusion. Nevertheless, over that 54-hour period, their customer churn rate was 5 times higher than usual, resulting in an estimated $110 million in lost revenue during the next 12-month period alone, not to mention long-term losses based on customer lifetime value.

In the era of viral media, it should not have taken TelCorp 48 hours to understand the gravity of the situation. Clearly, there was a need for enhanced sense-making capabilities. The TelCorp situation is not unique. There are many well-documented cases of organizations failing to appropriately make sense of employee and/or customer communications in internal and external-facing social media, resulting in significant financial consequences. Examples include employee relations at Wal-Mart (Berfield 2013), Gap's failure to understand customers' preferences during logo redesign (Halladay 2010), and Maker's Marks' production-related misstep (Lee 2013). In each of these incidents, sense-making from social media could have been used proactively to inform decision making, and/or reactively as part of a real-time monitoring strategy to mitigate damage. However, enhanced sense-making requires IT artifacts capable of effective text analytics. In the next section, we present an overview of LAP and describe how it can help improve the state-of-the-art for sense-making from social media. We also illustrate how the proposed LAP-based framework could facilitate enhanced sense-making in the context of TelCorp.

The Language-Action Perspective and Sense-Making in Online Discourse

Three important aspects of language are semantics, syntax, and pragmatics (Winograd and Flores 1986). Numerous prior technologies that support analysis of computer-mediated communication content have emphasized the semantics of language with particular focus on topics and sentiments of discussion; that is, what people are saying (Abbasi and Chen 2008). As new internet-enabled Web 2.0 based technologies gain widespread adoption in organizations, they are increasingly being used to facilitate communicative and discursive action involving employees, customers, partners, suppliers, etc. (Bughin and Chui 2010). While these technologies have great potential for supporting such activities, comprehensibility and clarity remain critical concerns: computer-mediated communication is highly incoherent (Herring 1999; Honeycutt and Herring 2009). Furthermore, the conventional Information System's perspective stresses the content of messages rather than the participants' interactive behavior (Aakhus 2007). There is a need for IT artifacts capable of accurately presenting pragmatic information such as communicative context and actions for enhanced sense-making (Schoop et al. 2006).

Design science provides concrete prescriptions for the development of IT artifacts, including constructs, models, methods, and instantiations (Hevner et al. 2004). Several prior studies have utilized a design science approach to develop business intelligence and analytics-related IT artifacts, including methods and instantiations (Abbasi and Chen 2008; Chau and Xu 2012; Lau et al. 2012). When creating IT artifacts in the absence of sufficient guidelines, design theories may help govern the development process (Storey et al. 2008; Walls et al. 1992). We use language-action perspective as a kernel theory to guide the development of the proposed framework and system (Winograd and Flores 1986).

The language-action perspective (LAP) emphasizes pragmatics; not what people say, but rather, what people do with language (Winograd and Flores 1986). LAP highlights "what people do by communicating, how language is used to create a common basis for communication partners, and how their activities are coordinated through language" (de Moor and Aakhus 2006, pp. 93-94). LAP's principles are based on several important theories, including speech act theory (Searle 1969), discourse analysis, and argumentation. Speech act theory (SAT) emphasizes the ordinary speaking view of language, where language is a social fact and its primary function is to promote sense-making in social interactions (Kuo and Yin 2011; Lyytinen 1985). Specifically, two LAP principles may provide important insights for the design and development of text analytics tools capable of improving sense-making from online discourse (Winograd and Flores 1986):

1. Conversation structures: LAP advocates considering messages in the context of the conversations in which they occur. Conversations encompass interactions between users and their messages. There are different types

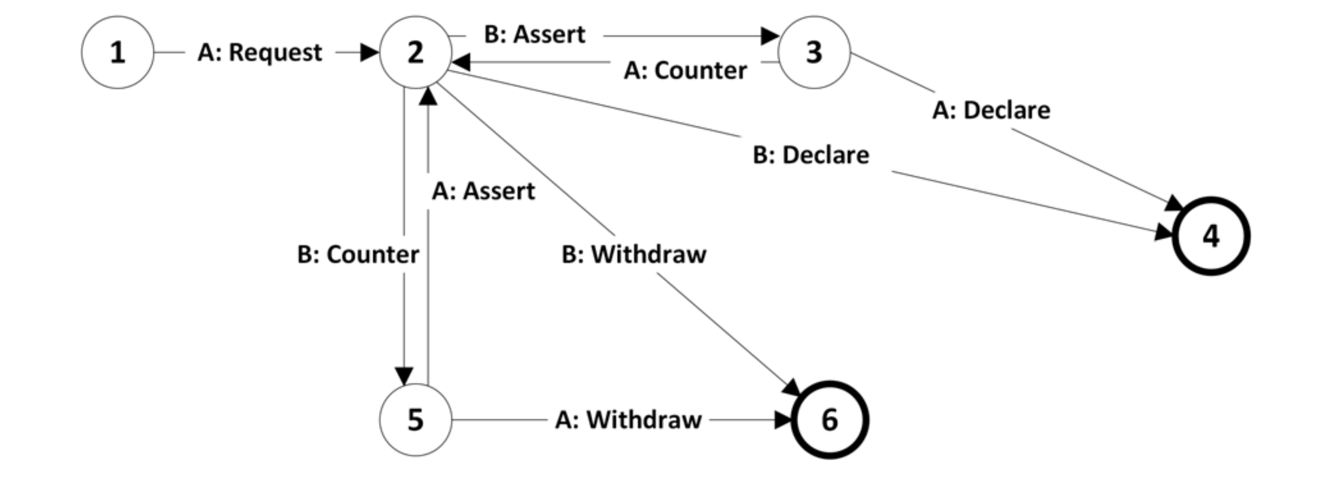


Figure 1. A Conversation for Clarification (adapted from Winograd and Flores 1986)

of conversations: conversations for action, conversations for clarification, conversations for possibilities, conversations for orientation, etc.

2. Actions and context: LAP advocates the pragmatic view, which can complement the semantic perspective by emphasizing actions, intentions, and communication context through consideration of speech acts.

Figure 1 presents a "conversation for clarification" example to illustrate the LAP principles, adapted from Winograd and Flores (1986). The example depicts two parties, A and B, and potential conversation sequences. For instance, A submits a request for information followed by B making an assertion, putting forth a counter request for additional information, declaring the issue resolved or inappropriate, or electing to withdraw from the conversation (and so on). The example shows a conversation template encompassing a collection of messages labeled with action information, multiple users, and their interactions (arrows). From an organizational social media analytics vantage point, the ability to analyze various types of conversations involving customers, employees, and other stakeholders can provide valuable sense-making capabilities which can complement the existing pervasive semantic view.

Despite the potential sense-making opportunities afforded by social media analytics guided by LAP, existing social media analytics tools used in organizational settings almost exclusively rely on semantics: analysis of topics and sentiments (Zabin et al. 2011). Accordingly, we propose a LAP-based framework for analyzing online discourse which emphasizes conversation structures, actions, and communication context (see Figure 2). The framework is predicated on the notion that methods which employ LAP principles can complement topic-sentiment-centric systems to facilitate enhanced sense-making through

1. Conversation disentanglement: the ability to accurately affiliate messages in discussion threads with their respective conversations. From a LAP perspective, conversations are an important unit of analysis that is presently not represented in text/social media analytics systems: messages are too atomic and threads encompass multiple intertwined conversations (Elsner and Charniak 2010).

2. Coherence analysis: the ability to infer reply-to relations among series of messages within a discussion thread (Nash 2005). Social media technologies make it difficult to accurately infer interrelations between messages (Honeycutt and Herring 2009), impacting quality of participant interaction and social network information (Aumayr et al. 2011; Khan et al. 2002).

3. Message speech act classification: the ability to infer the speech act composition of messages within discussion threads – for instance, assertions, questions, suggestions, etc. (Kim, Li, and Kim 2010).

Inclusion of these three components can be used to collectively improve sense-making capabilities by providing an enhanced representation of coherence relations and communication actions through the use of speech act trees (SATrees): the transformation of linear discussion threads into a series of conversations with reply-to relations and message speech act information. SATrees, and the information generated using

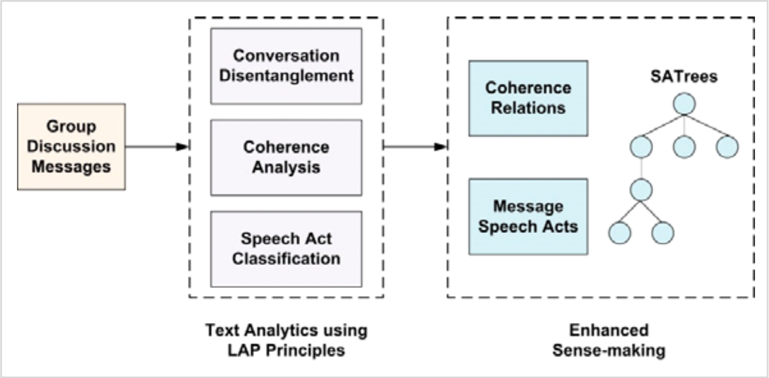

Group Discussion Messages
Conversation Disentanglement
Coherence Analysis
Speech Act Classification
Text Analytics using LAP Principles
Coherence Relations
Message Speech Acts
SATrees
Enhanced Sense-making

Figure 2. A LAP-Based Framework to Support Sense-Making in Online Discourse

LAP-based systems, can enable augmented support for key social media analytics use cases. The framework incorporates LAP concepts in two important ways. First, the composition and sequence of stages in the framework is closely aligned with LAP studies which emphasize conversations as the unit of analysis, interactions within these conversations, and the speech act composition of utterances (Winograd and Flores 1986). Second, within each component of the framework, principles from the LAP body of knowledge are used to prescribe design guidelines which are later operationalized through a LAP-based text analytics system. The proposed framework and related research questions are presented in the remainder of the section, along with discussion pertaining to the TelCorp example.

Conversation Disentanglement

A critical problem that arises in discourse are parallel, intertwined conversations (Elsner and Charniak 2010). Entangled conversations, which are highly prevalent in various forms of computer-mediated communication, occur as a result of multiple simultaneous conversations between two or more users appearing within a single discussion thread (Auramaki et al. 1992; McDaniel et al. 1996). In order to avoid thread confusion, disentanglement is widely regarded as an essential precursor for more advanced forms of discourse analysis (Adams and Martell 2008). It is especially important "when there are several streams of conversation and each stream must be associated with its particular feedback" (Te'eni 2001, p. 297). Consequently, in the proposed framework, disentanglement information/variables are key input for coherence analysis and speech act classification.

In order to illustrate the importance of conversation disentanglement, we revisit the TelCorp example. TelCorp examined sentiments in 2,000 discussion threads pertaining to its initiative. However, due to intertwined conversations, discussions threads are not the ideal unit of analysis (Honeycutt and Herring 2009). Figure 3 shows three initiative-related discussion threads taken from a web forum, Facebook, and Twitter, respectively. The threads were sampled from, and are representative of, the types of user-generated content found in the 2,000 threads pertaining to the initiative. In each thread, circles denote individual messages (e.g., a forum posting, a Facebook comment/reply, or a tweet). The vertical axes indicate thread turns, and the horizontal axes indicate conversations within the thread (with each column of circles signifying the messages in the same conversation). The arrows and boxes indicate the general topic of that particular conversation. As depicted in the figure, the web forum thread example encompassed six different conversations over a span of only 53 messages; the Facebook and Twitter threads, although shorter, also had 5 and 3 conversations, respectively. The initial conversations, which accounted for the majority of messages, were mostly positive expressions about the initiative—hence the positive thread-level sentiments observed by the monitoring team. However, some of the subsequent conversations drifted from positive, to questions, to criticisms, and even declarations of switching to other providers. Decomposing the threads to more meaningful semantic units by performing conversation-level analysis (Elsner and Charniak 2010) would have provided TelCorp's social media monitoring team with a better understanding of the situation.

This example underscores the importance of conversation disentanglement. Prior methods for disentanglement have mostly relied on single-pass clustering methods that compare newer messages against existing conversation clusters (e.g., Adams and Martell 2008; Shen et al. 2006; Wang and Oard 2009). While these methods utilize information regarding content similarity and spatial/temporal proximity between mes-

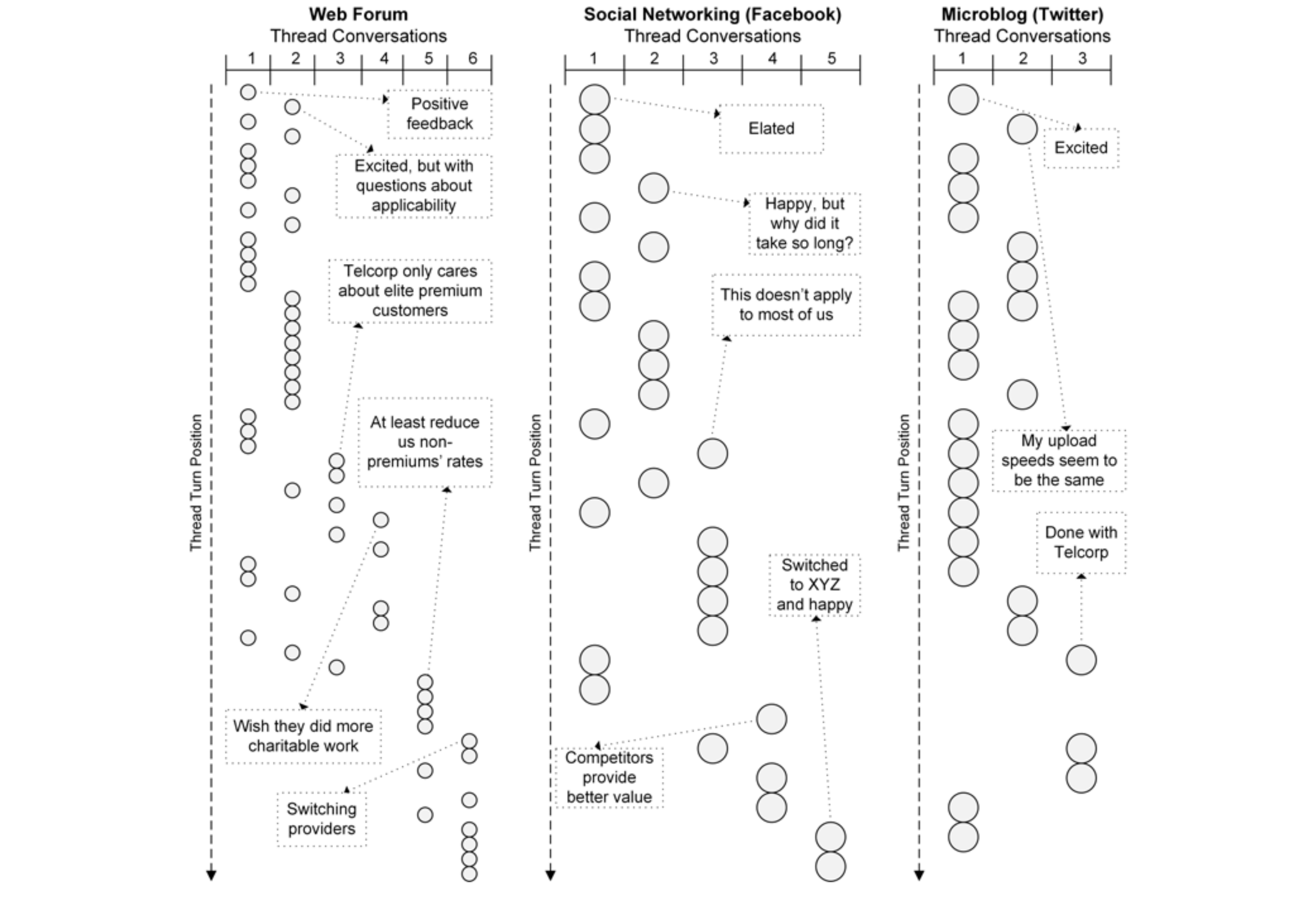

Web Forum
Thread Conversations
1
2
3
4
5
6
Thread Turn Position
Positive feedback
Excited, but with questions about applicability
Telcorp only cares about elite premium customers
At least reduce us non-premiums' rates
Wish they did more charitable work
Switching providers
Social Networking (Facebook)
Thread Conversations
Elated
Happy, but why did it take so long?
This doesn't apply to most of us
Switched to XYZ and happy
Competitors provide better value
Microblog (Twitter)
Thread Conversations
Excited
My upload speeds seem to be the same
Done with Telcorp

Figure 3. Illustration of Three Discussion Threads on Different Social Media Channels, Each with Multiple Conversations Related to TelCorp's Initiative

sages, they do not incorporate information pertaining to conversation structure. According to LAP, conversations are initiated by a specific illocutionary act, such as an assertion or a directive, subsequently followed by a finite sequence of acts (Kuo and Yin 2011; Winograd 1986). Hence, using LAP principles, a conversation can be decomposed into a beginning act succeeded by a series of "reacting" or "continuing moves" (Auramaki et al. 1992). A primitive message is a stand-alone assertion, and a derivative message is defined as a strictly logical or defeasible consequence of others (Raghu et al. 2001). Hence, primitive message identification is of great importance for disentanglement (Khan et al. 2002), as subsequent response messages are highly dependent upon it in terms of their illocutionary acts and propositional content (Kuo and Yin 2011; Winograd and Flores 1986). However, existing disentanglement methods do not attempt to explicitly identify primitive messages. Elsner and Charniak (2010, p. 405) used an empirical example to observe that a "detector for utterances which begin conversations could improve disentanglement scores." Given the importance of primitive messages, we pose the following question:

RQ1: *Will methods that emphasize conversation structure elements such as primitive message identification during the disentanglement process outperform existing techniques devoid of such information?*

Coherence Analysis

Text comprehension involves the construction of a coherent mental representation of situations described by texts. In online discourse, coherence is represented in terms of reply-to relationships between messages (Fu et al. 2008). However, communication technologies are susceptible to the sociotechnical gap—a gap between social requirements and technical feasibility (de Moor and Aakhus 2006). Jackson (1998) observed that there is a dichotomy between discourse practices

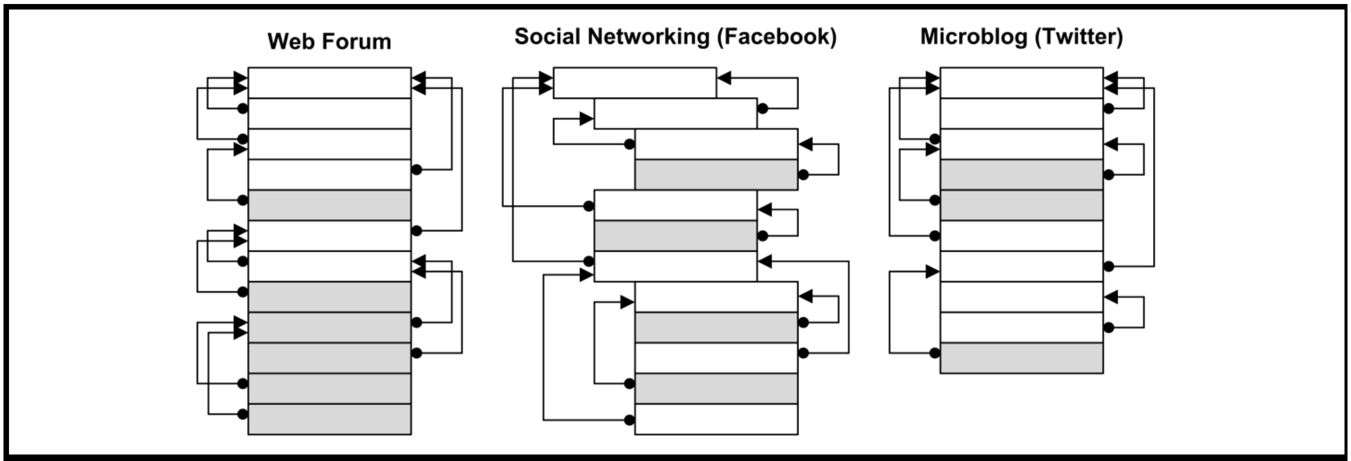


Figure 4. Reply-To Relations Between Messages in Web Forum, Facebook, and Twitter Discussions Pertaining to the TelCorp Initiative

and the tools intended to support online discussion. One such problem is "the imposition of a simple sequential ordering" (Jackson 1998, p. 192), which limits the effectiveness of temporal and spatial proximity-based system features. Consequently, social media discussions are highly susceptible to disrupted turn adjacency: a situation where adjacent messages in threads are often not related to one another, making threads highly incoherent (Herring 1999; Honeycutt and Herring 2009). For instance, 50% of messages in discussion threads do not respond to the previous or first post in the thread (Fu et al. 2008). Even in social networking sites such as Facebook, where users can comment on the original post or reply directly to prior comments, more than 30% of messages are incoherent (i.e., ambiguous with respect to reply-to relations). Similarly, microblogs such as Twitter, which were not originally designed to support conversations, are highly incoherent with respect to reply-to relations (Honeycutt and Herring 2009). Figure 4 shows examples of web forum, Facebook, and Twitter discussions pertaining to the TelCorp initiative. Each rectangle denotes a message; messages are ordered sequentially as they are generated (from top to bottom), while arrows indicate correct reply-to relations. Shaded messages are those deemed to be incoherent based on that particular social media channel's system-supported reply-to features. The illustrations only include the first 10 to 12 messages in the threads, and still 30% to 50% of the messages are out of place.

Coherence analysis attempts to offset the incoherent nature of online discourse by correctly reconstructing coherence relations among messages. Accurately attributing reply-to relations is critical to ensuring that participants' in-degree values are correct in social media-based social networks (Abbasi and Chen 2008; Anwar and Abulaish 2012). In the case of TelCorp, as later demonstrated, coherence analysis is critical to ensure proper sense-making of participant roles and centrality measures in online communities. Two important facets of coherence analysis are the features and techniques utilized. We review both and present a related research question in the remainder of the section.

Coherence Analysis Features

Three important categories of features used to identify coherence relations are system, linguistic, and conversation structure attributes. *System* features provide insights regarding the message context, including header (e.g., date/ime, message id, and subject/title) and quotation information (Abbasi and Chen 2008). For instance, Netscan extracted the "contents of Subject, Date, Organization, Lines, MessageID and Reference lines" to generate relationships in Usenet newsgroups, including conversation trees (Smith 2002). However, not all forms of group discussion contain a full range of system features, and the aforementioned sociotechnical gap hinders the utility of system features (Jackson 1998).

Linguistic features derived from message content can also provide important cues for coherence analysis. Common categories include direct address, co-reference, lexical relation, and semantic information (Donath 2002; Fu et al. 2008; Herring 1999; Nash 2005). Direct address occurs when a reply message includes the screen name of the author of a previous message (Donath 2002). Lexical relation is defined as a "cohesive relation where one lexical item refers back to another, to which it is related by having common referents" (Nash 2005). Co-reference also occurs when a lexical item refers to a previously posted lexical item; however, in this case the relation is implicit in that it can only be identified by the context (Soon et al. 2001). Nash (2005) divided co-reference into three subcategories: personal (e.g., use of pronouns), demonstratives, and comparatives (e.g., words such as "same" and "similar"). Examples of semantic information include opinions, emotions, synonymy information, parts-of-speech, etc. Such advanced NLP-based features have not been widely adopted (Abbasi and Chen 2008).

Group discussion is a repetitive process of subtopic/solution generation and evaluation. As previously alluded to, this process often results in simultaneous parallel conversations within a single discussion thread (Elsner and Charniak 2010). *Conversation structure features* are attributes that can shed light on the relations between messages and conversations within a discussion. Despite their importance for sense-making (McDaniel et al. 1996), conversation structure features have not been used much in previous coherence analysis research.

Coherence Analysis Techniques

Prior automated methods for coherence analysis include linkage, heuristic, and classification. *Linkage* methods construct interaction patterns using predefined rules that are primarily based on system features and assumptions regarding message sequences (Sack 2000). Most linkage methods employ two types of rules: direct linkage and naïve linkage (Fu et al. 2008). Direct linkage rules assume that users follow system features to post messages and clearly quote messages to which they respond. Naïve linkage rules are then applied to residual messages unidentified by direct linkage; these rules assume that all residual messages are responding to either the first message in the thread or the previous message (Comer and Peterson 1986). Linkage methods work fairly well with email-based discussion lists; however, as previously alluded to, social media is far less coherent. Nash (2005) manually analyzed 1,099 turns from Yahoo! Chat and found the lag between a message and its response to be as many as 100 turns. Herring and Nix (1997) concluded that nearly half of all turns were "off-topic." Consequently, linkage methods have performed poorly on web forums and chat (Abbasi and Chen 2008; Fu et al. 2008).

Heuristic methods rely on metrics derived from observations of online discourse (Fu et al. 2008). These metrics are based on a small, fixed, assumed set of communication patterns pertaining to system and/or linguistic features (Anwar and Abulaish 2012). For instance, the hybrid interactional coherence method uses an ordered list of heuristics, where messages unidentified by one heuristic are then evaluated by the next heuristic on the list (Fu et al. 2008). Khan et al. (2002) used finite state automata using linguistic features to identify interaction patterns in multi-person chat rooms. In many of these methods, the choice of heuristics (and their order) was based on prior observations of occurrence (Fu et al. 2008; Nash 2005). However, previous work has identified a plethora of different, context-specific discussion patterns and themes. In a group support system discussion involving 40 employees, Kuo and Yin (2011) noted that while 11 speech act patterns accounted for approximately 50% of the conversations, these patterns were very specific to, and dependent upon, the nature of the discussion topic. Similarly, Khan et al. (2002, p. 4) acknowledged the complexity caused by "factors such as number of participants, the topic(s) of chat, the familiarity of users with each other, etc." Consequently, the effectiveness of heuristic methods is predicated on the validity and generalizability of the set of heuristics incorporated.

Classification methods formulate coherence analysis as a binary classification problem (Aumayr et al. 2011). These techniques couple system and/or linguistic features with supervised machine-learning methods: predictive analytics algorithms that build models from a set of labeled training data (Wang et al. 2011). For example, in order to handle highly incoherent text from student online forums, Kim, Li, and Kim (2010) used supervised learning to classify discussion threads. Soon et al. (2001) adopted a machine learning approach to identify co-reference of noun phrases both within and across sentences which had been used for discourse analysis and language understanding.

The key gaps with respect to coherence analysis pertain to limited representational richness of feature sets and the need for classification methods capable of learning interaction patterns used in communication. Whereas few prior studies have used system, linguistic, and structure features in unison, as noted by prior studies based on LAP, linguistic and conversation structure features may help overcome the limitations of system features. Linguistic features allow users to assess relevance. Relevance is a critical component of a conversation; it requires "speakers to pick up elements from the preceding contributions appropriately and employ them in their own utterances" (Auramaki et al. 1992, p. 346). This process, which is analogous to leaving a trail of bread crumbs for fellow discussion participants, is essential for proper contextualization (Te'eni 2006). Similarly, conversation structure features that can help illuminate relations between messages and conversations are critical for identifying coher-

Table 2. Overview of Searle's Speech Acts

Speech Act	Description	Examples
Assertive	The speaker represents facts of the world.	statements that can be assessed as true or false
Commissive	The speaker commits to some future action.	agreement, support, disagreement, opposition, promises
Expressive	The speaker says something about his/her feelings or psychological attitudes.	apologies, congratulations, gratitude
Declarative	The speaker brings about changes in the world.	pronouncements, declarations, verdicts
Directive	The speaker gets the hearer to do something.	suggestions, questions, requests, commands, desires

ence relations (Auramaki et al. 1992; Winograd and Flores 1986). In summary, accurate identification of coherence relations necessitates the consideration of system, linguistic, and conversation information in conjunction with robust classifiers that can offer enhanced pattern recognition capabilities over linkage and heuristic methods (Wang et al. 2011).

RQ2: How extensively can classification methods that leverage conversation structure, linguistic, and system features outperform existing methods for coherence analysis?

Speech Act Classification

According to SAT, the minimal unit of an utterance is a speech act (Searle 1969). There are two distinct components of a speech act: the propositional content and the illocutionary force (Searle 1969). The propositional content is the topic of the utterance, while the illocutionary force describes the way in which it is uttered (Schoop 2001). Both elements must be considered in order to understand the speech act. Based on the illocutionary point, Searle (1969) defined five types of speech acts: assertive, directive, commissive, expressive, and declarative. Table 2 provides details regarding the five speech act categories.

Analysis of speech acts is useful for improving understanding of participant intentions (Te'eni 2006), an important problem for online discourse analysis (Mann 2011). While topic and sentiment analysis are essential components of any social media content analysis, they fail to capture underlying actions and intentions. Looking back at the TelCorp discussion threads depicted in Figure 2, the threads encompassed positive expressives in earlier conversations, followed by conversations comprised of questions, suggestions, assertions of indifference/negligence, negative expressives, and declarations of having switched to other providers. In other words, the threads encompassed many conversations for clarification (confusion) and conversations for action (churn) (Winograd and Flores 1986). Beyond what was being said, how and why were also important, especially with respect to customer confusion and churn.

Consequently, recent studies have explored automated methods for classifying speech acts in online discourse (Cohen et al. 2004; Kim, Wang, and Baldwin 2010; Moldovan et al. 2011). These methods have typically incorporated linguistic features such as bag-of-words and parts-of-speech tags in conjunction with machine-learning classification methods (e.g., Moldovan et al. 2011). However, speech acts are not individual unrelated events, but participate in larger conversational structures (Winograd and Flores 1986). While some prior methods leveraged basic information regarding speech act sequences (e.g., Carvalho and Cohen 2005), these studies failed to include a holistic representation of conversation structure such as that offered by conversation trees. Conversation trees have been used in prior social media analytics tools for visualizing conversation structures (Herring 1999; Smith 2002). They represent conversations as a tree comprised of coherence relations between parent, child, and sibling messages. Conversation trees can effectively represent the structure and flow of various conversations occurring within a discussion thread, thereby enabling enhanced representation of the relations dependencies among message speech acts.

RQ3: Will methods that utilize conversation trees attain enhanced speech act classification performance over existing methods that do not include such information?

Sense-Making

When performing sense-making tasks, users evaluate relevant costs and benefits associated with support technologies, including time, effort, and information quality (Russell et al. 1993). Hence, evaluation of sense-making artifacts requires assessment of information quality, the impact on users' sense-

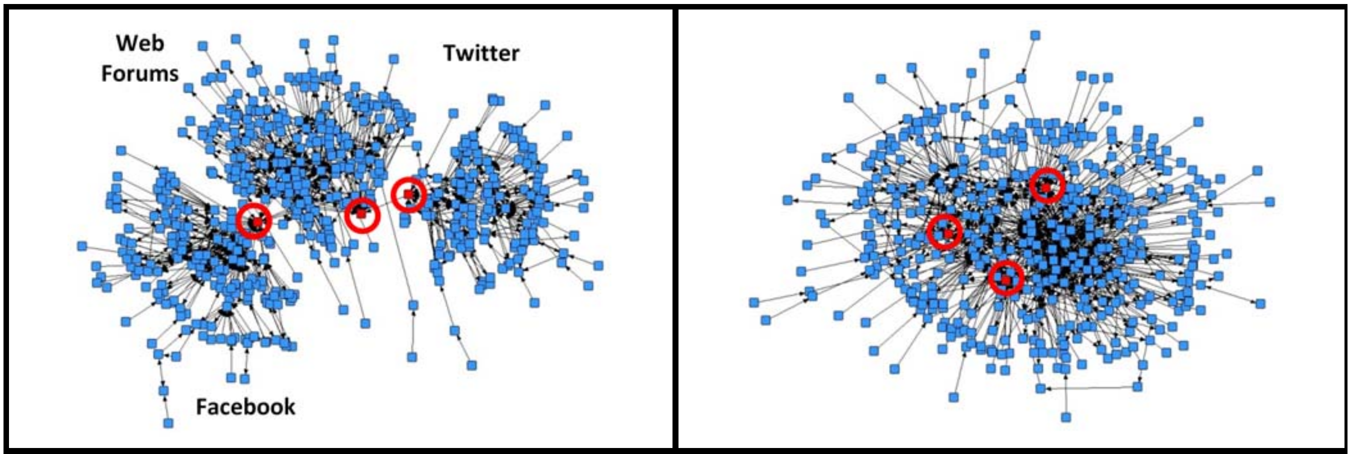


Figure 5. Social Media Social Networks for 50 TelCorp Initiative-Related Threads: Actual Network (left) and Constructed Network Using Existing Coherence Analysis Method (right)

making capabilities, and users' perceptions regarding costs and benefits (Pirolli and Card 2005).

Organizational use of social network analysis is on the rise (Mann 2013). From an organizational discourse perspective, important applications of social network analysis include identifying experts and influencers (de Moor and Aakhus 2006; Heracleous and Marshak 2004; Mann 2013). Given the prevalence of social network analysis in academia and industry, assessing the accuracy of social networks represents an important information quality evaluation for sense-making. For instance, the chart on the left in Figure 5 shows the *actual* social media interaction network for participants in 50 TelCorp initiative-related discussion threads encompassing web forums, Facebook, and Twitter. The interactions are generally intra-channel, with the exception of cross-channel links/mentions facilitated by three critical participants (circled). Interestingly, these three posted negative comments about the TelCorp initiative and garnered significant replies. Not surprisingly, these three discussants have the highest betweenness centrality values, as they serve as important bridges for the discussions occurring across the web forums, Facebook, and Twitter. However, in the interaction network *constructed* for the same threads (chart on the right Figure 5) using an existing state-of-the-art coherence analysis method, due to 30% misclassified reply-to relations, the network structure looks very different. In fact, the degree centrality measures in this constructed network for the *actual* top 20 discussants have mean absolute percentage error rates of over 40%, with over 50% of them not even being included in the top 20 of this network. Furthermore, the importance of the high-betweenness discussants (circled) is also significantly underestimated, with all three ranked outside the top 10 in terms of betweenness centrality in the network on the right. In this case, inadequate text analytic capabilities influenced TelCorp analysts' ability to identify key network members; a critical social media use case (Zabin et al. 2011).

As illustrated in this example, social networks derived from conversations can illuminate participant roles using measures such as degree centrality, betweenness, closeness, etc. (Fu et al. 2008). However, accurately computing these measures requires precise values for in-degree: the number of messages responding to a participant (Anwar and Abulaish 2012; Aumayr et al. 2011). Otherwise participant roles can be distorted; either exaggerated for some or understated for others (Fu et al. 2008).

RQ4: How extensively will enhanced coherence analysis attributable to LAP-based methods improve representation of social network centrality measures for discussion participants?

Ultimately, enhanced sense-making entails user involvement to reap the benefits of better text analytics (Russell et al. 1993; Weick et al. 2005). Visualization of discussion thread structure can coherently show the dynamics of communicative interaction and collaboration, and depict disentangled conversations (Donath 2002; Smith 2002). Similarly, depicting the speech act composition of messages can alleviate discourse ambiguity, a situation where participants are unclear as to the propositional content and/or illocutionary force of a message (Auramaki et al. 1988). However, demonstrating efficacy entails presenting the conversation, coherence, and speech act results to users. Accordingly, we employ SATrees: visualization of conversation trees where message nodes are labeled with their respective speech act information. As input, SATrees use methods for identifying conversations, coherence relations, and speech acts inspired by LAP principles.

It is important to note that our focus is not to develop a new visualization technique, but rather, to illustrate the utility of the underlying conversation disentanglement, coherence analysis, and speech act classification text analytics, which provides invaluable *input* for the SATree. Effective visualization is in itself a large research area (Donath 2002; Sack 2000; Smith 2002), beyond the scope of this paper. SATrees are merely labeled conversation trees (Honeycutt and Herring 2009) intended to provide a visual representation of coherence relations and illocutionary acts attributed to messages, allowing better understanding of conversation structure and flow, as well as participant intentions and group dynamics. Given the significance of information quality and coherence for sense-making (Weick et al. 2005), we present the following question:

RQ5: *Can SATrees facilitate enhanced user sense-making of online discourse compared to conversation trees generated using existing methods or the sequential message ordering approach commonly used by communication technologies?*

Further examining the sense-making value of an artifact within organizational settings, beyond short-term sense-making potential, entails field experimentation over an extended period of time. When performing sense-making tasks using supporting technologies longitudinally, users evaluate the utility of available methods in terms of their time/effort and information quality tradeoffs (Pirolli and Card 2005). "Collectively, these factors and tradeoffs form a cost structure guiding choices made during sense-making, including future usage of decision aids" (Russell et al. 1993, p. 269).

RQ6: *Will systems incorporating LAP-based text analytics garner greater perceived usefulness, actual usage, and productivity improvements over time than systems devoid of such information?*

A LAP-Based Text Analytics System for Sense-Making in Online Discourse

In the design science paradigm, kernel theories can be used to guide requirements for the design artifact, and both the theory and requirements can be used to inform design (Walls et al. 1992). Using LAP principles, in the previous section we presented the requirements: a framework for enhanced sense-making based on effective conversation disentanglement, coherence relations, and speech act classification. In this section we propose a design instantiation of the framework: a LAP-based text analytics system (LTAS) for sense-making in online discourse (Figure 6). LTAS has three major components: conversation disentanglement, coherence analysis, and speech act classification. For each discussion thread, the key outputs of the conversation disentanglement component are predictions of conversation beginnings and inter-message conversation affiliations, which serve as important conversation structure variables for the coherence analysis and speech act classification components. Within each discussion thread, the coherence analysis component leverages conversation structure information provided by the disentanglement component and basic speech act information, along with system and linguistic features, to output conversation trees encompassing finalized conversation affiliations and message reply-to relations. The output of the first two components is also leveraged by the speech act classification component, which uses conversation tree information to assign speech act labels to each message. The collective output of the system is an SATree, showing disentangled conversations within a discussion thread, with reply-to relations among messages that are labeled with their respective speech acts. As previously noted, SATrees signify the rich *types* of information offered by LTAS; this information can enable enhanced support for various social media analytics use cases as later demonstrated through user studies and a field experiment.

Prior LAP studies have emphasized close interrelatedness among conversations, coherence, and speech act compositions (Winograd and Flores 1986). In LAP, conversations form the building block for deeper analysis of interactions and speech act exchanges (Kuo and Yin 2011). Accordingly, LTAS considers the interplay of conversations, coherence, and speech acts. The output of the conversation disentanglement component is part of the input for coherence relations, since interactions are highly dependent on conversation context (Auramaki et al. 1992). Similarly, reply-to relations inform speech act classification since speech act composition for future messages within a conversation is dependent on those messages which precede them (Schoop 2001; Winograd and Flores 1986). Furthermore, each of the three components of LTAS leverages several important concepts from the discourse analysis and argumentation literature that have been incorporated into prior LAP-based studies, as summarized in Table 3. These concepts include context, relevance, conversation-beginning identification, thematization, discourse ambiguity, conversation structure elements, and message and conversation-level speech act composition. The three components of the system are discussed in the remainder of this section.

Conversation Disentanglement

The conversation disentanglement component of LTAS uses a two-stage approach. First, candidate primitive messages

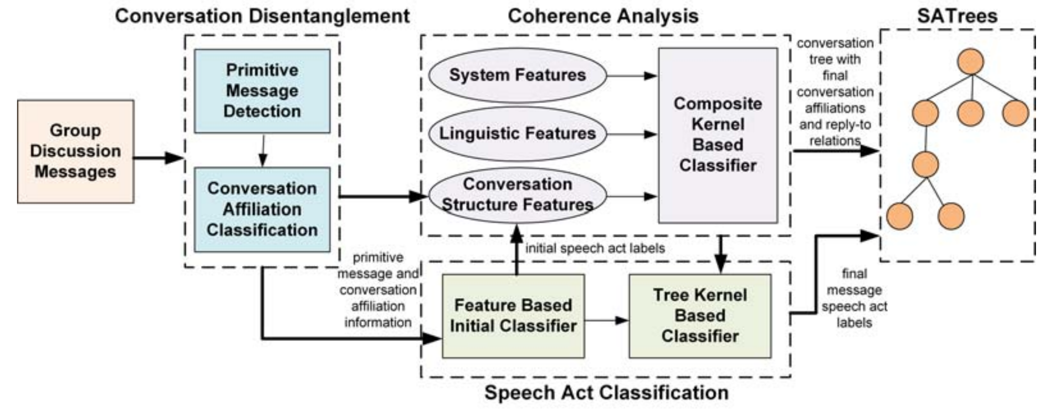

Conversation Disentanglement
Coherence Analysis
SATrees
Group Discussion Messages
Primitive Message Detection
Conversation Affiliation Classification
System Features
Linguistic Features
Conversation Structure Features
Composite Kernel Based Classifier
conversation tree with final conversation affiliations and reply-to relations
initial speech act labels
primitive message and conversation affiliation information
Feature Based Initial Classifier
Tree Kernel Based Classifier
final message speech act labels
Speech Act Classification

Figure 6. A LAP-Based Text Analytics System (LTAS) to Support Sense-Making in Online Discourse

Table 3. Select LAP-Based Principles Guiding Design of LTAS

LAP-Based Principle	Design Implications for LTAS
Interplay between conversations, interactions, and message acts (Winograd and Flores 1986)	Inclusion of three key system components, sharing of information between components for enhanced performance.
Importance of conversation beginnings as drivers of conversation structure, coherence relations, and conversation speech act composition (Auramaki et al. 1992; Winograd and Flores 1986)	Inclusion of the primitive message detection stage which provides key features to disentanglement, coherence analysis, and speech act classification components.
Contextualization and lexical chaining (Te'eni 2006)	Use of rich similarity measures between messages for conversation disentanglement and coherence analysis.
Thematization for uncovering conversation elements (Auramaki et al. 1992)	Inclusion of similarity bins from different regions to perform thread-level thematization for conversation affiliation classification.
Interdependency among speech acts (Auramaki et al. 1988; Kuo and Yin 2011; Winograd and Flores 1986)	Utilization of conversation tree-based message sequence patterns for speech act classification.

(i.e., conversation beginnings) are identified by using linguistic features to compute inter-message similarity. The features and output of the primitive message detection stage are then used as input for the second disentanglement stage. As previously discussed, prior conversation disentanglement studies have mostly used unsupervised clustering methods (e.g., Adams and Martell 2008; Wang and Oard 2009) and, to a lesser extent, supervised classification techniques with clustering overlaid (e.g., Elsner and Charniak 2010). We used supervised classification to garner enhanced precision and recall, and since conversation affiliations are not finalized until the coherence analysis component. The key outputs of our conversation disentanglement component are primitive message classifications and a pairwise message-to-message conversation affiliation classification (i.e., whether two messages belong to the same conversation), which serve as key conversation variables in the subsequent coherence analysis and speech act classification components. Details regarding the two-stage approach follow.

Primitive Message Detection

Participants in the same discussion thread often use contextualization to allow others to more easily understand conversation and coherence relations associated with their message (Te'eni 2006). One common approach for contextualization is lexical chains: the use of terms that are semantically related to terms appearing in prior messages within the same conversation (Auramaki et al. 1988). Therefore, an important cue regarding the conversation affiliation of a particular message is the degree of relevance between the message and topical themes of the existing conversations (Auramaki et al. 1992). Within a discussion thread, conversation beginnings (i.e., primitives) are messages that significantly deviate from existing conversations with respect to their topical themes (Aumayr et al. 2011; Khan et al. 2002). They are characterized by low topical similarity with messages that precede them, and high similarity with some of the messages that follow (Elsner and Charniak 2010). Conversely, non-primitive messages are likely to have higher similarity with at least some prior messages. Furthermore, while research has shown that as many as 20% of successive conversation messages can be separated by more than 10 turns within a forum thread (Nash 2005), or 5 tweets in a Twitter conversation (Honeycutt and Herring 2009), similarity between messages that are closer, both preceding and following, is typically of greater importance. For instance, many conversations exhibit topic drift: a gradual deviation from the starting point of a topic (Herring and Nix 1997). One implication of topic drift is that non-primitive messages may have higher max similarity with prior messages that are closer in proximity. Hence, message proximity and sequential trends are also important considerations for both primitive message detection in particular and conversation disentanglement in general.

The primitive message detection stage, depicted in Figure 7, leverages these important insights. It treats primitive message detection as a binary classification problem: predicting whether or not a given message within the discussion thread is a primitive. Let X represent a message in turn position p within a discussion thread of length l. All messages preceding X are placed into n roughly equal-sized bins, with each bin containing $(p\text{-}1)/n$ messages on average. Similarly, all messages following X within the thread are placed into n bins, each of size $(l\text{-}p)/n$ messages on average. Binning is used since discussion thread lengths vary and due to the fact that messages occur at different turns within a thread. Bins provide a consistent mechanism for representing message feature vectors in the statistical learning theory-based kernel function employed, while facilitating the inclusion of thematic trend information and proximity-sensitive similarity measurement. While the use of fixed-sized bins does present some limitations, as later discussed in the results section and Appendix C, binning also facilitates enhanced primitive message detection performance. Next, in order to capture information about lexical chains, we compute the average and max similarity scores between message X and messages within its surrounding $2n$ bins. For a given bin B_i, if $i \le n$, the average similarity $\text{Ave}\left\{\text{Sim}\left(X,B_i\right)\right\} = \sum_{Y\in B_i} \frac{\text{Sim}(X,Y)}{(p-1)/n}$, where Y is one of the $(p-1)/n$ messages in B_i. It is worth noting that for threads where $l < 2n$, $\text{Sim}(X, B_i) = 0$ if Bi is empty.

Many prior conversation disentanglement studies have used the vector space model (VSM) to represent the similarity between messages (Adams and Martell 2008; Wang and Oard 2009). In VSM, documents are typically represented with vectors of *tfidf*: term frequency multiplied by inverse document frequency (Adams and Martell 2008; Shen et al. 2006). *tfidf* downgrades the weight attributed to common terms. Similarities between *tfidf* document vectors are computed using the cosine similarity measure, with values ranging from 0 to 1, and higher values indicating greater similarity. Sim (X, Y) uses a document similarity measure with two important refinements: the use of parts-of-speech (POS) tag and synonymy information. Research has shown that noun phrases and verb phrases carry most of the important topical meaning in a sentence (i.e., the "bread crumbs" in the lexical chain), while conjunctions, adverbs, and adjectives are less important (Soon et al. 2001). Thus, we define meaningful terms to be nouns, noun compounds, named entities, verbs, and verb phrases. Instead of taking into consideration every term within a document, we only focus on ones with these POS tags, thereby narrowing the feature space to those terms most relevant to the lexical chain. Additionally, in group discussion text, users tend to use different words to express the same thing (Nash 2005). In other words, the "bread crumbs" in the lexical chain are not simply keyword repetition. A traditional VSM will treat synonyms or hypernyms as unrelated entries (Adams and Martell 2008). We take such information into consideration by computing a similarity value s_{tr} between two terms, which is incorporated into the *tfidf* calculation, thereby allowing better representation of semantic relations between messages. Accordingly, the similarity score between a pair of messages X and Y is as follows:

$$\text{Sim}\left(X,Y\right) = \frac{\sum_{t=1}^{k} w_{xt} \max_{r}\left(w_{yr}s_{tr}\right) + \sum_{r=1}^{j} w_{yr} \max_{t}\left(w_{xr}s_{tr}\right)}{2\sqrt{\sum_{t=1}^{k} w_{xt}^2}\sqrt{\sum_{r=1}^{j} w_{yr}^2}}$$

Where $w_{xt} = tf_{xt}idf_t$, t is one of the k unique terms in X, r is one of the j unique terms in Y, t and r are nouns, verbs, noun/verb phrases, or named entities, and s_{tr} is the similarity between t and r based on the shortest path that connects them in the is-a (hypernym/hypnoym) taxonomy in WordNet (Miller 1995). The set of nouns and verbs in WordNet includes many noun compounds, such as "prescription drug," and verb phrases, such as "give in" and "throw up." However, some noun compounds may not be present. In such cases, we compare the individual components of the noun compounds, and calculate s_{tr} as the average of the component-level similarities (Kim and

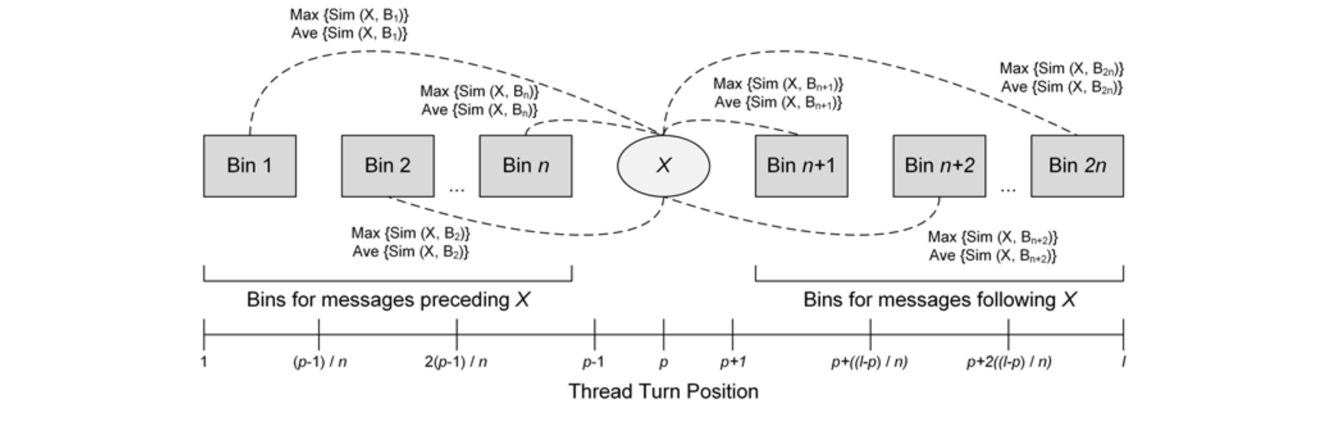


Figure 7. Illustration of Bins and Similarity Scores Used in Primitive Message Detection Stage

Baldwin 2005). For example, let's assume t = "customer service" and r = "client support." Assuming neither compound is present in WordNet, we compare the two head nouns "service" and "support" to one another, and two modifiers "customer" and "client." If the noun compound contains more than one modifier, the product of the similarities among various modifier combinations in tr is used (Kim and Baldwin 2005). A similar approach is taken for the verb phrases "intend switch" and "am leaving" from the statements "I intend to switch" and "I am leaving TelCorp." Appendix L empirically demonstrates the viability of our WordNet-based approach versus alternative state-of-the-art methods.

In the training data set, for each message X, the max and average Sim(X, B_i) are computed, resulting in a feature vector of length $4n$. These feature vectors constitute rows in the training data matrix, appended with class labels indicating primitive or non-primitive. Due to the class-imbalance, with non-primitives significantly outnumbering primitives, a moving threshold was adopted (Fang 2013). Such an approach has been shown to outperform traditional minority class over-sampling and majority class under-sampling methods in prior research (Fang 2013). See Appendix A for details. In this case, given classes i (X is not a primitive message) and j (X is a primitive message), let $p(X)$ represent the true classification probability of an unclassified instance X belonging to class i. Given training data set T, with each instance's class label $\in \{i, j\}$, and let c(i) denote the number of elements of T with class label equal to i, the classification Z = i if $p(X) \geq \frac{c(j)}{c(i)+c(j)}$, and Z = j otherwise (Fang 2013).

On each data set, we trained a support vector machine (SVM) classifier with a linear kernel on T, and applied it to each test instance X to generate p(X).

Conversation Affiliation Classification

Guided by prior LAP-based studies, stage two of the conversation disentanglement approach performs conversation affiliation classification. Traditionally, thematization has been proposed as a mechanism for linearizing a conversation to sequentially uncover important themes within a single conversation (Auramaki et al. 1992). The conversation affiliation classification stage performs what can be considered discussion thread-level thematization by utilizing conversation segments to infer whether two given messages are part of the same conversation (illustrated in Figure 8). Two critical components of this thematization strategy are inclusion of similarities from messages in surrounding regions to the two messages of interest and inclusion of primitive message information. The intuition for the proposed method is as follows. Conversations are collections of messages. Consequently, many prior methods have employed clustering methods for grouping messages based on inter-message similarity (e.g., Adams and Martell 2008). In addition to the similarity between two messages themselves, similarity to other messages within the thread "can provide further evidence to the semantics" (Wang and Oard 2008, p. 204). Given that message lengths in social media may introduce sparsity in linguistic feature vectors, which can impact similarity assessments, evaluating similarity with other messages can improve robustness, acting as a message similarity evidence "expansion" strategy (Wang and Oard 2008). Primitive message information is included since similarity relative to conversation beginnings is a key conversation affiliation cue, providing insights into discussion schisms, topic drift, and floor tracking (Elsner and Charniak 2010). Consequently, the successful inclusion of such information is believed to be capable of boosting affiliation classifications by at least 5% to 10%

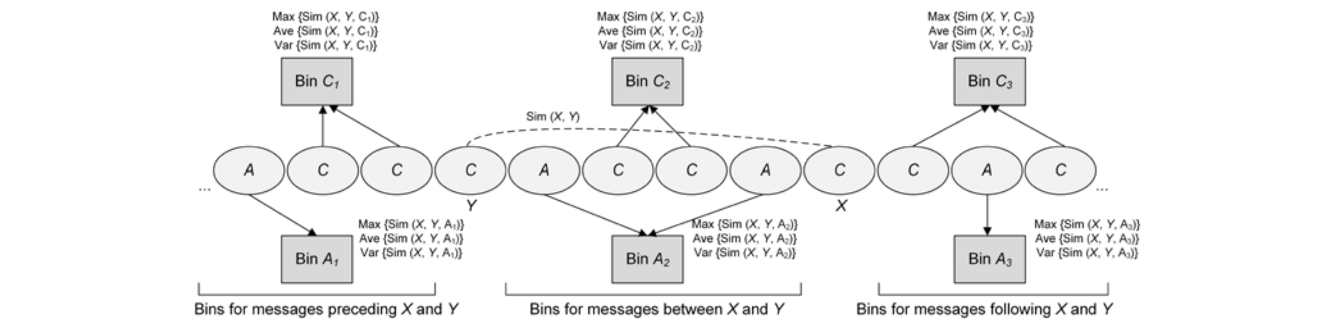


Figure 8. Illustration of Region, Bins, and Similarity Scores Used in the Affiliation Classification Stage

(Elsner and Charniak 2010). Our own experiment results presented later support the importance of primitive messages.

This intuition is operationalized as follows. Based on the output from the primitive message detection stage, all messages within the thread are labeled primitive or non-primitive (denoted by *A* and *C* in Figure 8, respectively). All message pairs within the thread are compared and classified as either belonging to the same conversation or not, as follows. For a given message pair *X* and *Y*, three conversation regions are derived: region 1 for messages preceding *X* and *Y*, region 2 for messages between *X* and *Y*, and region 3 for messages that follow *X* and *Y*. In addition to the similarity between X and *Y* (i.e., Sim (*X*,*Y*)), within these three regions, the difference in similarity between *X* and *Y* with respect to primitive (A_1, A_2, A_3) and non-primitive (C_1, C_2, C_3) message bins are leveraged using average, max, and variance measures. For a given bin C_i, the average similarity $\mathrm{Ave}\{\mathrm{Sim}(X,Y,C_i)\} = \sum_{Z \in C_i} \frac{|\mathrm{Sim}(X,Z) - \mathrm{Sim}(Y,Z)|}{d}$, where *Z* is one of the *d* messages in the non-primitive bin C_i. The maximum and variance measures are computed in a similar manner. For instance, $\mathrm{Max}\{\mathrm{Sim}(X,Y,C_i)\} = \max_{z}(|\mathrm{Sim}(X,Y) - \mathrm{Sim}(Y,Z)|)$. It is important to note that if *X* and *Y* are adjacent messages, Ave/Max/Var{Sim(*X*, *Y*, C_2)} and Ave/Max/Var{Sim(*X*, *Y*,A_2)} are all 0 since C_2 and A_2 are empty. The intuition for incorporating average and max similarity is based on the use of similar cluster centroid and nearest-neighbor style measures in past studies (Adams and Martell 2008; Shen et al. 2006; Wang and Oard 2009). Variance was included since the preceding, between, and following message region sizes can vary considerably as thread length increases, impacting average and max similarity values, and as a gauge for intertwined conversations within the region.

In the training data set, for each message pair *X* and *Y*, the max, average, and variance attributes from the three regions as well as Sim (*X*,*Y*) are derived, resulting in a feature vector encompassing 19 independent variables and the yes/no class label indicating whether *X* and *Y* belong to the same conversation. As with the primitive message detection stage, threshold moving was utilized for conversation affiliation classification to alleviate class imbalance for the linear SVM classifiers when applied to threads in the test set (Fang 2013). The output of the conversation disentanglement module of LTAS are two-fold: (1) classification of primitive messages within a thread and (2) classification of each message pairs' conversation affiliations (i.e., whether they belong to the same/different conversations). This information is leveraged extensively as input variables in the coherence analysis and speech act classification components of LTAS, as discussed in subsequent sections.

Coherence Analysis

Consistent with prior work (Kim, Li, and Kim 2010), the identification of coherence relations is modeled as a binary classification problem, where each message pair in the discussion thread either constitutes a reply-to relation or does not. The attributes used are three feature vectors for each message pair: system, linguistic, and conversation structure features. These feature vectors are inputted into a composite kernel function for an SVM classifier. Details are as follow.

Coherence Analysis Features

Table 4 shows the various system, linguistic, and conversation structure features derived for each message pair *X* and *Y*, where *X* precedes *Y* within the discussion thread. System features include those commonly used in prior studies, including the message proximity in turns (Nash 2005), temporal distance in minutes (Aumayr et al. 2011), and whether *Y* includes system-generated quoted content from *X* (Abbasi and Chen 2008; Smith 2002). Messages closer in turn or temporal prox-

Table 4. Features of Candidate Message Pairs

Category	Feature	Description
System Features	Turn Proximity	Turn index of message *Y* – turn index of message *X*
	Temporal Distance	Timestamp of message *Y* – timestamp of message *X* (in minutes)
	Quoted Content	Whether *Y* contains system-generated quoted content from *X*
	Reply-To	Whether *Y* contains system-generated reply to *X* in header, subject, or title
Linguistic Features	Lexical Relation	Sim (*X*,*Y*) based on formulation presented in Section 4.1
	Direct Address	Whether *Y* references screen name of author of *X*
	Co-reference	Whether *X* and *Y* have personal pronouns and comparatives (4 features)
	Sentiment Polarity	Whether *X* and *Y* are objective or subjective (2 features)
	Length Difference	Length of *X* (in words) – length of *Y*
Conversation Structure Features	Message Status	Whether messages *X* and *Y* are primitive messages (2 features)
	Conversation Status	Whether messages *X* and *Y* are part of the same conversation
	Between Status	Number of primitive messages between *X* and *Y*
	Prior Status	Number of primitive messages prior to *X* and *Y*
	Speech Act	Speech act classifications for messages *X* and *Y* (2 features)
	First Message	Whether *X* or *Y* are the first message in the discussion thread

imity are more likely to have a reply-to relation between one another (Aumayr et al. 2011; Honeycutt and Herring 2009; Nash 2005). While turn proximity has been shown to provide utility in prior coherence analysis studies (Fu et al. 2008), its effectiveness is diminished by the sociotechnical gap; in this case through the imposition of a simple, sequential ordering (Jackson 1998).

As previously alluded to, linguistic features are important for understanding contextual elements and lexical relations between messages (Auramaki et al. 1992; Te'eni 2006), and therefore have important implications not only for conversation disentanglement, but also for coherence analysis. We use several important linguistic features. The lexical relation between messages (Nash 2005) is derived using the Sim(*X*, *Y*) formulation described in the "Conversation Disentanglement" section. Direct address indicates whether message *Y* explicitly references the screen name of the author of message *X* (Fu et al. 2008). The four co-reference features indicate whether *X* and *Y* each include the following two implicit lexical chain elements: personal pronouns (e.g., your) and comparatives (e.g., worse) (Soon et al. 2001). The two sentiment polarity features indicate whether *X* and *Y* contain subjective or objective content, respectively. Subjective messages are those that have greater sentiment polarity (Abbasi and Chen 2008; Lau et al. 2012). Sentiment information is useful since users often express their opinion towards a prior message with positive polarity (e.g., "I like your idea") or negative polarity ("I think that's a terrible suggestion"). Sentiment lexicons such as SentiWordNet provide an effective mechanism for inferring sentiment polarity (Esuli and Sebastiani 2006). We adopt a straightforward approach to determine whether a message is subjective or objective, where each term in a message is compared against items in the sentiment lexicon to compute a subjectivity score on a 0–1 scale (with higher values indicating greater subjectivity). SentiWordNet contains a positive, negative, and neutral polarity score ranging from 0 to 1 for each term. Our sentiment feature is the average, across all terms in the message, of each term's (positive + negative score)/2. Message length information can be a useful coherence relation cue, especially when combined with speech act features. For instance, shorter agreement messages are less likely to be responded to by lengthier messages (Kim, Wang, and Baldwin 2010).

As noted in prior LAP and discourse analysis studies, coherence relations and salient underlying interaction cues are highly dependent upon conversation context (Fu et al. 2008; Khan et al. 2002). Conversation disentanglement information is essential in order to reduce the likelihood of creating coherence relations between messages from different conversations (Elsner and Charniak 2010). Since interactions are highly dependent on the context surrounding the conversations in which they occur (Winograd and Flores 1986), six types of conversation structure features are utilized based on the conversation disentanglement component described earlier. The two message status attributes are the primitive/nonprimitive message classifications from the primitive message detector. Obviously, if message *Y* is deemed primitive, it is less likely to be responding to *X*. However, if *X* is a primitive and *Y* is not, the likelihood of a reply-to relation increases since conversation beginnings typically attain more responses than non-primitive messages (Elsner and Charniak 2010; Fu et al. 2008). Similarly, the conversation status feature is the conversation affiliation classification for *X* and *Y*. The primitive message detector is also the basis for the between status and prior status attributes. Since primitive messages attain more replies, greater between and prior status may reduce the likelihood of a reply-to relation. As previously alluded to, conversations, interactions, and speech acts are closely interrelated (Winograd and Flores 1986). Hence, the speech acts for *X* and *Y* are included as attributes, predicted using the "initial classifier" described later in the section "Initial Classifier."

Coherence Analysis Technique

Consistent with prior work (Kim, Li, and Kim 2010), the training corpus is comprised of all positive and negative (i.e., non-reply-to cases) reply-to cases encompassed in a collection of conversations. For a given message, negative cases are all previous messages with which it does not have a reply-to relation. The number of negative cases considerably exceeds the number of positive cases, warranting the use of threshold moving as done in the conversation disentanglement experiments (Fang 2013).

Once the features between all message pairs in the training set discussion threads have been extracted, a composite kernel is used to leverage the system, linguistic, and conversation structure feature categories in an ensemble-like manner (Szafranski et al. 2010). In part, the beauty of kernel-based methods such as SVM lies in their ability to define a custom kernel function *K* tailored to a given problem, or to use the standard predefined kernels (e.g., linear, polynomial, radial basis function, sigmoid, etc.). When dealing with classification tasks involving diverse patterns, composite kernels are well-suited to incorporate broad relevant features while reducing the risk of over-fitting (Collins and Duffy 2002; Szafranski et al. 2010). In our case, diversity stems from differences in the occurrence of system, linguistic, and conversation structure features across users, social media channels, and/or industries. In Appendix K we present further background on kernel methods and empirically demonstrate the proposed composite kernel's effectiveness versus a single SVM classifier.

Let s_i, l_i, and c_i represent the system, linguistic, and conversation structure feature vectors for a given message pair *X* and *Y*. We define a combinatorial ensemble of kernels $K = \{K_1, \ldots, K_Q\}$ encompassing all combinations of linear composite kernels involving s, l, and c (here Q = 7 due to $2^3 - 1$ total combinations). Given two instance rows in the training data matrix, their similarity is defined based on the inner product between all combinations of their three vectors s_1, l_1, c_1, and s_2, l_2, and c_2. For instance,

$$K_1(s_1,s_2) = \frac{\langle s_1,s_2\rangle}{\sqrt{\langle s_1,s_1\rangle\langle s_2,s_2\rangle}}, \quad K_2(l_1,l_2) = \frac{\langle l_1,l_2\rangle}{\sqrt{\langle l_1,l_1\rangle\langle l_2,l_2\rangle}},$$

$$K_4(s_1+l_1,s_2+l_2) = \frac{\langle s_1,s_2\rangle}{\sqrt{\langle s_1,s_1\rangle\langle s_2,s_2\rangle}} + \frac{\langle l_1,l_2\rangle}{\sqrt{\langle l_1,l_1\rangle\langle l_2,l_2\rangle}},$$

$$K_5(s_1+c_1,s_2+c_2) = \frac{\langle s_1,s_2\rangle}{\sqrt{\langle s_1,s_1\rangle\langle s_2,s_2\rangle}} + \frac{\langle c_1,c_2\rangle}{\sqrt{\langle c_1,c_1\rangle\langle c_2,c_2\rangle}}.$$

The composite kernel K_σ is the combination of these Q kernels: $K_\sigma = \sum_{q=1}^{Q} \frac{K_q}{Q}$. The SVM classifier trained using this kernel outputs a prediction confidence score for each instance (scores are real numbers), where negative numbers indicate a non-reply-to classification and values greater than or equal to zero indicate positive reply-to relation classifications. Hence, for a message *Y* in a discussion thread, we attain predictions for each message *X* that precedes it. Since a given message in a conversation may reply to multiple prior messages, in theory, if *Y* is preceded by 10 messages in the discussion thread, the classifier outputs may predict 0 to 10 reply-to relations originating from *Y*. However it is worth noting that in our data sets as well as in prior research, multi-replies happen very infrequently (in less than 1% or 2% of instances). Though not done in this study, some prior research has used a fixed "single reply-to relation from a message" rule to reduce false positives. Irrespective, to evaluate coherence analysis relations, metrics such as precision and recall of positive reply-to relation classifications are typically adopted.

The output of the coherence analysis component is a conversation tree encompassing the finalized disentangled conversations and message reply-to relations within the discussion threads. Most studies represent conversations as trees with a single parent for each child node (Herring 1999; Smith 2002). In order to leverage a tree structure here as well, we create a duplicate node for each message (and its subtree) with multiple reply-to relations, under each of its respective parent nodes (as illustrated in Appendix F).

Speech Act Classification

Within a conversation, speech act occurrences are closely related to one another, with subsequent speech acts highly dependent upon those speech acts which precede them

(Stolcke et al. 2000; Winograd and Flores 1986). In order to represent these interdependencies, prior methods incorporated information regarding the transition probabilities between speech act pairs (Carvalho and Cohen 2005). While such information is highly useful, speech acts are part of the larger overall conversation structure (Winograd and Flores 1986). To represent such information more holistically, the speech act classification component of LTAS uses a two-stage approach comprised of an initial classifier and a tree kernel-based classifier. The initial classifier employs attributes derived using system, linguistic, and conversation structure information to provide an initial speech act label for each message in the conversation tree. The kernel method then uses this labeled tree as input to improve performance by leveraging important facets of conversation structure.

Initial Classifier

The feature set used by the initial classifier consists of content attributes and contextual attributes. The content attributes include (1) binary/presence vector for all nouns and verbs appearing at least three times in the training corpus, lemmatized with their part-of-speech information; (2) whether or not the message has sentiment; and (3) whether or not the message is deemed a primitive message by the classifier described earlier. Emphasis is placed on nouns and verbs since prior research has shown that these two parts-of-speech are strong indicators of message speech act composition (Carvalho and Cohen 2005 Cohen et al. 2004; Stolcke et al. 2000). Sentiment information is often present in commissive and expressive speech acts (Kuo and Yin 2011).

The contextual attributes extracted for each message pertain to primitive message and thread length and proximity information: (4) the distance from the closest preceding primitive message in the thread, in message turns, as a percentage of total messages in the thread; (5) the total number of preceding primitive messages in the thread; (6) the total number of messages in the thread; and (7) the position of the message in the thread, as a percentile. These attributes are intended to capture basic conversation context information from the discussion thread. For instance, depending on the context, certain speech acts such as assertives and directives are more likely to begin a new conversation, whereas expressives often appear later in conversations (Kuo and Yin 2011). Other studies have also noted the varying occurrence probabilities of certain speech acts at different stages of a conversation (Carvalho and Cohen 2005; Winograd and Flores 1986). Similarly, lengthier threads are more likely to have commissive and directive speech acts that extend the discussion through agreement, disagreement, follow-up questions, etc. (Rowe et al. 2011). The position of a message in the thread, as a percentile, has been shown to be a useful contextual attribute for speech act classification (Wang et al. 2011).

The features are input into a series of linear SVM classifiers. Since SVMs are binary-class classifiers, for each pair of speech act combinations (e.g., assertives and expressives, assertives and commissives, etc.), a separate SVM classifier is constructed. Test messages are evaluated by each of the binary classifiers and assigned to the classes receiving the highest aggregate prediction scores across classifiers (Szafranski et al. 2010). The output of the initial classifier is a speech act category prediction for each test message.

Labeled Tree Kernel-Based Classifier

Conversation structures vary considerably depending upon their speech act compositions. For example, conversations for action often begin with a declarative, followed by a series of commissives, declaratives, and assertives (Winograd and Flores 1986). Similarly, conversations for clarification, possibilities, and orientation each have distinct structural and composition-related elements. Coherency is important for understanding the stage structure of a discourse, and consequently, the relations between speech acts (Auramaki et al. 1988). In order to leverage coherence relations, we propose a novel labeled tree kernel classifier (Figure 9). Kernel-based methods are useful since custom kernels can incorporate rich structural information into the learning process (Abbasi et al. 2010; Collins and Duffy 2002). As input, the classifier uses a labeled conversation tree constructed using coherence relations and message speech act labels. The coherence relations are based on the coherence analysis component of LTAS, while message speech act labels are generated using the initial classifier. For illustrative purposes, let's assume our speech act label set L = {A, C, D, E} for assertive, commissive, declarative, and expressive.

For each message y_i in the test set Y, we extract a sub-tree Sy_i comprised of parent, child, and sibling nodes. Figure 9 illustrates how the sub-tree for the test message originally labeled "D" by the initial classifier is extracted. Parent message is the one that D replies to, child messages are ones replying to D, and sibling messages are ones that share the same parent message as D. In the extracted sub-tree, the label for the message of interest is always changed to "?".

For each message x_i in the training set X, we extract sub-tree Sx_i. Training sub-trees are also derived by applying the initial classifier and coherence analysis classifier using 10-fold cross-validation on the training data. While we could simply incorporate the gold-standard coherence relations and message speech act labels for the training sub-trees, we found that

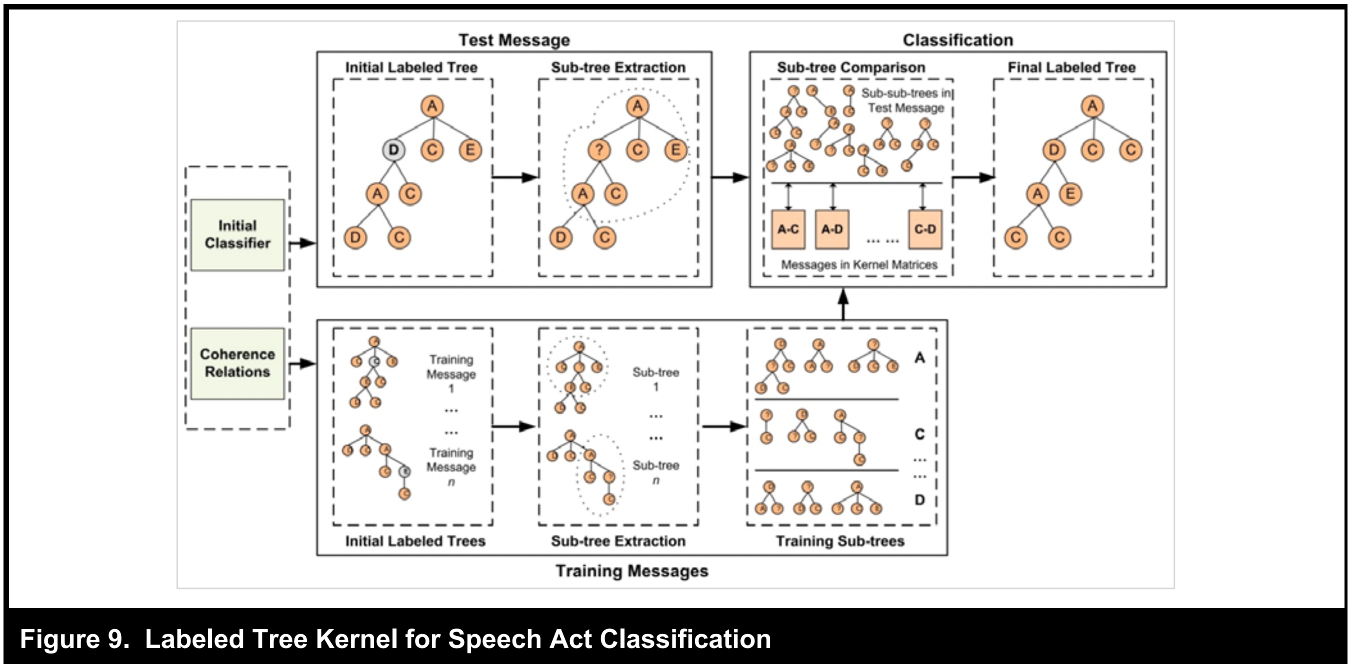


Figure 9. Labeled Tree Kernel for Speech Act Classification

using the same classifiers on the training/testing data improved performance by allowing input classifier biases to be incorporated into the kernel classifier's learning process. This process results in a collection of training message sub-trees for each speech act class, as depicted in the "Training Sub-trees" component of Figure 9.

Classifier training is performed as follows. For each pair of speech act classes in L, a separate kernel matrix K is constructed on the training data. For instance, K_{AC} is comprised of similarity scores K_{AC} (x_i, x_j) between each pair of training messages in X_{ac}, the subset of X with class label assertive or commissive, intended to learn patterns to differentiate assertives from commissives. K_{AC} (x_i, x_j) is a similarity measure between Sx_i and Sx_j computed by comparing all tree fragments in Sx_i and Sx_j, where a fragment is defined as any sub-graph containing more than one node (Collins and Duffy 2002). K_{AC} (x_i, x_j) is simply equal to two times the number of common fragments in Sx_i and Sx_j, divided by the total number of fragments in Sx_i and Sx_j. Formally, let $h_k(x_i)$ denote the presence of the k^{th} tree fragment in Sx_i (where $h_k(x_i) = 1$ if the kth tree fragment exists in x_i) such that Sx_i is now represented as a binary vector $h(x_i) = (h_1(x_i),\ldots,h_n(x_i))$:

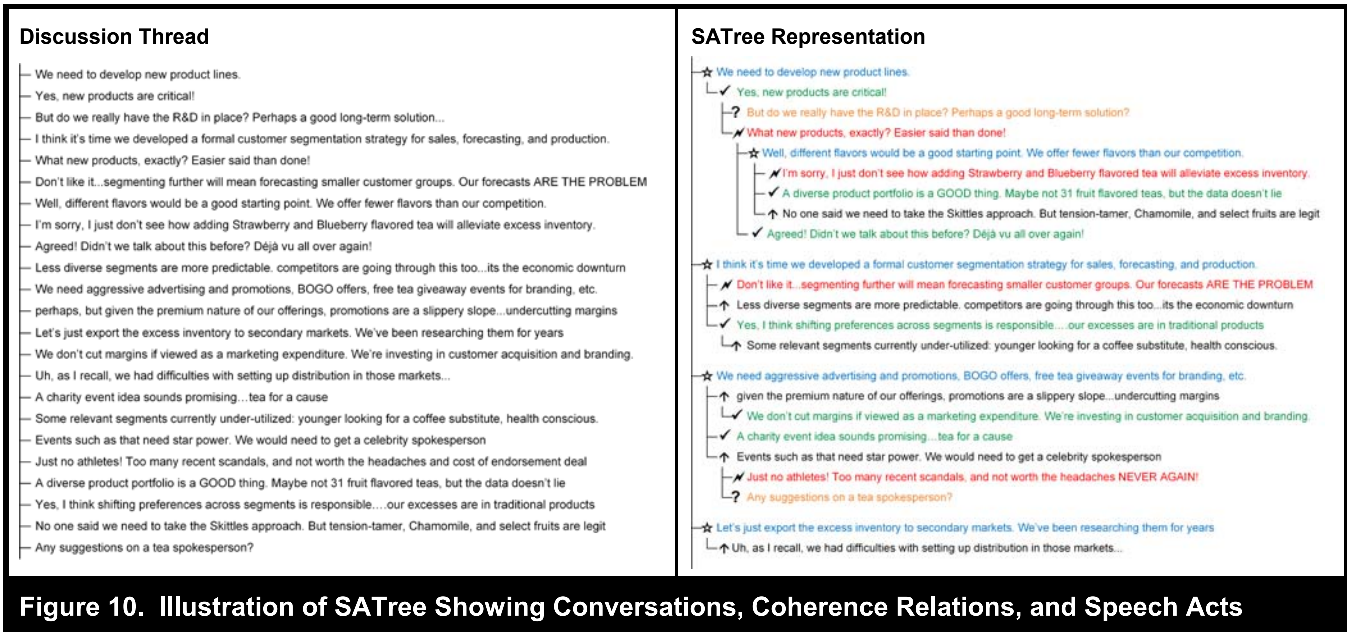
$$K_{AC}\left(x_i,x_j\right)=\frac{2\sum_{k=1}^{n}h_k\left(x_i\right)h_k\left(x_j\right)}{\sum_{k=1}^{n}h_k\left(x_i\right)+\sum_{k=1}^{n}h_k\left(x_j\right)}$$

Similar to the process described in the "Coherence Analysis Technique" section with respect to the coherence analysis classifier, each K is used to build a separate binary classifier for each speech act label pair using SVM Light (Joachims 1999). In Figure 9, the trained models are depicted by boxes in the classification section (e.g., A-C, A-D).

Test message y_i is classified by all of the trained binary SVM models, each of which takes a vector of sub-tree comparison-based similarity scores as input. For instance, the A-C classifier would take $(K_{AC}(x_i, y_i), \ldots, \mathrm{K}_{\mathrm{AC}}(x_z, y_i))$ as input, where $|X_{ac}|= z$, and output a prediction score. Voting across the binary classifiers is used where the final speech act label for each y_i is the class receiving the highest aggregate prediction score. The eventual outcome is a final labeled tree for each conversation in the test set.

Speech Act Tree (SATree)

The conversation disentanglement, coherence relation, and speech act classification components of LTAS are combined to create an SATree for each group discussion. Figure 10 presents an example of an SATree. In the tree, each branch represents a conversation; nodes under those branches represent messages in the conversations. Symbols to the left of each message are used to indicate speech act composition; for example, assertions ↑, directive-suggestions ☆, directive-questions **?**, commissives ✓, and expressives ↗. Even from this small example, it is apparent that this particular discus-



Discussion Thread

- We need to develop new product lines.
- Yes, new products are critical!
- But do we really have the R&D in place? Perhaps a good long-term solution...
- I think it's time we developed a formal customer segmentation strategy for sales, forecasting, and production.
- What new products, exactly? Easier said than done!
- Don't like it...segmenting further will mean forecasting smaller customer groups. Our forecasts ARE THE PROBLEM
- Well, different flavors would be a good starting point. We offer fewer flavors than our competition.
- I'm sorry, I just don't see how adding Strawberry and Blueberry flavored tea will alleviate excess inventory.
- Agreed! Didn't we talk about this before? Déjà vu all over again!
- Less diverse segments are more predictable. competitors are going through this too...its the economic downturn
- We need aggressive advertising and promotions, BOGO offers, free tea giveaway events for branding, etc.
- perhaps, but given the premium nature of our offerings, promotions are a slippery slope...undercutting margins
- Let's just export the excess inventory to secondary markets. We've been researching them for years
- We don't cut margins if viewed as a marketing expenditure. We're investing in customer acquisition and branding.
- Uh, as I recall, we had difficulties with setting up distribution in those markets...
- A charity event idea sounds promising…tea for a cause
- Some relevant segments currently under-utilized: younger looking for a coffee substitute, health conscious.
- Events such as that need star power. We would need to get a celebrity spokesperson
- Just no athletes! Too many recent scandals, and not worth the headaches and cost of endorsement deal
- A diverse product portfolio is a GOOD thing. Maybe not 31 fruit flavored teas, but the data doesn't lie
- Yes, I think shifting preferences across segments is responsible….our excesses are in traditional products
- No one said we need to take the Skittles approach. But tension-tamer, Chamomile, and select fruits are legit
- Any suggestions on a tea spokesperson?

SATree Representation

- ☆ We need to develop new product lines.
  - ✓ Yes, new products are critical!
    - ? But do we really have the R&D in place? Perhaps a good long-term solution?
    - ⚡ What new products, exactly? Easier said than done!
      - ☆ Well, different flavors would be a good starting point. We offer fewer flavors than our competition.
        - ⚡ I'm sorry, I just don't see how adding Strawberry and Blueberry flavored tea will alleviate excess inventory.
        - ✓ A diverse product portfolio is a GOOD thing. Maybe not 31 fruit flavored teas, but the data doesn't lie
        - ↑ No one said we need to take the Skittles approach. But tension-tamer, Chamomile, and select fruits are legit
      - ✓ Agreed! Didn't we talk about this before? Déjà vu all over again!
- ☆ I think it's time we developed a formal customer segmentation strategy for sales, forecasting, and production.
  - ⚡ Don't like it...segmenting further will mean forecasting smaller customer groups. Our forecasts ARE THE PROBLEM
  - ↑ Less diverse segments are more predictable. competitors are going through this too...its the economic downturn
  - ✓ Yes, I think shifting preferences across segments is responsible….our excesses are in traditional products
    - ↑ Some relevant segments currently under-utilized: younger looking for a coffee substitute, health conscious.
- ☆ We need aggressive advertising and promotions, BOGO offers, free tea giveaway events for branding, etc.
  - ↑ given the premium nature of our offerings, promotions are a slippery slope...undercutting margins
    - ✓ We don't cut margins if viewed as a marketing expenditure. We're investing in customer acquisition and branding.
  - ✓ A charity event idea sounds promising...tea for a cause
  - ↑ Events such as that need star power. We would need to get a celebrity spokesperson
    - ⚡ Just no athletes! Too many recent scandals, and not worth the headaches NEVER AGAIN!
    - ? Any suggestions on a tea spokesperson?
- ☆ Let's just export the excess inventory to secondary markets. We've been researching them for years
  - ↑ Uh, as I recall, we had difficulties with setting up distribution in those markets...

Figure 10. Illustration of SATree Showing Conversations, Coherence Relations, and Speech Acts

sion encompasses multiple conversations, some of which have elaborate interaction patterns and diverse message speech act compositions. Appendix O presents an extended illustration of how the conversation structure, reply-to relation, and message speech act composition information encompassed in SATrees can support key social media use cases such as identifying issues, suggestions, and key participants. It is also important to reiterate that our focus is not to develop a new visualization technique, but rather, to illustrate the utility of the underlying conversation disentanglement, coherence analysis, and speech act classification text analytics encompassed in LTAS, which provides invaluable *input* for the SATree based on LAP. Effective visualization is in itself a large research area (Donath 2002; Sack 2000). The visualization style employed for SATree was inspired by visual dynamic topic analysis diagrams (Honeycutt and Herring 2009).

Evaluation

Consistent with Hevner et al. (2004), a series of experiments were conducted to evaluate the effectiveness of various components of our LTAS text analytics system and underlying LAP-based framework. The six experiments, which were closely aligned with the questions presented earlier, were set up in three parts. Part 1 was intended to demonstrate the superiority of the proposed LAP-based system (LTAS) relative to state-of-the-art methods for text-based sense-making through a series of data mining experiments which follow. Experiment 1 assessed the effectiveness of the conversation disentanglement component (RQ1). Experiment 2 evaluated the usefulness of using linguistic and conversation structure features in conjunction with system features and a robust classification method (RQ2). Experiment 3 assessed the speech act component of the system (RQ3).

Part 2 showed the efficacy of LTAS for sense-making, through data and user experiments involving sense-making tasks, conducted in four different organizational settings. Specifically, experiment 4 empirically demonstrated enhancements in information quality for social network centrality measures (RQ4), while experiment 5 illustrated how SATrees could allow practitioners to improve sense-making from online discourse as compared to existing methods (RQ5).

In part 3 of the evaluation, we used a field experiment (experiment 6) to demonstrate the business value of LTAS over a 4-month period, where the social media monitoring team members using LTAS garnered enhanced issue identification capabilities estimated by TelCorp to be worth millions of dollars. Collectively, these three forms of evaluation demonstrate the art of the possible, practical, and valuable for text analytics grounded in the pragmatic view.

Working closely with our industry collaborators, the experiments related to parts 1 and 2 were performed on 10 group discussion data sets spanning four industries: telecommunications, health, security, and manufacturing. The 10 data sets encompassed several important social media channels used routinely for both intra-organizational and customer-facing communication, collaboration, and engagement, including web

Table 5. Overview of Test Bed

Domain or Industry	Channel	Description	No. of Threads	Messages		Particip. Per Thread	Convo. Per thread
				Total	Per Thread		
Telecom	Web Forum	Telus forum postings on DSLReports	69	2608	37.8 (20.0)	18.7 (9.9)	4.3 (2.7)
	Social Network	Telus Facebook fan page comments	208	3209	15.4 (4.1)	4.5 (1.1)	2.6 (0.9)
	Microblog	Telus-related tweets	228	2403	10.5 (2.3)	4.0 (1.0)	1.8 (0.6)
Health	Web Forum	Prescription drug posts on Drugs.Com	66	2764	41.9 (28.4)	13.2 (10.4)	6.2 (4.8)
	Social Network	Drug comments on PatientsLikeMe	128	2026	15.8 (5.4)	9.5 (3.3)	1.7 (1.3)
	Microblog	Prescription drug-related tweets	383	2905	7.6 (2.1)	3.1 (0.9)	1.3 (0.5)
Security	Web Forum	McAfee posts on Bleeping Computer and Malwarebytes	65	3491	53.7 (23.3)	25.2 (13.9)	6.1 (3.3)
	Social Network	McAfee Facebook fan page comments	180	2471	13.7 (3.5)	5.3 (2.0)	2.1 (0.7)
	Microblog	McAfee-related tweets	268	2445	9.1 (2.4)	3.5 (0.9)	1.6 (0.6)
Manufacturing	Chat	Comments on tea bag over-production	20	835	41.8 (14.0)	4.0 (0.0)	6.8 (3.1)
Total			**1,615**	**25,157**			

*A separate training set encompassing a similar quantity of data per domain/channel was used by LTAS/comparison methods (Appendix H).

forums, social networking sites, micro-blogs, and group chat (Bughin and Chui 2010; Mann 2013). Table 5 provides an overview of the data sets, including the number of discussion threads, total number of messages, and messages/participants/conversations per thread (mean and standard deviation). The total test bed included over 25,000 messages associated with 1,615 discussion threads. Looking at Table 5, we make a few observations about the test bed. Web forum discussion threads tend to be lengthier (and involve more participants) than those appearing in social networking sites such as Facebook and Patients Like Me, or on microblogs like Twitter (Fu et al. 2008; Honeycutt and Herring 2009). As later observed, these channels also varied considerably in conversation structure, dynamics, interaction patterns and cues, and speech act composition. These differences made inclusion of a variety of industries and channels important to ensure a robust evaluation test bed.

The telecommunications data sets pertained to Telus, one of the three largest telecommunications service providers in Canada. In the telecommunications industry, customer churn is a big problem (ACSI 2014). Consequently, industry leaders such as Telus rely heavily on social media monitoring and analytics for brand reputation management, better understanding pain points, and to derive customer-related insights (Kobielus 2011). Since Telus' social media presence and their online mentions span several channels, three different data sets were included. The Telus forum on DSLReports.com allows current, past, and prospective customers to discuss services and issues pertaining to Telus' cable and high-speed internet offerings. Visitors of Telus' Facebook fan page post comments regarding the company's community outreach initiatives, on-going promotions, and their personal experiences with Telus' mobile, home phone, and cable/Internet services. The third telecommunications data set was comprised of Twitter discussion threads mentioning Telus and/or the company's products and services.

The health data sets were social media discussions of prescription drug offerings from Merck KGaA's major competitors. The three data sets included threads from the Drugs.com web forum, Twitter, and the social networking site Patients Like Me. In these social media channels, users talk about their experiences, potential side-effects, other adverse reactions, ask questions, and seek advice. As post-marketing drug surveillance using social media gains popularity, organizations also seek to leverage such information for competitive intelligence and demand forecasting (Adjeroh et al. 2014; Zabin et al. 2011).

The security data sets were comprised of web forum postings, Facebook fan page comments, and tweets related to McAfee, Inc. and its security software, respectively. In the discussion threads, customers talk about observed strengths and weaknesses, problems encountered, and their overall experiences with McAfee's B2C offerings, as well as those of competitors. Insights derived from analysis of such social media content have important implications for operations and product strategy (Mann 2011; Zabin et al. 2011).

The manufacturing discussion test bed was derived from a series of group support system (GSS) chat-based discussions. The data was comprised of 20 discussion threads involving 4 participants each; 80 total participants that were all experienced with the GSS software employed. Each of the 20 threads focused on the discussion topic of how to best address the overproduction problem for a tea bag manufacturer. Subjects were told to discuss solutions. Whereas the other nine data sets were derived from external-facing web forums, social networking sites, or micro-blogs, this data set differed one important way: it was comprised of chat sessions with a more internal-facing perspective.

It is important to note that due to the need for manually annotating a gold standard for each thread/message, most *labeled* social media and/or text document test beds used in prior studies appearing in top IS journals have typically used 5,000 documents/messages or fewer (e.g., Abbasi and Chen 2008; Lau et al. 2012). From that perspective, the test bed incorporated in this study is fairly extensive and robust with respect to the total volume of data as well as the variety of industries, domains, and social media channels incorporated. Consistent with prior studies (Fu et al. 2008; Lau et al. 2012; Kuo and Yin 2011), all data sets in the test bed were rigorously labeled by two independent human annotators with backgrounds in linguistics and experience in discourse analysis (Honeycutt and Herring 2009; Nash 2005). Additionally, these annotations were further validated by practitioner social media analysts. See Appendix H for details.

Experiment 1: Conversation Disentanglement

In the first experiment, we evaluated the effectiveness of the conversation disentanglement component of LTAS, which utilizes primitive message detection as a precursor to conversation affiliation classification. LTAS was compared against several existing disentanglement methods, most of which utilized VSM-based features to compute similarity between messages, which were then used as input for clustering methods. Choi (2000) performed segmentation using VSM applied to bag-of-words and clustering based on the Euclidean distance between messages. Wang and Oard (2009) also used VSM on bag-of-words and single-pass clustering. However, they incorporated information regarding the author and temporal and conversational context (e.g., posting author information, time between messages, and direct address). Shen et al. (2006) used VSM applied to bag-of-words coupled with additional linguistic features and messages weighted by time as input for a single-pass clustering algorithm. Adams and Martell (2008) used VSM with bag-of-words, hypernym information, a message distance penalty, as well as direct address information. Elsner and Charniak (2010) performed disentanglement using word repetition and discourse-based features, time windows, and direct address as input for a maximum entropy algorithm. For all comparison methods, parameters were tuned retrospectively in order to yield the best possible results. See Appendix H for details. Consistent with prior work, micro-level precision, recall, and f-measure were used as our performance measures (Shen et al. 2006).

Table 6 shows these f-measures. Precision and recall values can be found in Appendix N. LTAS outperformed all five comparison methods by a wide margin on all ten data sets. The performance lift was consistent for precision, recall, and f-measure. In most cases, LTAS was 15% to 20% better than the best competing methods. Paired t-tests were conducted to evaluate LTAS against the comparison methods. The tests were performed on the f-measures for the 1,615 discussion threads (i.e., n = 1,615). LTAS significantly outperformed all five comparison methods (all p-values < 0.001). The results presented here (RQ1), as well as further analysis presented in Appendices B, C, and E, underscore the efficacy of the primitive message detection-oriented LTAS method as a viable method for conversation disentanglement.

LTAS performed better across all 10 data sets spanning different industries and social media channels. Figure 11 shows the f-measures for LTAS and comparison methods across each of the 1615 discussion threads. The chart on the left shows mean f-measures for threads encompassing 1 to 10+ conversations. The chart on the right shows mean f-measures by thread length percentile rankings (with lower percentile values on the horizontal axis indicating shorter thread lengths). Not surprisingly, the f-measures of all techniques declined as the number of conversations and messages per thread increased. Interestingly, although LTAS performed better across the board, the performance margins were greater on threads with a higher number of conversations and/or messages (i.e., the right half of each of the two charts in Figure 11). Whereas the average f-measures of the two best comparison methods dipped by 22% to 35% or more, LTAS's performance dropped by only about 15% to 18%. The enhanced performance was largely attributable to LTAS's emphasis on identifying primitive messages (i.e., conversation beginnings). Analysis revealed that LTAS correctly identified approximately 85% of the primitive messages whereas com-

Table 6. F-Measures for Conversation Disentanglement Experiment on Various Channels

	Telco			Health			Security			Manu.
Method	**Forum**	**Social**	**Twitter**	**Forum**	**Social**	**Twitter**	**Forum**	**Social**	**Twitter**	**Chat**
LTAS*	**70.6**	**84.2**	**88.5**	**69.0**	**72.6**	**87.0**	**72.5**	**78.6**	**90.3**	**68.0**
Elsner and Charniak (2010)	45.9	62.6	73.6	48.8	59.9	78.6	46.0	59.2	72.7	37.7
Adams and Martell (2008)	48.4	61.6	64.2	44.3	51.9	68.1	48.3	56.7	63.7	44.6
Shen et al. (2006)	37.3	58.7	61.8	40.6	58.9	65.2	37.1	55.0	65.2	28.9
Choi (2000)	26.8	51.9	53.7	24.4	56.6	52.5	26.3	51.1	52.5	24.3
Wang and Oard (2009)	30.9	40.3	45.8	28.9	59.8	43.1	30.4	42.6	43.1	33.0

*Significantly outperformed comparison methods, with all p-values < 0.001

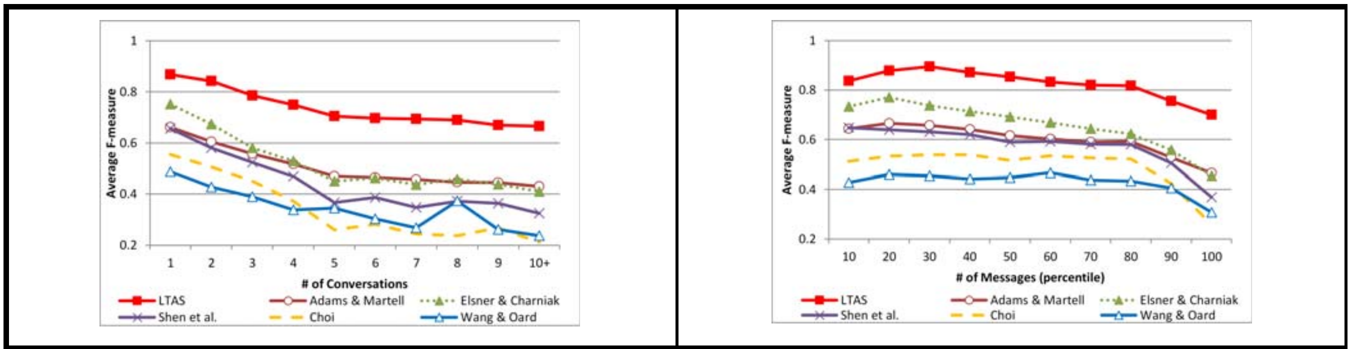


Figure 11. Average F-Measures for LTAS and Comparison Methods across Discussion Threads Grouped by Number of Conversations (left) and Number of Messages (right)

parison methods typically only detected 60% of primitives. LTAS was also more accurate at identifying marginal messages. Another factor was that LTAS only included terms with noun or verb parts-of-speech to compute similarity between messages, whereas the comparison methods did not incorporate parts-of-speech information. These factors resulted in better conversation disentanglement, with margins being more pronounced as the number of conversations and messages per discussion thread increased.

Experiment 2: Coherence Analysis

In the second experiment, we evaluated the effectiveness of the coherence analysis component of LTAS against existing classification, heuristic, and linkage techniques. LTAS uses system, linguistic, and conversation structure features for coherence analysis, as described earlier. While few studies have leveraged system, linguistic, and conversation structure features in concert, we examined the use of all three feature categories in conjunction with a robust classification method embodying LAP principles. Consistent with prior work, we treated this as a binary classification problem: whether the latter message in a pair replied to the earlier one or not. However, in this classification problem, we were only interested in those message pairs that were classified as having a reply-to relation. While the number of pairs that were classified as having no reply-to relationships was much larger, including these instances in the performance evaluation would have artificially inflated precision and recall rates for all experiment settings. Thus, our precision and recall metrics were based only on correctly classified reply-to relationships.

We compared LTAS against existing heuristic, linkage, and classification methods for coherence analysis. The heuristic-based method (Fu et al. 2008) relied on three linguistic features derived from the message body: direct address, lexical similarity, and residual match. The direct address match identified coherence relations based on references to user/screen names. Lexical similarity between messages was derived using VSM. A naïve linkage-based residual match rule was applied to the remaining messages (Comer and Peterson 1986; Fu et al. 2008).

The classification-based method used linguistic and system features (Kim, Li, and Kim 2010). We extracted four types of features from the message pairs: "time_gap" and "dist" were the interval of time and distance between message pairs, respec-

Table 7. F-Measures for Coherence Analysis Technique Comparison Experiment

	Telco			Health			Security			Manu.
Method	Forum	Social	Twitter	Forum	Social	Twitter	Forum	Social	Twitter	Chat
LTAS*	**81.1**	**87.2**	**91.0**	**78.7**	**80.1**	**86.4**	**81.0**	**83.7**	**92.5**	**84.8**
Heuristic	59.0	51.5	71.6	52.2	53.4	73.8	54.4	59.7	74.5	56.1
Classification	58.0	57.4	78.8	50.9	56.8	81.6	50.7	65.4	78.4	43.5
Linkage-Previous	38.9	44.6	71.1	33.1	38.2	70.3	29.9	53.9	69.0	21.7
Linkage-First	35.9	32.6	52.2	26.2	32.0	61.9	27.2	42.1	51.3	13.7

*Significantly outperformed comparison methods, with all p-values < 0.001

tively. "repeatNoun" was the number of repeated nouns between message pairs, and "viewer_timeGap" examined the time interval for messages pairs from the same author. The linkage methods used available system features and assumed all residual messages (i.e., ones not containing any system-based interaction cues) were replying to either the previous message (Linkage-Previous) or the first message (Linkage-First).

Table 7 shows the f-measures. Precision and recall values can be found in Appendix N. LTAS outperformed the comparison heuristic, linkage, and classification methods by a wide margin in terms of thread-level f-measures (all paired t-test p-values < 0.001, n = 1,615). With respect to comparison methods, the poor performance of the linkage methods was attributable to disrupted turn adjacency and lack of system-based interaction cues. Particularly in the case of the web forums and chat data sets, over 70% of the time adjacent messages in the discussion thread did not have a reply-to relationship with one another. Furthermore, many messages in these data sets were not replying to the first message. Consequently, Linkage-Previous and Linkage-First yielded poor results on web forums and chat. The comparison classification method also attained lower precision and recall. This was attributable to limitations in the coverage provided by the classifier's rules, which were mostly based on system features related to message proximity and time gaps. The limited use of linguistic features and lack of conversation structure attributes contributed to the classification method's low recall. While the heuristic method performed better than the classification method on web forums and chat, its performance was adversely affected by the utilization of discourse pattern-related assumptions that did not hold as well, particularly in the context of social networking sites and Twitter.

Figure 12 shows the f-measures for LTAS and comparison methods across each of the 1,615 discussion threads. The chart on the left shows mean f-measures for threads encompassing 1 to 10+ conversations. The chart on the right shows mean f-measures by thread length percentile rankings (with lower percentile values on the horizontal axis indicating shorter thread lengths). As with the conversation disentanglement results presented in the previous section, all coherence analysis techniques' f-measures declined as the number of conversations and messages per thread increased. However, once again, although LTAS performed better across the board, the performance margins were greater on threads with a higher number of conversations and/or messages. Whereas the average f-measures of the two best comparison methods dipped by 15% to 30% or more, LTAS's performance dropped by 10% or less. The was partly attributable to the inclusion of conversation structure features which allowed lengthier threads to be "decomposed" into smaller conversations, making accurate coherence analysis classifications more feasible (see Appendices D and F for further details). The results demonstrate the efficacy of the proposed coherence analysis method, which combines system, linguistic, and conversation structure features with a robust classification method.

Experiment 3: Speech Act Classification

Speech acts are important for understanding communicative actions and intentions (Janson and Woo 1996; Te'eni 2006). Consistent with prior work, the annotators labeled six categories of speech acts using the approach previously described (Moldovan et al. 2011; Stolcke et al. 2000): assertives, suggestions and questions (directives), expressives, commissives, and declaratives. The final annotation results are presented in Figure 13. Across the various data sets in the test bed, messages were concentrated along the assertive, directive, commissive, and expressive speech acts. In other words, messages were primarily statements, suggestions, questions, agreement/disagreement, and sentiments/affects. Interestingly, due to the problem-solving nature of discussion in the web forums, suggestions were more prevalent and expressives occurred less frequently relative to prior studies (e.g., Kuo and Yin 2011; Twitchell et al. 2013). Conversely, in Facebook and Twitter discussions, expressives such as opinions,

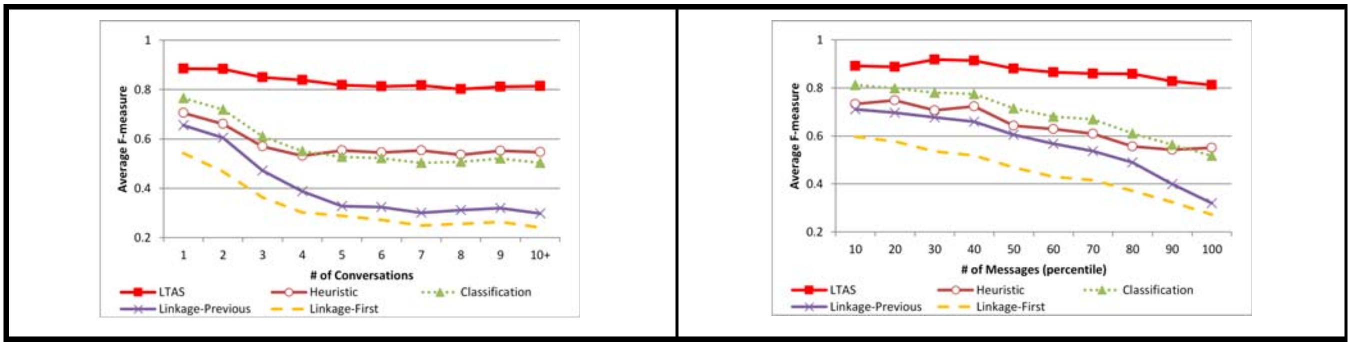


Figure 12. Average F-Measures for LTAS and Comparison Methods across Discussion Threads Grouped by Number of Conversations (left) and Number of Messages (right)

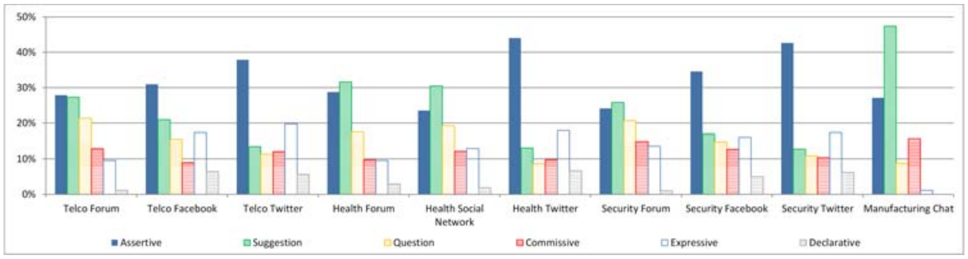


Figure 13. Speech Act Composition across Data Sets in Test Bed

sentiments, and emotional content were more prevalent. The tea manufacturing group chat discussions involved an ideation task; such discussions are generally rich in questions and suggestions (Kuo and Yin 2011). Declaratives accounted for less than 5% of messages in most data sets. Their limited occurrence is consistent with previous work (Kuo and Yin 2011). Speech act annotation details appear in Appendix H.

We compared the speech act classification component of LTAS against several existing methods. For all methods, the settings yielding the best results were reported. The *n*-Word method extracts the first *n* tokens and their associated POS tags for each message, where *n* ranges between 2 and 6 (Moldovan et al. 2011). These attributes are then used as input for a decision tree classifier. In our experiments, we set *n* to 2 since it yielded the best results. The n-gramSVM method proposed by Cohen et al. (2004) attained the best results on our test bed when using unigrams (i.e., single words) and bigrams (i.e., word pairs) with a linear SVM classifier. Kim, Wang, and Baldwin (2010) used lexical and conversation context features that included the frequency of lemmatized token and POS tag combinations, message position relative to thread length, and whether the posting author was the thread initiator. These features were input into a conditional random fields (CRF) classifier. Collective classification iteratively improves speech act predictions using a series of underlying local classifiers that rely on bag-of-words and relational features such as the speech act labels of parent/child nodes (Carvalho and Cohen 2005). Joint classification utilizes a conditional random field meta-learner with an embedded dependency parsing classifier as well as conversation context, semantic, and message relation attributes (Wang et al. 2011).

The evaluation measures employed were overall accuracy (i.e., percentage of total messages' speech acts correctly classified) and speech act class-level recall: percentage of total messages associated with a particular speech act that were correctly classified. Table 8 shows the experiment results for accuracy. LTAS's Labeled Tree kernel-based speech act classification component attained the best overall accuracy across all 10 data sets in the test bed, outperforming all comparison methods by at least 15% to 20%. Paired t-test results for accuracy were significant (all p-values < 0.001, n = 1,615). Appendix N includes the class-level recall values for the two best comparison methods (joint classification and collective classification) on four of the highly prominent speech acts: assertive, suggestion, question, and commissive. LTAS's Labeled Tree kernel outperformed both comparison methods for all speech acts across the 10 data sets. Moreover,

Table 8. Accuracies for Speech Act Classification Experiment

	Telco			Health			Security			Manu.
Method	**Forum**	**Social**	**Twitter**	**Forum**	**Social**	**Twitter**	**Forum**	**Social**	**Twitter**	**Chat**
LTAS – Labeled Tree*	**92.1**	**92.5**	**93.3**	**93.6**	**93.0**	**95.5**	**91.9**	**90.4**	**93.7**	**90.7**
Collective Classification	76.1	74.6	76.1	74.9	74.5	77.8	74.5	70.7	76.0	72.3
Joint Classification	72.4	69.7	75.3	72.0	72.4	75.5	71.9	70.5	74.2	68.4
CRF	61.1	66.7	67.9	64.0	70.2	73.8	61.8	66.3	69.0	64.2
n-gramSVM	64.1	67.9	68.3	64.4	66.1	66.8	65.6	68.4	67.6	64.8
n-Word Method	61.9	64.0	64.5	59.5	62.1	62.4	61.3	63.4	63.7	57.9

*Significantly outperformed comparison methods, with all p-values < 0.001

it performed fairly consistently across speech acts, with recall rates ranging from 86.5% to 98.8%. Labeled Tree's enhanced performance was attributable to the amalgamation of coherence tree structure and system, linguistic, and conversation attributes in a kernel-based method (see Appendix G). Interestingly, the joint classification and collective classification comparison methods, which also utilized coherence information, also performed markedly better than methods that relied primarily on message-level attributes (e.g., Cohen et al. 2004; Moldovan et al. 2011).

Experiment 4: Information Quality for Sense-Making

An experiment was conducted to evaluate the quality of information generated using LTAS as compared to existing methods (RQ4). Inaccurate coherence relations can distort representations of participants' roles in online group discussions. This has implications for social media use cases such as identification of key discussion participants (Zabin et al. 2011), as well as broader social network analysis using social media. Differences between actual and projected social network centrality measures can shed light on the level of distortion (Aumayr et al. 2011; Fu et al. 2008). Three commonly used measures are degree centrality, closeness centrality, and betweenness centrality. Degree centrality is the total number of out links (sent messages) and in links (received/reply-to messages) associated with a discussant; it is a measure of a discussant's level of participation and interaction within a discussion thread (Aumayr et al. 2011). Closeness centrality is a measure of the level of interaction between participants within a group, with greater interaction between discussants indicating greater closeness. Betweenness centrality is an important measure of how critical an individual is for the flow of communication among other discussants in a conversation (Fu et al. 2008). For a given discussant, it is computed as the proportion of shortest paths between discussants in the network that include the given discussant. We examined the mean absolute percentage error on degree, closeness, and betweenness centrality for the LTAS coherence analysis module and the comparison heuristic, linkage, and classification methods. The values were computed for each of the 10 data sets in our test bed. The results for closeness and betweenness appear in Appendix N.

Table 9 shows the experiment results for degree centrality. LTAS had the smallest mean absolute percentage errors across all data sets in the test bed, with error percentages of less than 7%. Error rates for LTAS were typically two to four times better than for those of comparison methods. Regarding RQ4, the differences were statistically significant (with all p-values < 0.001). With respect to the comparison methods, heuristic and classification each had error rates ranging from 10% to 25% for degree on most data sets. The linkage methods typically had mean absolute percentage errors in excess of 20%. Consistent with, and proportional to, the coherence analysis experiment results, centrality measure error rates were lowest on Twitter and social networking websites relative to web forums and group chat.

Figure 14 depicts the gold standard social network (top left chart), along with results generated by LTAS, heuristic, and linkage methods, for one of the discussions in the Telus (telecom) forum data set. In order to allow easier comparison, the node placements in all four charts are identical, node sizes are proportional to degree centrality, and reply-to links/ties obviously vary for the different ICA methods. Looking at the four charts, it is apparent that LTAS most closely resembles the gold standard in terms of links between nodes and node sizes. Conversely, the linkage method (bottom right) tends to exaggerate the degree centrality of many nodes (e.g., Wonton Noodle, beachside, BadMagpie, zod5000, etc.). This is consistent with prior studies, which have also observed that linkage methods inflate degree centrality (by over-attributing in-degree) for discussants with greater posting frequency (Fu

Table 9. Mean Absolute Percentage Error for Degree Centrality Measure

	Telco			Health			Security			Manu.
Method	**Forum**	**Social**	**Twitter**	**Forum**	**Social**	**Twitter**	**Forum**	**Social**	**Twitter**	**Chat**
LTAS*	**4.9**	**4.3**	**2.6**	**6.1**	**6.2**	**3.3**	**4.7**	**4.3**	**2.1**	**7.9**
Heuristic	15.2	14.0	13.7	17.2	17.1	10.3	15.2	13.7	8.9	16.9
Classification	18.3	15.9	14.9	18.0	16.5	8.7	15.9	12.5	8.0	17.1
Linkage-Previous	25.2	29.9	23.9	27.8	26.2	16.9	26.6	19.6	14.7	41.3
Linkage-First	37.0	34.8	35.8	37.9	35.6	23.7	42.2	30.2	26.1	55.7

*Significantly outperformed comparison methods, with all p-values < 0.001

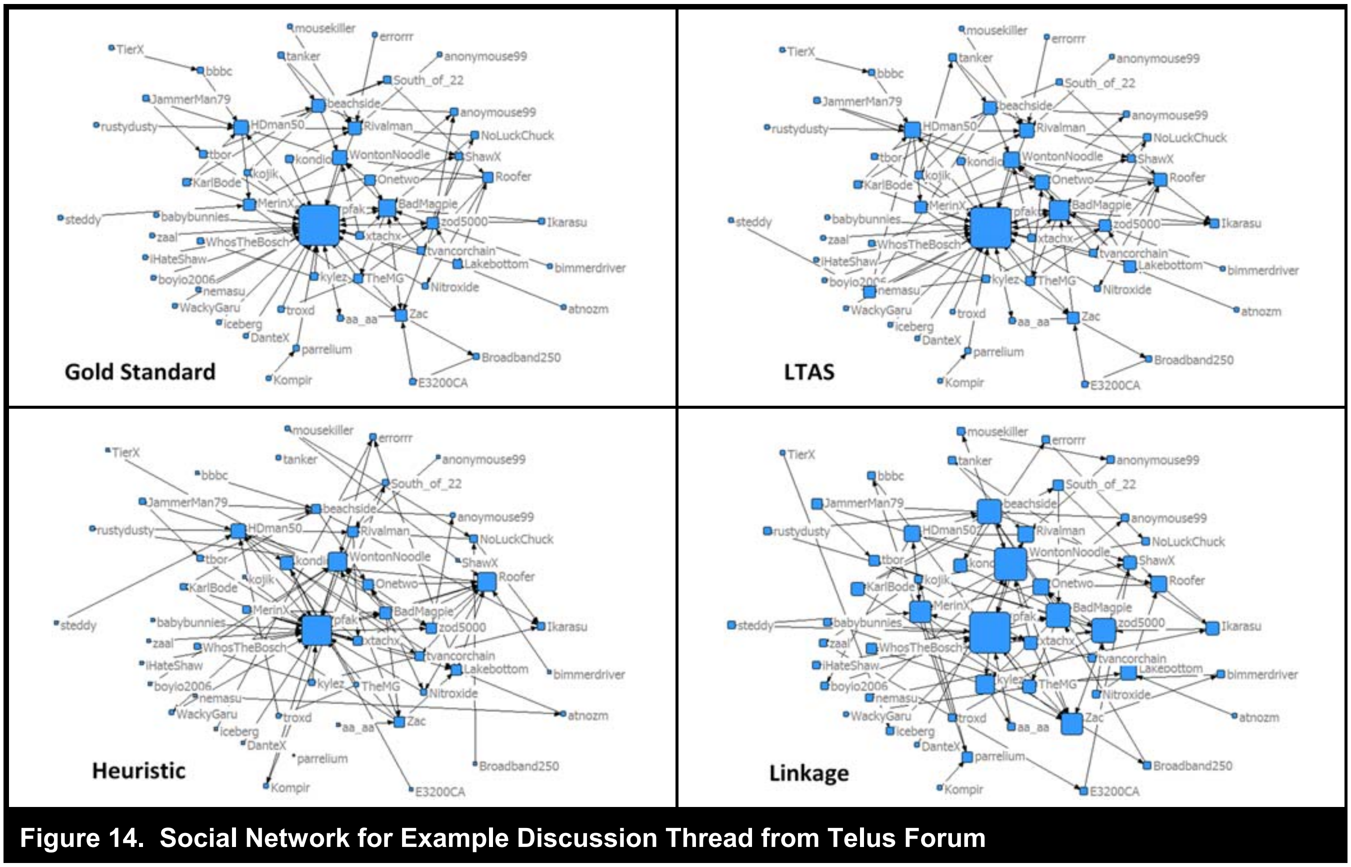


Figure 14. Social Network for Example Discussion Thread from Telus Forum

et al. 2008). Similarly, the heuristic method exaggerated degree centrality for some nodes while understating it for others (bottom left of Figure 14). The figure visibly illustrates how lower coherence analysis performance can significantly hurt the quality of a social media thread discussion's network. When applied across entire forums and social media channels, these effects become even more pronounced (as shown earlier in Figure 5). Overall, the results from the experiment suggest that LTAS is less likely to inflate or underestimate the perceived importance of discussion participants (in terms of centrality). Given that over 75% of organizations surveyed consider identification of influential participants as one of the most important use cases for social media analytics (Zabin et al. 2011), the results further demonstrate the usefulness of the LTAS system.

Experiment 5: User Sense-Making

The prior experiments demonstrated information quality enhancements, an important prerequisite for user sense-making (Weick et al. 2005). Ultimately, for these enhancements to be

Table 10. Overview of Participants in User-Sense-Making Experiment

Dimension	Telecom	Health	Security	Manufacturing
Number of Participants	120	103	85	132
Organization	TelcoInc	HealthInc	SecurityInc	Three companies and university
% Female	37%	31%	35%	43%
Bachelor's Degree	96%	97%	98%	99%
Master's Degree	41%	64%	59%	67%

meaningful, users must be able to derive knowledge and insights. Accordingly, we evaluated the effectiveness of SATrees generated by LTAS in assisting users with sense-making (RQ5) in comparison with three additional experiment settings: (1) A conversation tree comprised of **gold standard** coherence relations and human expert tagged speech acts; (2) a conversation tree comprised of **best benchmark** methods for coherence analysis (classification) and speech act classification (joint classification); and (3) **sequential order**, chronologically ordered discussion messages without coherence relation information or speech act tags. The methodology used was a controlled experiment; participants were assigned to one of the four experiment settings and asked to answer sense-making questions.

The experiments were performed in the four industry contexts previously described in the evaluation section: telecommunications, health, security, and manufacturing. Table 10 summarizes the experiment participants. For the telecom, health, and security contexts, the participants were practitioners in three large North American telecommunications, health, and security companies, respectively. These practitioners included members of social media monitoring teams, customer relationship management team members, marketing analysts, marketing managers, product design team members, etc. For the manufacturing data set, participants were recruited by email invitations to employees at three companies, graduate students, and faculty members from the school of management at a major university.

User Experiment Design

We selected two representative discussion threads from our test bed for each of the four industry contexts depicted in Table 10. The threads were presented to the participants using the aforementioned presentation formats to which they were assigned, through a web-based interface. Four sense-making questions were used in the experiment. The questions were closely aligned with some of the major social media use cases alluded to in the introduction, namely identifying issues and ideas. The questions were tailored to each industry context, but entailed similar sense-making tasks and cognitive effort (Klein et al. 2006). Appendix I provides details about the questions and thread topics used for each industry context.

Here we describe the four questions for the tea manufacturing context. The first was a *general* sense-making question: users were asked to list all the solutions proposed in the discussion. Following Heracleous and Marshak's (2004) work pertaining to analyzing discourse, we employed three additional sense-making questions associated with *action*, *situated action*, and *symbolic action* as they involve differing levels of data fusion (Klein et al. 2006). In the first of these three questions (*action*), we asked which solutions a particular discussant supported. The second (*situated action*) question asked the participants to identify the solution that resulted in the greatest amount of conflict among discussants in the entire discussion thread (i.e., one creating the largest dichotomy between support and opposition). The third (*symbolic action*) question asked participants to sense certain discussants' characteristics based on their utterances and interactions in the discussion (e.g., level of enthusiasm toward others' ideas).

Participants were required to structure their answers as bulleted lists. Responses were evaluated using theme identification, an approach that has been used to evaluate user performance in complex information retrieval tasks when a correct answer contains multiple themes (Zhou et al. 2006). A theme was considered correct if it matched any of the themes identified by experts; evaluators were used to determine what constituted a match. By examining the themes that participants derived using different representation tools, we were able to evaluate how effectively each experimental setting aided subjects with sense-making.

The experiment protocol was pretested with 2 doctoral students and a pilot study was conducted with a total of 12 doctoral and master's students. Based on their feedback, we clarified the wording in questions and refined the experiment process and instructions. Each participant was randomly assigned to one of the four experimental settings. All partici-

Table 11. Results Across All Eight Sense-Making Questions for User Experiment

	Telecom			Health		
Technique	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Gold Standard	80.4	74.1+	77.1+	79.0	74.1	76.4
SATree	**77.8***	**72.6***	**75.1***	**75.5***	**71.0***	**73.2***
Best Benchmark	63.3	59.9	61.5	61.5	56.4	63.9
Sequential Order	58.7	53.4	55.9	54.0	47.4	50.2
	Security			**Manufacturing**		
Technique	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
Gold Standard	84.8+	80.0+	82.3+	67.8+	57.5+	60.9+
SATree	**84.7***	**80.5***	**82.5***	**66.5***	**55.7***	**58.4***
Best Benchmark	70.0	72.0	71.0	45.8	36.2	38.8
Sequential Order	61.1	64.7	62.7	48.0	35.6	39.2

*Significantly outperformed Best Benchmark and Sequential Order methods, with all p-values < 0.001

+Did not significantly outperform SATree

pants answered all four questions for both discussion threads, resulting in eight total questions and answers per participant. The order in which the two threads were presented was randomized to avoid biases. For each thread, participants had 5 minutes to familiarize themselves with the discussion's messages before they started answering the questions. During the experiment, the tasks performed by participants were timed. All answers were cross-judged by two domain experts. In order to measure participant's sense-making capabilities, theme precision, recall, and f-measure were calculated (Pirolli and Card 2005). Participants who failed to answer one or more of the eight total questions or those that failed to follow instructions were removed from the data. In each of the four contexts, the number removed was less than 4% (i.e., two from telecom, four from health, three from security, and five from manufacturing).

User Experiment Results

Table 11 depicts the average theme precision, recall and f-measure across all questions for the four experiment settings, on the four industry contexts. As expected, subjects using the Gold Standard conversation tree attained the best overall results. Interestingly, however, this gain was not significantly better than the performance for subjects that used SATree on three of the four data sets: telecom, security, and manufacturing. This result suggests that in many cases SATree may provide somewhat comparable support for sense-making as compared to gold standard coherence relations and speech act composition information. Furthermore, SATree yielded significantly better performance than the best benchmark and sequential ordering for all four contexts (all pair-wise t-test p-values < 0.001). Participants leveraging SATree attained precision and recall that were 20 percentage points higher than status quo sequential ordering, and more than 10 percentage points better than the best benchmark. These results demonstrate the transference of the proposed LAP-based systems' improved information quality representations into augmented user sense-making performance. Two critical criteria for analytical technologies that support sense-making are information quality and time (Pirolli and Card 2005). Although not reported here, the three conversation tree-based representations (gold standard, SATree, and best benchmark) also had significantly lower participant response times than the sequential ordering method on the telecom, health, and security settings. In other words, those using SATrees were not only markedly more accurate, they were also faster than participants using the sequential ordering method.

Table 12 shows the f-measure results for the four questions across the two discussion threads for all four industry contexts. Consistent with the overall results, SATree significantly outperformed best benchmark and sequential order for all questions, suggesting that it is better suited to support sense-making for the issue/idea identification and participant analysis use cases. Participants using the gold standard did not perform significantly better than those using SATree on 7 of the 16 questions, further underscoring the relative lack of information degradation when using the LAP-based system. Overall, the results presented in Tables 11 and 12 lend credence to the notion that text analytics systems guided by LAP-based principles may facilitate enhanced sense-making in online discourse.

Table 12. Results by Question Type in User Experiment

	Telecom				Health			
Technique	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Gold Standard	75.3	82.4+	77.7	72.8	76.2	80.3	77.5	71.7
SATree	**73.0***	**81.5***	**75.3***	**70.6***	**71.9***	**77.3***	**73.7***	**69.8***
Best Benchmark	59.6	65.7	61.1	59.8	62.6	65.7	65.4	62.0
Sequential Order	54.8	60.6	56.4	51.9	50.7	51.6	46.4	52.3
	Security				**Manufacturing**			
Technique	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Gold Standard	83.4+	85.0+	82.1+	78.7+	69.0	46.6+	82.8	55.3+
SATree	**84.5***	**84.8***	**82.0***	**78.7***	**60.8**	**48.4***	**77.1***	**55.8***
Best Benchmark	72.0	74.1	71.5	66.2	48.5	30.9	50.4	34.7
Sequential Order	63.2	63.6	64.4	59.8	51.9	33.6	53.3	32.0

*Significantly outperformed Best Benchmark and Sequential Order methods, with all p-values < 0.001

+Did not significantly outperform SATree

Field Experiment

For novel IT artifacts, field experiments are useful for demonstrating value in organizational settings. Accordingly, for RQ6 we conducted a 4-month field experiment at TelCorp to show the utility of the information provided by the proposed LAP-based system (LTAS). The experiment was performed using members of TelCorp's large social media monitoring team, encompassing 23 analysts. This team previously used a customized version of a popular social media analysis tool provided by a major vendor in the space. The tool presented tables and charts, searching, and browsing features at different levels of granularity: social media channels, discussants, messages, and threads. The browsing capability presented threads using existing channel-system features (i.e., they appeared as they would in the actual forum, social networking chat, and/or microblog). Analytics features included topic (keyword) and sentiment analysis, which could be used as filters/dimensions in the existing search, browsing, and visualization capabilities. TelCorp's engineering team had developed custom dashboards on top of the tool to support their internal reporting needs pertaining to various use cases, including issues, ideas, and key participants.

A/B testing is a commonly used method to concurrently examine the performance of alternative artifacts or design settings. The key outputs of LTAS are conversation affiliations, coherence relations, and message speech acts. Treating the existing system used by TelCorp as setting A, we worked with the TelCorp's IT staff to develop setting B. In order to test our premise that the pragmatic view can enrich analytical capabilities over the pervasive semantic perspective, this setting entailed inclusion of coherence relation, conversation, and speech act information on top of the existing system already supporting topics and sentiments. For the B system setting, LTAS was embedded into TelCorp's real-time analysis pipeline adding conversation affiliation, reply-to relation, and speech act labels to all messages. Furthermore, participant importance rankings were computed using these revised social network analysis metrics. In the custom dashboards, sequential ordering was complemented with an SATree option. Conversation and speech acts were added as additional filters/dimensions for search, browsing, and visualization.

Members of TelCorp's monitoring team were randomly assigned to one of the two settings. One team member left the company during the 4-month experiment, resulting in 12 employees being assigned to A and 10 being assigned to B. Each team member had access only to their respective system setting for the duration of the experiment; they were asked to perform all daily monitoring tasks using this system. Using prior research as guidelines, a longitudinal data collection schedule was used (Venkatesh et al. 2003). Surveys were utilized to capture all users' perceptions about system A, one week of training on B for those assigned, followed by the use of surveys to capture user reactions for A and B at periodic intervals. After the one week period, user reactions were gathered again at the two month and four month marks, along with system usage data (Venkatesh et al. 2003). The user reaction constructs, which were adapted from Venkatesh et al. (2003), included perceived usefulness of the system, perceived usefulness of the information provided by the system, perceived ease of use of the system, perceived usefulness of the thread browsing capability, and perceived usefulness of the participant ranking capability. These were measured on a

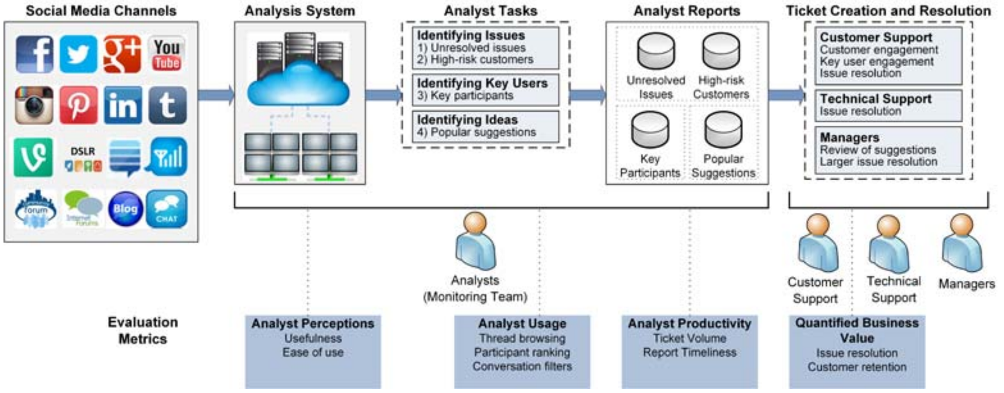


Figure 15. High-Level Overview of TelCorp's Business Process for Social Media Monitoring

1–10 continuous scale (see Appendix J for further details). The system usage measurements were captured through system logs and transformed to a 1–10 scale using a simple range transformation. The system automatically logged off inactive users after 10 minutes to reduce idle time in usage logs.

Figure 15 provides an overview of TelCorp's social media monitoring team workflow. Further details appear in Appendix M. TelCorp's monitoring team focuses on three key social media monitoring tasks: identifying issues, identifying key users, and identifying suggestions. Identifying issues encompasses (1) unresolved issues and (2) high-risk customers. TelCorp defines unresolved issues as events that adversely impact a set of customers. A good, extreme example is the one presented the second section of this article on the need for sense-making. Two other examples that arose during the 4-month field experiment include an error in the billing system which caused customers in three U.S. states to receive excess charges on their monthly statements, and a technical issue with the installation software of a new, integrated router-plus-modem which caused tens of thousands of customers to experience random Internet outages. High-risk customers are customers that may possibly churn due to what TelCorp considers "standard operational issues." Examples include an individual upset about call center wait times, or a customer considering switching to another carrier due to price differences. While issue identification is the primary use case for TelCorp's monitoring team, they also look to identify key discussion participants based on social network centrality—these include key positive/negative influencers, brand advocates, etc. Additionally, analysts in the monitoring team seek to identify popular suggestions. Examples include ideas about fund-raising events, charities valued by existing and prospective customers, requests for new product and/or service offerings, and suggestions on how to enhance the customer web portal and mobile app.

For the field experiment, four types of evaluation metrics were incorporated. The first two were analyst perceptions and actual system usage (measured through the process described in the prior paragraph). The other two were analyst productivity and quantified business value. The first two sections in Table 13 shows mean values for survey responses and actual usage, at the four-month mark. Users of system B responded much higher for perceived usefulness of the system, its information for identifying issues, thread browsing capability, as well as actual usage of thread browsing, participant ranking, and thread/conversation-level analysis. The increased perceived usefulness and actual usage of the thread browsing capability is attributable to the SATree-based browsing feature in system B. The participant ranking capability based on LTAS coherence relations also garnered higher perceived usefulness and actual usage. Various characteristics, including speech act composition, contributed to higher perceived usefulness of information for identifying issues. Furthermore, the use of conversations in B was higher than the use of threads in A (even though thread capability was also available in B).

Ultimately tangible value results from observed increases in productivity leading to quantifiable business value. Using the system, analysts submit reports, with each report including a description, severity level, and associated social media discussants, conversations, and/or threads. These reports are routed to customer support representatives, technical support, and/or managers. For a subset of reports, tickets are created indicating cases requiring action. Customer support reps attempt to engage with high-risk customers with the goal of reducing attrition. They also reach out to key users in order to preemptively garner brand advocacy or mitigate negative influence. Tech support reps work to resolve technical issues. Managers review suggestions and may also be involved in resolution of larger issues. Since Systems A and B were run

Table 13. Results of Field Experiment at TelCorp

Dimension		System A Status Quo N = 12	System B with LTAS N = 10
Analyst Perceptions	Usefulness of system (1–10)	7.9	**8.7**
	Ease of system use (1–10)	**8.1**	7.8
	Usefulness of information for identifying issues (1–10)	7.6	**8.5**
	Usefulness of thread browsing capability (1–10)	6.0	**7.2**
	Usefulness of participant ranking capability (1–10)	7.9	**8.2**
System Usage	Usage of thread browsing capability (1–10)+	7.1	**8.0**
	Usage of participant ranking capability (1–10)	8.2	**8.6**
	Usage of thread/conversation filters and charts (1–10)*	7.9	**8.8**
Analyst Productivity	Mean timeliness of reports (in minutes)	84.3	**30.7**
	Ticket volume—unresolved issues: total Ticket volume—unresolved issues: non-overlapping	19,040 1,548	**28,263** **10,771**
	Ticket volume—high-risk customers: total Ticket volume—high-risk customers: non-overlapping	9,520 1,415	**15,073** **6,968**
	Ticket volume—suggestions: total unique Ticket volume—suggestions: unique non-overlapping	452 54	**1,153** **755**
	Ticket volume—key participants: total Ticket volume—key participants: non-overlapping	492 134	**640** **302**
Quantified Business Value	Issue resolution	$9,139,200	**$13,566,000**
	Customer retention	$4,569,600	**$7,235,200**

*Measured thread-level usage for A versus conversation-level for B

+System B users also significantly higher for web forums, social networking sites, and microblogs

in parallel using non-overlapping teams, reports generated by users of each system were tracked, resulting in two sets of reports. The first of the two productivity measures incorporated by TelCorp was *timeliness* of overlapping reports created by users of both systems: in other words, the timeliness delta between report submission timestamps. The second productivity measure was *ticket volume*. Only those reports deemed to be the most important are converted to tickets by the customer/technical support reps or managers. For TelCorp, the total number of generated tickets, as well as non-overlapping tickets attributable to reports submitted by users of System A versus System B signified important productivity measures. Business value stems from *better* identifying issues, key participants, and ideas in a *timelier* manner. Appendix M offers further details. For the field experiment, TelCorp chose to quantify business value primarily in terms of identified issues, including the value of resolving issues on customer churn reduction (i.e., for those impacted by the issue), and successfully engaging and retaining high-risk customers. Hence we report business value metrics related to these use cases.

Looking at the productivity metric rows in Table 13, it is apparent that analysts using System B were able to generate reports resulting in a much larger number of total tickets for unresolved issues and high-risk customers. Furthermore, looking at the unique ticket volumes, users of System A produced fairly few tickets that were not covered in the set generated by users of System B. Based on customer/technical support rep and manager follow-up, the quantified value of these tickets to TelCorp in terms of post-issue customer retention or standard churn avoidance was over $7 million during the 4-month field experiment. Similarly, System B garnered higher ticket volumes for suggestions—more than double those attributable to users of System A (with few unique tickets in System A). Additionally, System B also resulted in greater tickets for key participants. The findings highlight the potential utility of information generated by the proposed LAP-based system in an organizational setting. In fact, TelCorp was so pleased with the field experiment results that, moving forward, they have adopted System B as their full-time analysis tool for the entire monitoring team. Overall, the analyst perceptions, system usage, productivity results, and quantified business value over an extended period of time further bolster external validity (Russell et al. 1993).

Results Discussion

Following Walls et al. (1992), we used a kernel theory to govern requirements and design, each of which was carefully tested. Each phase of the LAP-based framework is intended

Table 14. Summary of Results for Research Questions

Eval. Part	RQ	Result
(1) LAP-Based Methods	1	Conversation disentanglement methods explicitly incorporating detection of conversation beginnings (primitives) able to significantly outperform state-of-the art techniques.
	2	Coherence analysis methods incorporating conversation structure information in conjunction with system and linguistic cues able to markedly outperform existing methods, which are devoid of conversation structure information.
	3	Speech act classification methods leveraging conversation trees and kernel-based methods able to markedly boost classification capabilities.
(2) Sense-making	4	Improved coherence analysis can significantly enhance social network analysis centrality measures over existing methods that primarily rely on system-generated features.
	5	Sense-making user experiments in multiple organizations, with several hundred practitioners, revealed significantly higher precision and recall for sense-making tasks, relative to benchmark methods.
(3) Business Value	6	Four-month field experiment at TelCorp revealed that social media team members' perceptions, usage, and productivity were higher when using a system with LAP-based information relative members relying on existing social media analytics systems, resulting in significant quantified business value.

to improve sense-making while simultaneously serving as an input refinement mechanism for other phases of the framework. The conversation disentanglement component produces the conversation structure attributes used as part of the input feature set for the coherence analysis component. Results from the conversation disentanglement and coherence analysis components are used to enhance speech act classification. The coherence relations and message-speech act information is used to create SATrees. Consistent with design science principles (Hevner et al. 2004), we used a series of experiments to rigorously test each component of the proposed IT artifacts. The experiment results, summarized in Table 14, demonstrate the efficacy of LTAS and its underlying LAP-based framework.

Regarding the first part of our evaluation, experiments 1 through 3 demonstrated the effectiveness of the conversation disentanglement, coherence analysis, and speech act classification components of LTAS relative to benchmark methods (RQs 1–3). In the second part of the evaluation, experiment 4 showed how the LTAS components collectively resulted in augmented information quality in the context of social networks (RQ4). Based on experiment 5 (RQ5), LTAS facilitated demonstratively better sense-making than comparison methods, allowing users to better understand discussion elements pertaining to social media use cases. Experiment 6 (RQ6) presented results from a 4-month field experiment at TelCorp where the use of LTAS-based information enhanced social media monitoring team members' perceptions, system usage, and productivity, resulting in considerable quantified business value.

The findings have important design implications for text/social analytics artifacts, which is a growing body of literature in IS (e.g., Abbasi and Chen 2008; Chau and Xu 2012; Lau et al. 2012). The results also provide insights for the broader social media analysis researcher and practitioner communities. Key takeaways include:

- *Consider making conversations more of a focal point*—the interplay between conversations, coherence relations, and speech act composition of messages in social media is valuable for enhanced sense-making. For instance, conversation structure, including conversation beginnings, message conversation affiliation information, and conversation trees received limited attention in prior work despite their ability to dramatically enhance coherence analysis and speech act identification. Conversations may serve as a more meaningful unit of analysis than system-generated aggregate discussion threads, or stand-alone messages devoid of communication context.
- *Proceed with caution when performing social network analysis using system-generated reply-to relations*—social networks constructed purely based on system features and naïve linkage methods in web forums, social

networking sites, and microblogs can distort important centrality measures such as degree and betweenness for key network members by 15% to 50%. Enhanced coherence analysis methods are essential for ensuring information quality in social media-based networks.

- *Incorporating the pragmatic view in monitoring systems can enhance sense-making*—the semantic view of language is pervasive in text/social media analytics, and for good reason. Topic, sentiment, and affect analysis are incredibly important and valuable analysis dimensions (Abbasi and Chen 2008). However, also incorporating the pragmatic view in text/social analytics systems (e.g., conversation structure, coherence relations, and speech act information) can significantly improve users' social media sense-making capabilities. We observed increases of 20 to 40 percentage points for various tasks in four organizations, with hundreds of practitioners. Based on field experiment results, these findings also translated into enhanced analyst perceptions, usage, and productivity, resulting in meaningful quantified business value.

- *Developing and/or utilizing advanced machine learning and data science-based IS design artifacts can further the state-of-the-art for text/social media analytics*—our proposed LTAS artifact demonstrated the utility of advanced kernel-based methods, including tree and ensemble kernel-based approaches. As data science continues to play a bigger role in IS research geared toward deriving economic and societal value from unstructured Big Data (Abbasi et al. 2016; Saar-Tsechansky 2015), exploration of advanced machine learning-based constructs, methods, and instantiations seems advantageous.

Conclusions

Our contributions are three-fold. First, we presented several key findings relevant to the design of text analytics artifacts and to the social media analysis research and practitioner communities (summarized in the previous section). Additionally, our two design science contributions are as follows. Second, we described how a framework based on LAP principles can be used to inform the design of text analytics systems for enhanced sense-making. Third, we developed LTAS, which adopted these principles in its feature sets and techniques for conversation disentanglement, coherence analysis, and speech act classification. LTAS employed several important concepts that have been incorporated into prior LAP-based studies, including context, relevance, thematization, discourse ambiguity, conversation structure elements, and message and conversation-level speech act composition. In order to effectively incorporate structural, linguistic, and interaction information, novel kernel-based classifiers were developed. A series of experiments were used to illustrate the efficacy of various components of LTAS. User studies and a field experiment demonstrated the external validity of the proposed design artifacts. With respect to recent design science guidelines, our research contribution represents an "improvement": a novel and holistic solution to an established, important problem (Goes 2014; Gregor and Hevner 2013).

Analytical technologies that support enhanced sense-making from online discourse constitute an increasingly critical endeavor as comprehension lays the foundation for reasoning and decision-making (Weick et al. 1995). The results of our work have important implications for social media analytics. As intra-organizational and external-facing communication via social media becomes increasingly pervasive (Bughin and Chui 2010), sense-making remains a paramount concern (Honeycutt and Herring 2009). The results can shed light on interaction dynamics in intra-organizational communication, corporate blogs and wikis, and group support systems. Furthermore, organizations are increasingly interested in understanding customer actions and intentions expressed via social media; that is, going beyond the *what* to uncover contextual elements such as the *why* and *how* (Mann 2013). Some specific, important use-cases for social media analytics are identifying issues and important participants (Zabin et al. 2011). While topic and sentiment analysis remain essential semantic forms of analyses, as shown in the TelCorp and other examples, the pragmatic view emphasized by LAP provides considerable complementary value to allow better understanding of issues through examination of interactions and speech acts within conversations. Furthermore, enhanced coherence analysis enables meaningful representation of social media social networks, making identification of key discussion participants more feasible.

Future work can extend this study in various ways. LAP-based text analytics systems for sense-making could be evaluated in other contexts, on other discussion topics, languages, and communication modes. LTAS could be improved via adaptive learning where components iteratively improve one another, or via automated detection of conversation types. Additionally, the SATrees in LTAS signify the key outputs of systems using the LAP-based framework. As done in our field experiment, these outputs can be leveraged with alternative visual formats, or for other social media use cases as an information/feature space refinement, such as social media for predicting adverse events, financial metrics, health-related outcomes, etc. Nevertheless, the system and underlying framework presented demonstrate the viability of applying LAP concepts, which advocate the pragmatic perspective centered around conversations and actions as complementary to the pervasive semantic view, enabling enhanced text analytics for sense-making. Given the ubiquitous nature of online discourse, the results of our work constitute an important and timely endeavor; one which future research can build upon.

Acknowledgments

We would like to thank our industry partners in the telecommunications, health, and security industries for their invaluable expertise, financial support, and assistance with various facets of the paper, including research motivation examples, test bed construction, evaluation, user study, and field experiment. This work was funded by NSFC (91646205, 71421002, 71601124, 71171131), and NSF IIS-1553109, IIS-1236970, BDS-1636933, CCF-1629450, ACI-1443019.

References

Aakhus, M. 2007. "Communication as Design," *Communication Monographs* (74:1), pp. 112-117.

Abbasi, A., and Chen, H. 2008. "CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811-837.

Abbasi, A., Sarker, S., and Chiang, R. H. L. 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the AIS* (17:2), pp. i-xxxii.

Abbasi, A., Zhang Z., Zimbra, D., and Chen, H. 2010. "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly* (34:3), pp. 435-461.

ACSI. 2014. "ACSI Telecommunications and Information Report 2014," The American Customer Satisfaction Index, University of Michigan.

Adams, P. H., and Martell, C. H. 2008. "Topic Detection and Extraction in Chat," in *Proceedings of the IEEE International Conference on Semantic Computing*, pp. 581- 588.

Adjeroh, D., Beal, R., Abbasi, A., Zheng, W., Abate, M., and Ross, A. 2014. "Signal Fusion for Social Media Analysis of Adverse Drug Events," *IEEE Intelligent Systems* (29:2), pp. 74-80.

Anwar, T., and Abulaish, M. 2012. "Mining an Enriched Social Graph to Model Cross-Thread Community Interactions," in *Proceedings of the 3rd International Workshop on Mining Social Media*, Milwaukee, WI, pp. 35-38.

Aumayr, E., Chan, J., and Hayes, C. 2011. "Reconstruction of Threaded Conversations in Online Discussion Forums," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pp. 26-33.

Auramaki, E., Hirschheim, R., and Lyytinen, K. 1992. "Modelling Offices Through Discourse Analysis: The SAMPO Approach," *The Computer Journal* (35:4), pp. 342-352.

Auramaki, E., Lehtinen, E., and Lyytinen, K. 1988. "A Speech-Act Based Office Modelling Approach," *ACM Transactions on Office Information Systems* (6:2), pp. 126-152.

Berfield, S. 2013. "OUR Walmart Agrees to Stop Picketing for 60 Days," *Bloomberg Businessweek*, February 1.

Bughin, J., and Chui, M. 2010. "The Rise of the Networked Enterprise: Web 2.0 Finds its Payday," *McKinsey Quarterly*, December.

Carvalho, V. R., and Cohen, W. W. 2005. "On the Collective Classification of Email 'Speech Acts,'" in *Proceedings of the 28th Annual ACM SIGIR Conference*, Salvador, Brazil, pp. 345-352

Chau, M., and Xu, J. 2012. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly* (36:4), pp. 1189-1216.

Choi, F. Y. Y. 2000. "Advances in Domain Independent Linear Text Segmentation," in *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics*, San Francisco, pp. 26-33.

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. 2004. "Learning to Classify Email into 'Speech Acts,'" in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, pp. 309-316.

Collins, M., and Duffy, N. 2002. "Convolution Kernels for Natural Language," in *Advances in Neural Information Processing Systems*, T. G. Diettrich, S. Becker, and Z. Ghahramani (eds.), Cambridge, MA: MIT Press, pp. 625-632.

Comer, D., and Peterson, L. 1986. "Conversation-Based Mail," *ACM Transactions on Computer Systems* (4:4), pp. 200-319.

de Moor, A., and Aakhus, M. 2006. "Argumentation Support: From Technologies to Tools," *Communications of the ACM* (49:3), pp. 93-98.

Donath, J. 2002. "A Semantic Approach to Visualizing Online Conversations," *Communications of the ACM* (45:4), pp. 45-49.

Elsner, M., and Charniak, E. 2010. "Disentangling Chat," *Computational Linguistics* (36:3), pp. 389-409.

Esuli, A., and Sebastiani, F. 2006. "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining," in *Proceedings of the 5th Conference on Language Resources and Evaluation*, pp. 417-422.

Fang, X. 2013. "Inference-Based Naive Bayes: Turning Naive Bayes Cost-Sensitive," *IEEE Transactions on Knowledge and Data Engineering* (25:10), pp. 2302-2313.

Fu, T., Abbasi, A., and Chen, H. 2008. "A Hybrid Approach to Web Forum Interactional Coherence Analysis," *Journal of the American Society for Information Science and Technology* (59:8), pp. 1195-1209

Goes, P. 2014. "Editor's Comments: Design Science Research in Top IS Journals," *MIS Quarterly* (38:1), pp. iii-viii.

Gregor, S., and Hevner, A. R. 2013. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2), pp. 337-355.

Halladay, J. 2010. "Gap Scraps Logo Redesign after Protest on Facebook and Twitter," *The Guardian*, Marketing & PR, October 12.

Halper, F., Kaufman, M., and Kirsh, D. 2013. "Text Analytics: The Hurwitz Victory Index Report," Hurwitz and Associates, Needham Heights, MA.

Heracleous, L., and Marshak, J. R. 2004. "Conceptualizing Organizational Discourse as Situated Symbolic Action," *Human Relations* (57:10), pp. 1285-1312.

Herring, S. C. 1999. "Interactional Coherence in CMC," *Journal of CMC* (4:4).

Herring, S. C., and Nix, C. 1997. "Is 'Serious Chat' an Oxymoron? Academic vs. Social Uses of Internet Relay Chat," paper presented at the American Association of Applied Linguistics, Orlando, FL.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.

Honeycutt, C., and Herring, S. C. 2009. "Beyond Microblogging: Conversation and Collaboration in Twitter," in *Proceedings of the 42nd Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society Press.

Jackson, S. 1998. "Disputation by Design," *Argumentation* (12), pp. 183-198.

Janson, M. A., and Woo, C. C. 1996. "A Speech Act Lexicon: An Alternative Use of Speech Act Theory in Information Systems," *Information Systems Journal* (6:4), pp. 301-329.

Joachims, T. 1999. "Making Large-Scale SVM Learning Practical," In *Advances in Kernel Methods: Support Vector Learning*, B. Scholkopf, C. Burges, and A. Smola (eds.), Cambridge, MA: MIT Press, pp. 169-184.

Khan, F. M., Fisher, T. A., Shuler, L., Wu, T., and Pottenger, W. M. 2002. "Mining Chat-Room Conversations for Social and Semantic Interactions," Technical Report LU-CSE-02-011, Lehigh University, Bethlehem, PA

Kim, J., Li, J., and Kim, T. 2010. "Towards Identifying Unresolved Discussions in Student Online Forums," in *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Los Angeles, CA, pp. 84-91.

Kim, S. N., and Baldwin, T. 2005. "Automatic Interpretation of Noun Compounds Using WordNet Similarity," in *Natural Language Processing – IJCNLP 2005*, R. Dale, K. F. Wong, J. Su, and O. Y. Kwong (eds.), Berlin: Springer, pp. 945-956.

Kim, S. N., Wang, L., and Baldwin, T. 2010. "Tagging and Linking Web Forum Posts," in *Proceedings of the 14th Conference on Computational Natural Language Learning*, Uppsala, Sweden, pp. 192-202.

Klein, G., Moon, B. M., and Hoffman, R. R. 2006. "Making Sense of Sensemaking 1: Alternative Perspectives," *IEEE Intelligent Systems* (21:4), pp. 70-73.

Kobielus, J. 2011. "Telcos Tune Customer Experiences with Behavior Analytics," Forrester Research, June 30.

Kuechler, W. L. 2007. "Business Applications of Unstructured Text," *Communications of the ACM* (50:10), pp. 86-93.

Kuo, F. Y., and Yin, C. P. 2011. "A Linguistic Analysis of Group Support Systems Interactions for Uncovering Social Realities of Organizations," *ACM Transactions on MIS* (2:1), Article 3.

Lau, R., Liao, S., Wong, K. F., and Dickson, K. 2012. "Web 2.0 Environmental Scanning and Adaptive Decision Support for Business Mergers and Acquisitions," *MIS Quarterly* (36:4), pp. 1239-1268.

Lee, K. K. 2013. "Maker's Mark Apologizes for Almost Diluting its Bourbon," *Forbes*, February 17.

Lyytinen, K. 1985. "Implications of Theories of Language for IS," *MIS Quarterly* (9:1), pp. 61-74.

Mann, J. 2011. "Hype Cycle for Business Use of Social Technologies," Gartner Research, August 25.

Mann, J. 2013. "Hype Cycle for Social Software," Gartner Research, July 16.

McDaniel, S., Olson, G., and Magee, J. 1996. "Identifying and Analyzing Multiple Threads in Computer-Mediated and Face-to-Face Conversations," in *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work*, Cambridge, pp. 39-47.

Miller, G. A. 1995. "WordNet: A Lexical Database for English," *Communications of the ACM* (38:11), pp. 39-41.

Moldovan, C., Rus, V., and Graesser, A. R. 2011. "Automated Speech Act Classification for Online Chat," in *Proceedings of the 22nd Midwest AI and Cognitive Science Conference*, Cincinnati, Ohio.

Nash, C. M. 2005. "Cohesion and Reference in English Chatroom Discourse,"in *Proceedings of the 38th Hawaii International Conference on System Science*, Los Alamitos, CA: IEEE Computer Society Press.

Pirolli, P., and Card, S. 2005. "The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis," in *Proceedings of the of International Conference on Intelligence Analysis*, McLean, VA.

Raghu, T. S., Ramesh, R., Chang, A. M., and Whinston, A. B. 2001. "Collaborative Decision Making: a Connectionist Paradigm for Dialectical Support," *Information Systems Research* (12:4), pp. 363-383.

Rowe, M., Angeletou, S., and Alani, H. 2011. "Anticipating Discussion Activity on Community Forums," in *Proceedings of the Third IEEE International Conference on Social Computing,* pp. 315-322

Russell, D. M., Stefik, M. J., Pirolli, P., and Card, S. K. 1993. "The Cost Structure of Sensemaking," in *Proceedings of the ACM Conference on Computer–Human Interaction*, pp. 269-276.

Sack, W. 2000. "Conversation Map: An Interface for Very Large-scale Conversations," *Journal of Management Information Systems* (17:3), pp. 73-92.

Saar-Tsechansky, M. 2015. "Editor's Comments: The Business of Business Data Science in IS Journals," *MIS Quarterly* (39:4), pp. iii-vi.

Schoop, M. 2001. "An Intro to the Language-Action Perspective," *SIGGROUP Bulletin* (22:2), pp. 3-8.

Schoop, M., de Moor, A., and Dietz, J. 2006. "The Pragmatic Web: A Manifesto," *Communications of the ACM* (49:5), pp. 75-76.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*, Cambridge, UK: Cambridge University Press.

Shen, D., Yang, Q., Sun, J. T., and Chen, Z. 2006. "Thread Detection in Dynamic Text Message Streams," in *Proceedings of the 29th International ACM SIGIR Conference*, Seattle, WA, pp. 35-42.

Smith, M. 2002. "Tools for Navigating Large Social Cyberspaces," *Communications of the ACM* (45:4), pp. 51-55.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. 2001. "A Machine Learning Approach to Coreference Resolution of Noun Phrases," *Computational Linguistics* (27:4), pp. 521-544.

Stolcke, A., Ries, K., Jurafsky, D., and Meteer, M. 2000. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistic* (26:3), pp. 339-373.

Storey, V., Burton-Jones, A., Sugumaran, V., and Purao, S. 2008. "CONQUER: A Methodology for Context-Aware Query Processing on the World Wide Web," *Information Systems Research* (19:1), pp. 3-25.

Szafranski, M., Grandvalet, Y., and Rakotomamonjy, A. 2010. "Composite Kernel Learning," *Machine Learning* (79:1-2), pp. 73-103.

Te'eni, D. 2001. "*Review*: A Cognitive-affective Model of Organizational Communication for Designing IT," *MIS Quarterly* (25:2), pp. 251-312.

Te'eni, D. 2006. "The Language-Action Perspective as a Basis for Communication Support Systems," *Communications of the ACM* (49:5), pp. 65-70.

Twitchell, D., Jensen, M. L., Derrick, D. C., Burgoon, J. K., and Nunamaker Jr., J. F. 2013. "Negotiation Outcome Classification using Language Features," *Group Decision and Negotiation* (22:1), pp. 135-151.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425-478.

Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant EIS," *Information Systems Research* (3:1), pp. 36-59.

Wang, L., Lui, M., Kim, S. N., Nivre, J., and Baldwin, T. 2011. "Predicting Thread Discourse Structure over Technical Web Forums," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, pp. 13-25.

Wang, L., and Oard, D. 2009. "Context-Based Message Expansion for Disentanglement of Interleaved Text Conversations," in *Proceedings Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Boulder, CO, pp. 200-208.

Weick, K. E., Sutcliffe, K. M., and Obstfeld, D. 2005. "Organizing and the Process of Sensemaking," *Organization Science* (16:4), pp. 409-421.

Winograd, T., and Flores, F. 1986. *Understanding Computers and Cognition*, Norwood, NJ: Abex Publishing.

Zabin, J., Nail, J., and Wilder, S. K. 2011. "Gleansight Social Intelligence," *Gleanster Quarterly Reports*.

Zeng, D., Chen, H., Lusch, R., and Li, S. 2010. "Social Media Analytics and Intelligence," *IEEE Intelligent Systems* (25:6), pp. 13-16.

Zhou, Y., Qin, J., and Chen, H. 2006. "CMedPort: An Integrated Approach to Facilitating Chinese Medical Information Seeking," *Decision Support Systems* (42:3), pp. 1431-1448.

About the Authors

Ahmed Abbasi is Murray Research Professor of Information Technology in the McIntire School of Commerce at the University of Virginia. He is Director of the Center for Business Analytics and coordinator for McIntire's Executives-on-Grounds program. Ahmed received his Ph.D. in Information Systems from the University of Arizona, where he also worked as a project-lead in the Artificial Intelligence Lab. He attained an M.B.A. and B.S. in Information Technology from Virginia Tech. Ahmed's research interests relate to predictive analytics and natural language processing with applications in security, health, social media, and customer analytics. His research has been funded through multiple grants from the National Science Foundation. He has also received the IBM Faculty Award, AWS Research Grant, and Microsoft Research Azure Award for his work on Big Data. Ahmed has published over 70 peer-reviewed articles in journals and conferences, including top outlets such as *MIS Quarterly, Journal of Management Information Systems, ACM Transactions on Information Systems, IEEE Transactions on Knowledge and Data Engineering,* and *IEEE Intelligent Systems*. One of his articles was considered a top publication by the Association for Information Systems. He has also won best paper awards from *MIS Quarterly* and WITS. Ahmed's work has been featured in various media outlets, including the *Wall Street Journal*, the Associated Press, WIRED, CBS, and Fox News. He serves as senior editor for *Information Systems Research,* and associate editor for *ACM Transactions on MIS* and *IEEE Intelligent Systems*. He also previously served as associate editor at *Information Systems Research* and *Decision Sciences Journal*, and is a senior member of the IEEE.

Yilu Zhou is an associate professor of Information Systems in the Gabelli School of Business at Fordham University. She received a Ph.D. in Management Information Systems at the University of Arizona, where she also was a research associate at the Artificial Intelligence Lab. She received her B.S. in computer science from Shanghai Jiaotong University. Before joining Fordham University, Yilu was an assistant professor at George Washington University. Her research interests include business intelligence, web/text/data mining, multilingual knowledge discovery and human–computer interaction. Specifically, she investigates and explores computational, intelligent, and automatic ways to discover interesting and useful patterns in news articles, web sites, forums and other social media.

Shasha Deng is an assistant professor of Information Management and Information Systems in the School of Business and Management at Shanghai International Studies University. She holds a Ph.D. in Management Information Systems from Shanghai Jiaotong University, and an M.S. and B.Sc. in computer science and technology from Central South University. Her research interests include decision support systems, business intelligence, big data analytics and text mining. Specifically, her research focuses on the pattern discovery in social media. Her research has been supported by grants from the NSFC.

Pengzhu Zhang is Chair Professor and Director of the Department of Management Information Systems, Antai College of Management and Economics, Shanghai Jiaotong University. Pengzhu received bachelor's and master's degrees from Shandong University of Science and Technology, and a Ph.D. from Xi'an Jiaotong University. His research focuses on the decision support in innovation and health management. His research has been published in many journals including *Decision Support Systems, Information & Management, Computers in Human Behavior, PLos One, Journal of the American Society for Information Science and Technology, Journal of Nanoparticle Research, International Journal of Information Technology & Decision Making, Government Information Quarterly,* and *Journal of Management Analytics*, as well as some important journals in Chinese.

MIS Quarterly

RESEARCH ARTICLE

TEXT ANALYTICS TO SUPPORT SENSE-MAKING IN SOCIAL MEDIA: A LANGUAGE-ACTION PERSPECTIVE

Ahmed Abbasi
McIntire School of Commerce, University of Virginia,
Charlottesville, VA 22908 U.S.A. {abbasi@comm.virginia.edu}

Yilu Zhou
Gabelli School of Business, Fordham University,
New York, NY 10023 U.S.A. {yzhou62@fordham.edu}

Shasha Deng
School of Business and Management, Shanghai International Studies University,
Shanghai, CHINA {shasha.deng@shisu.edu.cn}

Pengzhu Zhang
Antai College of Management and Economics, Shanghai Jiaotong University,
Shanghai, CHINA {pzzhang@sjtu.edu.cn}

Appendix A

Impact of Class Imbalance Resolution Methods on LTAS Performance

For the conversation disentanglement and coherence analysis experiments reported in the main paper, we used threshold moving to deal with the class imbalance issue. In order to illustrate that the LAP-based text analytics systems' (LTAS) results are robust even for the less effective random under-sampling approach, here we report the results for both threshold moving (LTAS-TM) and under-sampling (LTAS-US). For LTAS-US, several bootstrapping runs are utilized with the training data matrix for each run comprising balanced instances using random under-sampling of the majority class. In each bootstrap run, the training matrices are used to build linear SVM classifiers (same as for LTAS-TM). A simple voting scheme applied on top of the bootstrap classifiers' predictions is used to classify test cases as primitive or non-primitive using the soft ensemble method described in Zhou and Liu (2006).

Tables A1 and A2 present the results for LTAS-TM (the same as those reported in the main document), and LTAS-US. Consistent with prior work, the use of threshold moving improved performance over the random under-sampling method. However, LTAS-US also outperformed all benchmarking methods presented in the conversation disentanglement and coherence analysis experiments reported in the main document. The results suggest that the effectiveness of LTAS relative to prior methods is not based on the specific class imbalance resolution method adopted.

Table A1. Results for Conversation Disentanglement Using Threshold Moving Versus Under-Sampling

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS-TM*	**68.7**	**72.5**	**70.6**	**75.7**	**95.0**	**84.2**	**79.9**	**99.2**	**88.5**
LTAS-US*	68.5	71.9	70.2	74.6	89.6	81.4	79.8	98.4	88.1
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS-TM*	**63.6**	**75.4**	**69.0**	**66.4**	**80.1**	**72.6**	**77.4**	**99.4**	**87.0**
LTAS-US*	62.8	74.4	68.1	65.7	79.3	71.8	76.7	95.8	85.2
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS-TM*	**69.7**	**75.6**	**72.5**	**76.8**	**80.5**	**78.6**	**82.5**	**99.6**	**90.3**
LTAS-US*	69.4	74.5	71.8	76.5	79.2	77.8	80.8	95.3	87.5
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
LTAS-TM*	**64.0**	**72.7**	**68.0**						
LTAS-US*	63.0	72.0	67.2						

*Significantly outperformed comparison methods, with all p-values < 0.001

Table A2. Results for Coherence Analysis Using Threshold Moving Versus Under-Sampling

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS-TM*	**77.0**	**85.7**	**81.1**	**80.4**	**95.3**	**87.2**	**87.3**	**95.0**	**91.0**
LTAS-US*	74.4	83.5	78.7	76.4	92.0	83.4	84.3	93.1	88.5
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS-TM*	**72.3**	**86.4**	**78.7**	**71.8**	**90.6**	**80.1**	**84.3**	**88.5**	**86.4**
LTAS-US*	67.3	84.4	74.9	69.0	87.3	77.1	80.7	87.4	83.9
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS-TM*	**77.6**	**84.8**	**81.0**	**79.5**	**80.5**	**83.7**	**90.1**	**94.9**	**92.5**
LTAS-US*	75.0	82.8	78.7	78.1	79.2	81.9	88.4	94.8	91.5
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
LTAS-TM*	**79.4**	**91.0**	**84.8**						
LTAS-US*	76.0	86.7	81.0						

*Significantly outperformed comparison methods, with all p-values < 0.001

Appendix B

Analysis of Primitive Message Detection Classification Method's Design Elements

Two key aspects of the primitive message detection component of the conversation disentanglement module of LTAS are the use of a feature vector comprising message bins coupled with average and max similarity scores for messages preceding and following the message of interest. Collectively, the use of these design elements is intended to facilitate inclusion of proximity and thematic trend information indicative of topic drift and new conversation emergence. In order to test the efficacy of these two elements, we evaluated the proposed primitive message detection method (labeled Bins-Ave&Max here) against one devoid of message bins. This method, labeled NoBins-Ave&Max, used four features: average and max similarity from all prior and subsequent features, respectively to demonstrate the utility of the bin feature. To examine the usefulness of including both average and max similarity from messages preceding and following, as opposed to just focusing on average similarity from prior messages and max similarity with subsequent ones, two additional settings were included: Bins-Ave/Max and NoBin-Ave/Max. Overall, the 4 settings (2 × 2 design) were meant to shed light on the additive benefit of each of the two design elements.

Table B1. Results for Primitive Message Detection

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-Ave&Max	**60.7**	**74.7**	**67.0**	**68.2**	**93.8**	**79.0**	**62.4**	**94.3**	**75.1**
Bins-Ave/Max	58.2	71.7	64.3	65.2	93.4	76.8	60.9	91.5	73.1
NoBin-Ave&Max	55.3	71.7	62.5	65.7	90.8	76.3	59.2	89.3	71.2
NoBin-Ave/Max	54.6	70.4	61.5	62.7	86.6	72.7	58.5	91.5	71.4
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-Ave&Max	**63.3**	**69.6**	**66.3**	**57.6**	**89.2**	**70.0**	63.6	98.6	77.3
Bins-Ave/Max	59.4	66.4	62.7	55.9	87.8	68.3	**63.7**	**99.6**	**77.7**
NoBin-Ave&Max	54.3	64.0	58.7	52.7	83.3	64.6	62.8	98.4	76.7
NoBin-Ave/Max	53.4	63.5	58.0	51.6	81.4	63.2	57.6	97.2	72.3
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-Ave&Max	**71.6**	**84.4**	**77.5**	**62.0**	**87.6**	**72.6**	**59.9**	**95.5**	**73.6**
Bins-Ave/Max	69.0	81.9	74.9	60.3	87.6	71.5	59.3	92.8	72.4
NoBin-Ave&Max	63.8	80.4	71.1	59.3	87.4	70.6	56.6	92.8	70.3
NoBin-Ave/Max	61.8	79.6	69.6	58.9	86.8	70.2	57.6	97.2	72.3
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
Bins-Ave&Max	**66.9**	**70.6**	**68.7**						
Bins-Ave/Max	60.6	69.1	64.6						
NoBin-Ave&Max	58.0	66.9	62.1						
NoBin-Ave/Max	55.0	64.7	59.5						

The experiment results are presented in Table B1. The proposed method outperformed all three alternative settings on 9 of the 10 test beds, with performance gains ranging from 1% to 4% with respect to precision, recall, and f-measure. On the health tweets data set, the Bins-Ave/Max method, where the average similarity from preceding messages and the max similarity from subsequent messages was utilized, performed marginally better.

Figure B1 depicts the f-measures for Bins-Ave&Max and comparison methods across each of the 1615 discussion threads. The chart on the left shows mean f-measures for threads encompassing 1 to 10+ conversations. The chart on the right shows mean f-measures by thread length percentile rankings, with lower percentile values on the horizontal axis indicating shorter thread lengths. Whereas all four methods performed comparably on shorter threads and/or ones encompassing two or fewer conversations, the inclusion of bins and both average and max similarity for preceding and subsequent messages enabled Bins-Ave&Max to outperform comparison methods on lengthier threads or those with three or more conversations. In some cases, F-measures tended to be lower on shorter threads with only a single conversation due to lower precision rates since even a single false positive in a thread would drop the overall precision dramatically. It is also important to note that the NoBins-Ave&Max did outperform Bins-Ave&Max with respect to average f-measure on threads of length in the 20th percentile or lower. However, the markedly enhanced performance of Bins-Ave&Max on threads of above average length resulted in the better overall performance.

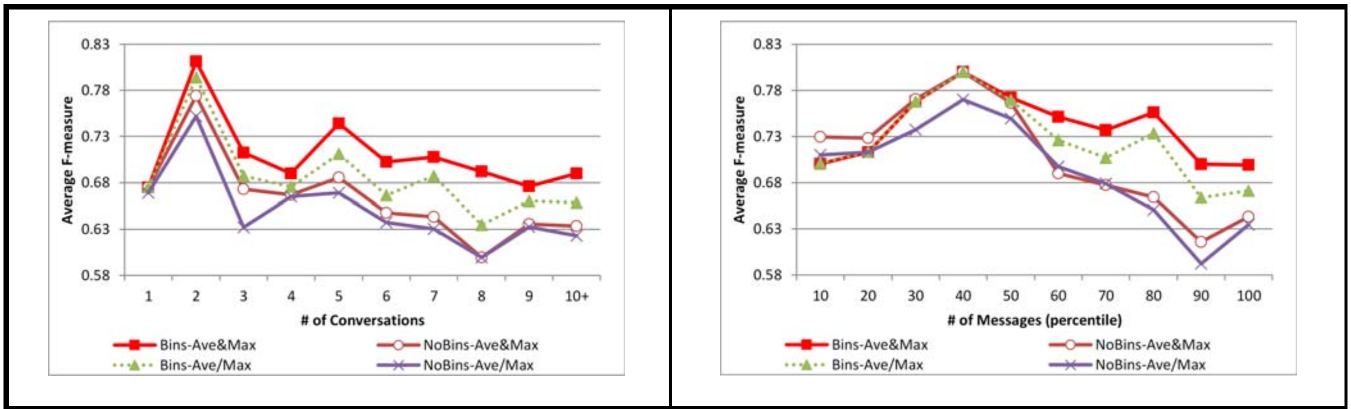
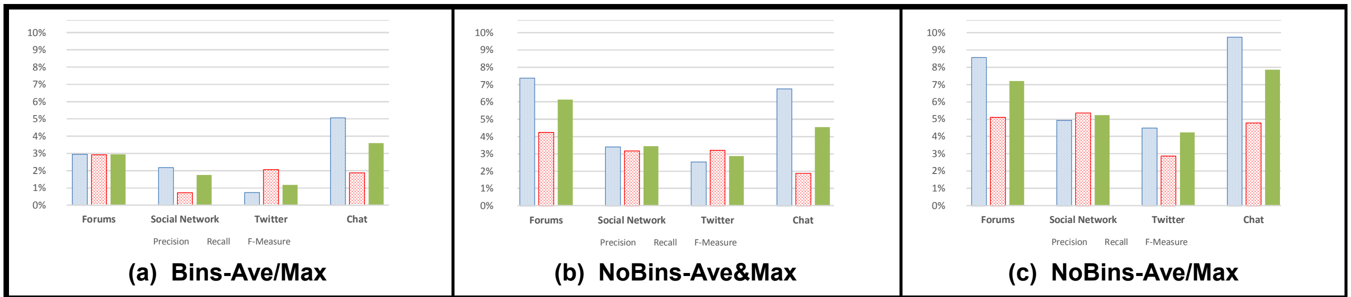


Figure B1. Average f-Measures for Proposed Bins-Ave&Max Method and Alternatives Across Discussion Threads Grouped by Number of Conversations (left) and Number of Messages (right)

Figure B2 shows the precision, recall, and f-measures for the three alternative settings relative to Bins-Ave&Max, aggregated by social media channel. The performance gains were most pronounced on the web forum and chat data sets, where average thread lengths and messages per conversation tend to be higher. However, even on the social networking and microblog data sets, the proposed method's primitive message detection rates were at least 2% to 5% higher.



(a) Bins-Ave/Max

(b) NoBins-Ave&Max

(c) NoBins-Ave/Max

Figure B2. Performance Deltas Relative to Bins-Ave&Max Method Used in LTAS Across Various Social Media Channels in Test Bed

Unlike the conversation affiliation classifier, the primitive message detection component only utilized average and max similarity. The rationale for including variance for the second stage of the disentanglement (i.e., affiliation classification) was to alleviate the impact of relying on three varying-sized bins (before, in-between, and after) which can become accentuated on lengthier threads, and to gauge the pervasiveness of intertwined conversations. Since the primitive message detection component uses fixed size bins and focuses on a different classification task,

Table B2. Impact of Including Variance Measures in Primitive Message Detection Component

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-AveMax	**60.7**	**74.7**	**67.0**	**68.2**	**93.8**	**79.0**	**62.4**	**94.3**	**75.1**
Bins-AveMaxVr	60.7	74.7	67.0	67.9	93.8	78.7	61.5	93.3	74.1
Health									
Technique	Web Forum			Social Network (Patients)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-Ave&Max	**63.3**	**69.6**	**66.3**	**57.6**	**89.2**	**70.0**	**63.6**	**98.6**	**77.3**
Bins-AveMaxVr	62.7	69.4	65.9	57.4	89.1	69.8	61.6	98.0	75.7
Security									
Technique	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-Ave&Max	71.6	84.4	77.5	**62.0**	**87.6**	**72.6**	**59.9**	**95.5**	**73.6**
Bins-AveMaxVr	**72.0**	**84.9**	**77.9**	62.0	87.6	72.6	59.3	93.3	72.5
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
Bins-Ave&Max	66.9	70.6	68.7						
Bins-AveMaxVr	**67.3**	**70.9**	**69.1**						

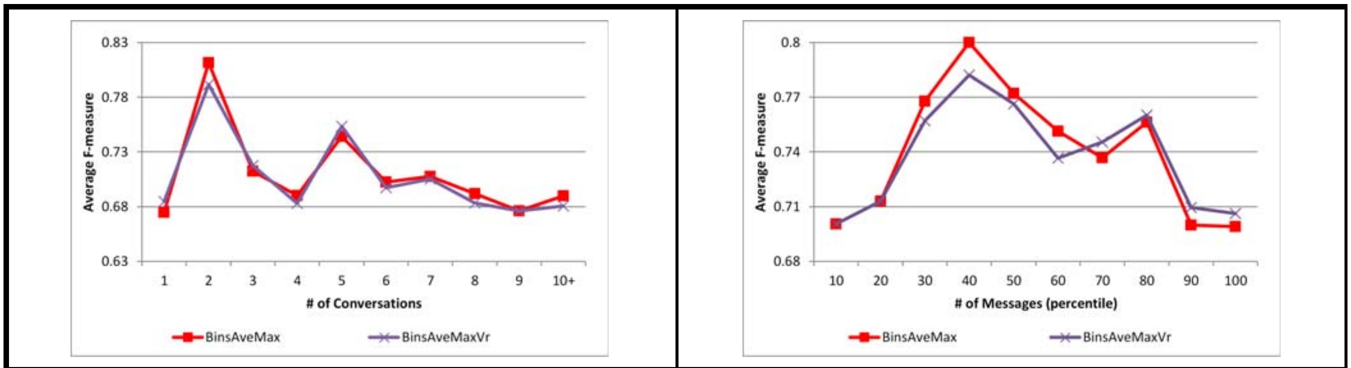


Figure B3. Average f-Measures for BinsAveMax Method and BinsAveMaxVr Alternative Across Discussion Threads Grouped by Number of Conversations (left) and Number of Messages (right)

whether a given message is the beginning of a new conversation, variance in similarities for messages in a bin did not seem as pertinent. Nevertheless, we empirically examined the impact of hypothetically adding variance to the primitive message detection component. The comparison results appear in Table B2. The inclusion of variance did improve f-measure by about 0.5% on the manufacturing chat and security forum data sets. It also resulted in comparable performance on the telecom forum and security social networking data. However in general, performance was either similar or marginally worse. The results suggest that variance measures useful in the affiliation classification phase may not be as valuable for primitive message detection due to differences in the problem task and design of the classification method.

In order to further illustrate this point, Figure B3 depicts the f-measures for the BinsAveMax approach utilized and the BinsAveMaxVr alternative across each of the 1,615 discussion threads. The chart on the left shows mean f-measures for threads encompassing 1 to 10+ conversations. The chart on the right shows mean f-measures by thread length percentile rankings, with lower percentile values on the horizontal axis indicating shorter thread lengths. From the chart on the left, we can see that the inclusion of variance for primitive message detection does not have any meaningful impact as the number of conversations per thread increases. Similarly, while variance information causes a slight lift on the lengthiest threads (i.e., in the 90th percentile or greater), this is offset by poorer performance on threads in the 30th through 60th percentile.

Appendix C

Impact of Fixed Binning on Primitive Message Detection Performance

The primitive message detection component uses a fixed bin approach due to representational constraints: all feature vectors instances in the training and testing set need to have input vectors of the same size since these vectors are converted using dot product, by the linear SVM kernel. As shown in Appendix B, the use of bins improves performance over methods that do not leverage bins since it enables inclusion of sequential trend and proximity-sensitive similarity measurement for enhanced primitive message detection. However, using fixed bin quantities for messages preceding and following a given message in a thread could create considerable variation in the quantities of messages per bin for two reasons: (1) differences in the positions of messages within a thread (for example, the last message in the thread would have 0 subsequent messages and a greater number of preceding messages per bin relative to all other messages in that thread) and (2) variation in the length of threads. For this latter point, Figure B1 in Appendix B already illustrates how the proposed method's f-measure is more than 10% lower on threads below the 20th percentile or above the 90th percentile with respect to number of messages.

Variation in the quantity of messages per bin is important to investigate since the average and max similarity measures per bin are features computed for each message in the training and testing set, and patterns based on these features are the basis for the primitive message detection model training and classification in the proposed method. In order to investigate the interplay between number of bins, message positions, and thread lengths, we plotted the bin message probability mass across all 1,615 threads and 25,157 messages in the test bed, for varying quantities of bins (i.e., primitive message detection with n = 1 through n = 6). Figure C1 presents the analysis results. In the figure, the charts' x-axes represent bin sizes in messages and the y-axes signify percentage of total bins. Looking at the six charts, it is apparent that the use of more bins dramatically decreases variation in the bin size distributions by compressing the range and converging towards fewer, higher occurrence likelihood bin sizes. This makes sense since the set of bin sizes following messages in a thread of length *l* for any value of n can be represented by $\left\{\left[\left[\frac{l-1}{n}\right],\left[\frac{l-1}{n}\right],\left[\frac{l-2}{n}\right],\left[\frac{l-2}{n}\right]\right]\cdots\left[0,\left[\frac{1}{n}\right]\right],\frac{0}{n}\right\}$. The results suggest that when incorporating information from surrounding messages that precede and follow a given message in a discussion thread, the use of larger bin sizes helps reduce variation in the quantity of messages per bin.

Next, we analyzed the impact of different values of n on primitive detection classification performance. The results appear in Figure C2. The left chart in Figure C2 shows mean f-measures by thread length percentile rankings, with lower percentile values on the horizontal axis indicating shorter thread lengths. Based on this chart, it is apparent that using fewer bins results in somewhat better f-measures on shorter threads (e.g., n = 1, n = 2, and n = 3), whereas larger values for n produce better results on lengthier threads (i.e., n = 5 and n = 6). However, the performance margins appear greater for higher values of n on lengthier threads relative to lower values on shorter ones, further underscoring the utility of bins. Interestingly, varying values of n create what is analogous to a "see-saw" effect with the pivot point being messages in the 50th and 60th percentiles, and larger values of n resulting in a positive slope, whereas smaller values create a negative one. In order to further illustrate this see-saw effect, the chart on the right side of Figure C2 depicts average f-measures across larger percentile ranges: 0–40th, 50th–60th, and 70th–100th. In this chart, the increasing f-measures for larger values of n on lengthier threads and corresponding decrease in f-measures for smaller values, and vice versa for shorter threads, is more readily apparent. Possibly due to the lesser variation in bin sizes, though not depicted here, the larger bin sizes (i.e., n = 5 and n = 6) had higher area under the curve values for the left chart in Figure C2, and higher overall f-measure for primitive message detection.

This finding is intuitive: in lengthier threads, using a larger number of bins helps to reduce variation in bin sizes. On the other hand, using a larger number of bins on shorter threads creates bins that are sparser with respect to number of messages. Overall, the results presented in the appendix further shed light on the value of using bins for primitive message detection, but also highlight some limitations of the approach; namely performance variations attributable to thread length. One future direction may be to use an ensemble of classifiers trained specifically on threads of a shorter length. For instance, based on some preliminary analysis, three classifiers for threads of length below the 40th percentile, above the 70th percentile, and in-between, each with their own respective value for n, could help enable an elevated, and "flatter" line for f-measure across thread lengths. We believe the analysis presented in the main document and appendices will set a foundation for future work that can further the state-of-the-art for primitive message detection oriented towards enhancing sense-making.

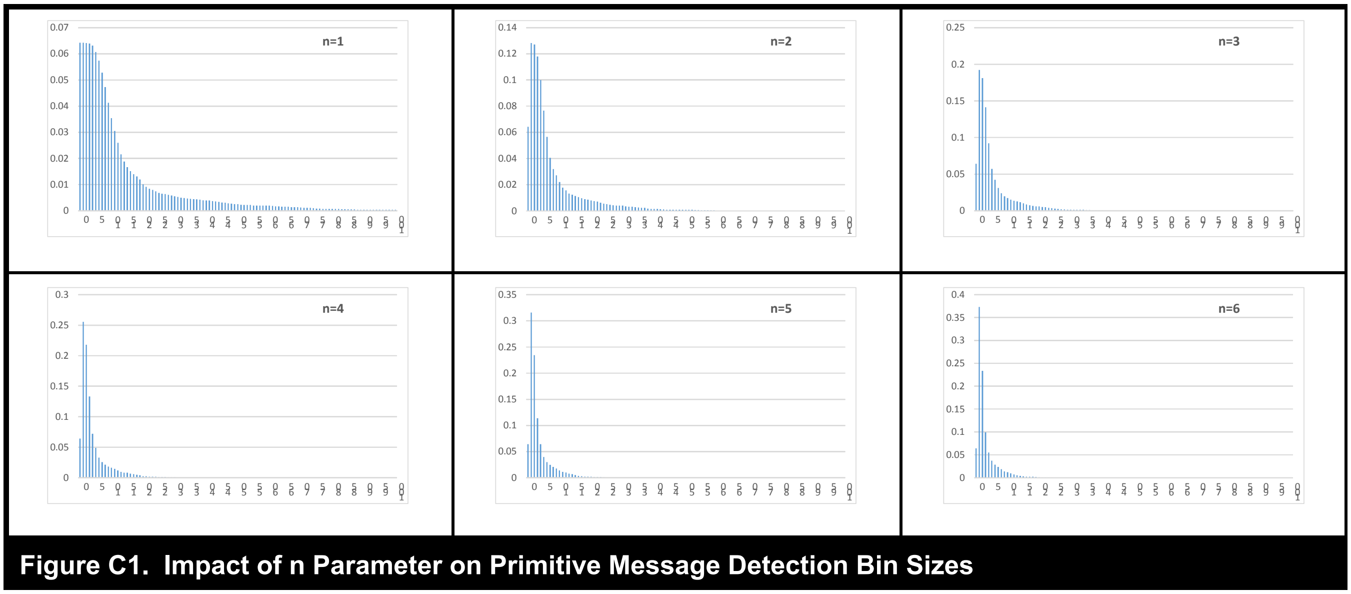


Figure C1. Impact of n Parameter on Primitive Message Detection Bin Sizes

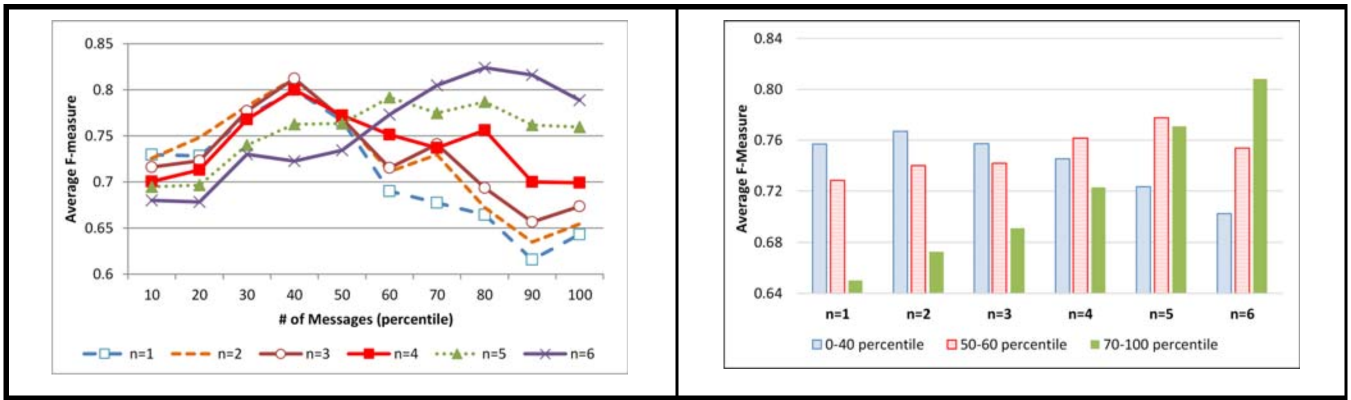


Figure C2. Average f-Measures for Primitive Message Detection Method Using Varying Values for n, Grouped by Thread Length Percentiles (left) and Percentile Range Aggregation (right)

Appendix D

Impact of Bins and Average/Max/Var Similarity on Conversation Affiliation Classification Performance

Similar to the results presented in Appendix B for analysis of the impact of similarity measures and bin usage on primitive message detection performance, here we examined the impact of the bins and the variance measure on conversation affiliation detection performance. The proposed method, labeled here as Bins-AvMxVr, was compared against four alternative variations: Bins-AvMx which was devoid of the variance measure; NoBins-AvMxVr and NoBinsAvMx which were devoid of bins, or bins and variance measure, respectively; and Bins-AvMxSz where variance was replaced with a "bin size" variable signifying the quantity of messages in that particular region.

Table D1. Impact of Region Bins and Similarity Measures on Conversation Disentanglement

Telecom									
Technique	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-AvMxVr*	**68.7**	**72.5**	**70.6**	**75.7**	**95.0**	**84.2**	**79.9**	**99.2**	**88.5**
Bins-AvMx	64.7	67.9	66.3	74.0	92.9	82.4	79.3	95.5	86.7
Bins-AvMxSz	65.6	68.8	67.2	73.6	91.8	81.7	76.1	95.5	84.7
NoBins-AvMxVr	61.9	64.9	63.4	72.2	89.6	80.0	73.9	92.8	82.3
NoBins-AvMx	60.0	63.8	61.9	71.6	89.2	79.4	73.7	92.8	82.2
Health									
Technique	Web Forum			Social Network (Patients)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-AvMxVr*	**63.6**	**75.4**	**69.0**	**66.4**	**80.1**	**72.6**	**77.4**	**99.4**	**87.0**
Bins-AvMx	60.0	71.3	65.2	63.9	76.5	69.6	75.9	96.9	85.1
Bins-AvMxSz	60.4	70.7	65.1	63.7	78.0	70.1	75.8	96.3	84.8
NoBins-AvMxVr	55.8	66.6	60.7	60.9	73.1	66.4	73.1	94.9	82.6
NoBins-AvMx	53.8	64.7	58.8	59.5	71.8	65.1	72.6	92.5	81.3
Security									
Technique	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Bins-AvMxVr*	**69.7**	**75.6**	**72.5**	**76.8**	**80.5**	**78.6**	**82.5**	**99.6**	**90.3**
Bins-AvMx	65.7	70.4	67.9	73.5	78.8	76.1	79.8	96.9	87.5
Bins-AvMxSz	66.1	70.0	68.0	74.3	77.9	76.1	78.0	95.2	85.8
NoBins-AvMxVr	62.6	66.4	64.4	72.5	77.2	74.8	78.8	95.2	86.2
NoBins-AvMx	60.7	64.4	62.5	71.3	75.8	73.5	77.9	94.4	85.4
Manufacturing									
Technique	Chat								
	Prec.	Rec.	F-Meas						
Bins-AvMxVr*	**64.0**	**72.7**	**68.0**						
Bins-AvMx	59.4	68.5	63.6						
Bins-AvMxSz	59.3	69.7	64.1						
NoBins-AvMxVr	55.7	63.6	59.4						
NoBins-AvMx	53.9	60.6	57.1						

*Significantly outperformed alternative methods, with all p-values < 0.001

The experiment results on the 10 test beds appear in Table D1. The use of the three region bins coupled with average, max, and variance in similarity measures outperformed alternatives by about 2% to 10% across methods and test beds. Excluding variance information caused performance on the lengthier thread-oriented web forum test beds to drop by about 4%, whereas the performance delta was about 2% on social networking and microblogging data sets. Replacing variance with a "bin size" variable did not help. The absence of bins had a more profound impact, with NoBins-AvMx-Vr outperformed by 4% to 6% on average by the proposed method. The results empirically underscore the utility of the key design elements for the conversation affiliation method incorporated.

In order to dig a bit deeper into the implications of including/excluding variance information on performance by thread length and number of conversations, we compared Bins-AvMxVr against Bins-AvMx, as well as the top-performing comparison method (i.e., Elsner and Charniak 2010). The left chart in Figure D1 shows mean f-measures by thread length percentile rankings, with lower percentile values on the horizontal axis indicating shorter thread lengths. From the figure it is evident that the inclusion of variance information in Bins-AvMxVr enabled better performance on threads in the 40th percentile of higher in terms of number of messages. The right chart depicts performance grouped by number of conversations per thread. Incorporating variance information in Bins-AvMxVr enabled better performance on threads encompassing three or more conversations. As noted in the paper, conversation disentanglement becomes more challenging as thread lengths and number of conversations increase, due to growth in the potential solution space and greater potential intertwining of conversations. In the main paper, we illustrated how the conversation disentanglement component in LTAS was more robust against performance drop-offs attributable to increasing length and quantity of conversations, relative to comparison methods. For instance, the best-performing comparison method (e.g., Elsner and Charniak 2010) observed f-measure drops of 28% and 34% across thread lengths and number of conversations, respectively. In contrast, the performance drops for the disentanglement method in LTAS were only 15% to 18%. The results depicted in Figure D1 suggest that while BinsAvMx was also effective relative to comparison methods, the inclusion of variance measures further enhanced the method's robustness.

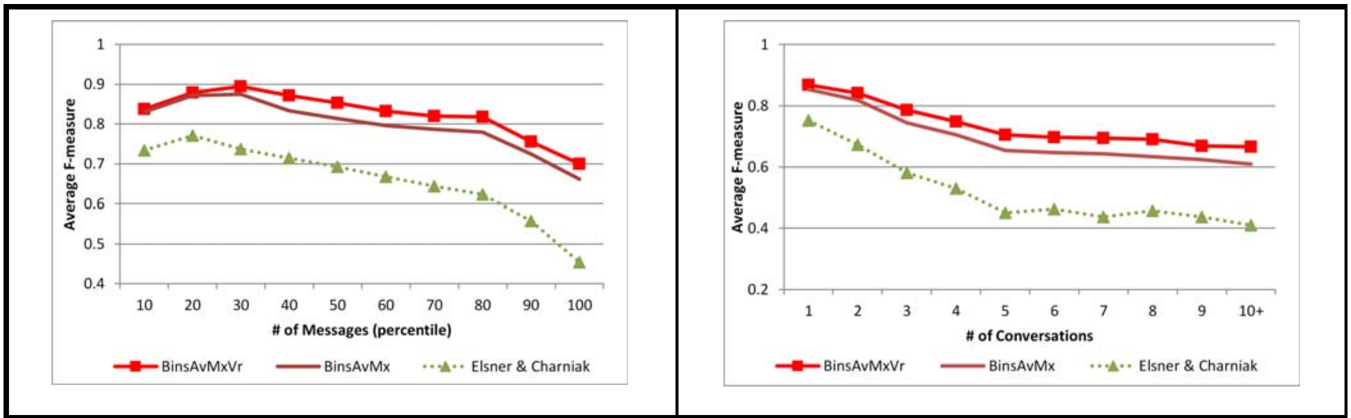


Figure D1. Average f-Measures for Conversation Disentanglement, Grouped by Number of Messages (left) and Number of Conversations (right)

Appendix E

Impact of Primitive Message Detection on Conversation Disentanglement and Coherence Analysis

In order to test the efficacy of the proposed primitive message detection component of LTAS, we examined the performance of the conversation disentanglement module without primitive message detection. In the absence of primitive message labels (i.e., messages labeled "A"), average, max, and variance between messages *X* and *Y* are only computed on the three message bins C_1, C_2, and C_3 (i.e., no A_1, A_2, and A_3 bins). Figure E1 shows the revised conversation disentanglement classification method devoid of primitive message detection, which can be contrasted with the actual LTAS method depicted in Figure 7 of the main document.

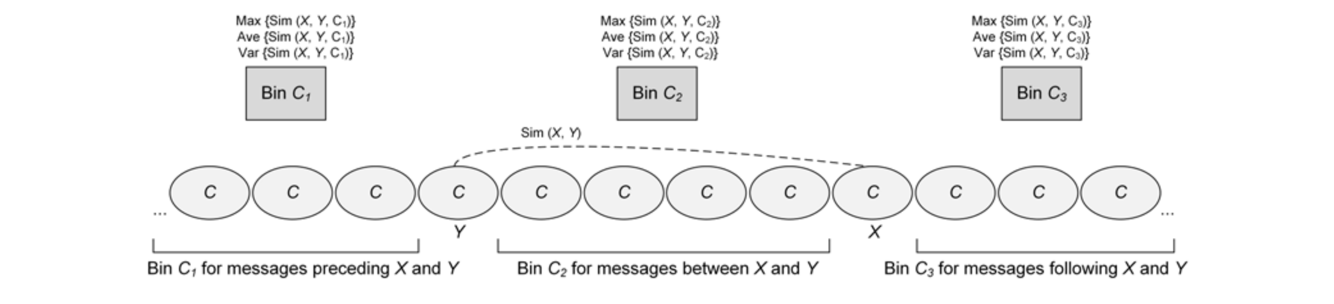


Figure E1. Illustration of Regions, Bins, and Similarity Scores Used in Affiliation Classification Stage Devoid of Primitive Message Detection Component

Table E1. Impact of Primitive Detection on Conversation Disentanglement Performance

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Primitive*	**68.7**	**72.5**	**70.6**	**75.7**	**95.0**	**84.2**	**79.9**	**99.2**	**88.5**
No Primitive	53.0	64.2	58.1	66.7	83.0	74.0	68.3	92.3	78.5
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Primitive*	**63.6**	**75.4**	**69.0**	**66.4**	**80.1**	**72.6**	**77.4**	**99.4**	**87.0**
No Primitive	49.2	66.8	56.6	54.2	72.9	62.1	66.6	95.6	78.5
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Primitive*	**69.7**	**75.6**	**72.5**	**76.8**	**80.5**	**78.6**	**82.5**	**99.6**	**90.3**
No Primitive	53.6	65.8	59.1	65.7	72.6	69.0	69.6	93.8	79.9
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
Primitive*	**64.0**	**72.7**	**68.0**						
No Primitive	49.0	59.8	53.9						

* Significantly outperformed conversation disentanglement method devoid of primitive message detection, with all p-values < 0.001.

Table E1 presents the experiment results for conversation disentanglement devoid of primitive message detection, relative to the primitive message detection-inclusive approach incorporated as part of LTAS. Including primitive message detection enabled a 10% boost in f-measure on average. While recall rates were 8% higher on average, the biggest gain was in precision (12% average across data sets), suggesting that the inclusion of primitive message labels during the affiliation classification phase helps reduce false positives.

The conversation disentanglement component provides many key conversation structure attributes used in the coherence analysis module of LTAS. In fact, four of the eight conversation structure attributes used in the coherence analysis module are explicitly based on primitive message detection: message status, between status, and prior status (see Table 3 in the main document for details). Furthermore, the diminished performance of the conversation affiliation method also impacts the quality of the conversation status attribute. In order to empirically examine the impact of not having primitive message detection on coherence analysis, Table E2 presents the coherence analysis results with and without primitive message detection. On average, the absence of primitive message detection reduced f-measures by 11%, with performance deltas as high as 16% to 18% on the security and telecommunications web forums. The results further underscore the efficacy of primitive message detection for conversation disentanglement and coherence analysis in social media.

Table E2. Results for Coherence Analysis Without Primitive Message Detection

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Primitive*	**77.0**	**85.7**	**81.1**	**80.4**	**95.3**	**87.2**	**87.3**	**95.0**	**91.0**
No Primitive	61.9	64.0	62.9	70.8	86.6	77.9	75.1	88.1	81.1
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Primitive*	**72.3**	**86.4**	**78.7**	**71.8**	**90.6**	**80.1**	**84.3**	**88.5**	**86.4**
No Primitive	60.2	74.3	66.5	65.7	82.2	73.0	72.5	81.9	76.9
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Primitive*	**77.6**	**84.8**	**81.0**	**79.5**	**88.3**	**83.7**	**90.1**	**94.9**	**92.5**
No Primitive	62.1	67.5	64.7	70.3	82.8	76.1	80.1	88.3	84.0
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
Primitive*	**79.4**	**91.0**	**84.8**						
No Primitive	67.2	80.6	73.3						

*Significantly outperformed coherence analysis method devoid of primitive message detection, with all p-values < 0.001.

Appendix F

Contribution of Linguistic and Conversation Structure Features to Coherence Analysis Performance

The enhanced performance of the LTAS coherence analysis module is largely attributable to the inclusion of conversation structure and linguistic attributes guided by LAP-based principles. In order to test the utility of these features, we analyzed the coherence analysis performance when using all system, linguistic, and conversation structure attributes (i.e., all depicted in Table 3 in the main document) versus combinations devoid of conversation structure (labeled Sys-Ling) and linguistic (labeled Sys-Constr) features. The same experiment design and settings as the original experiments presented in the main document were employed. The results across the 10 test beds appear in Table F1. The exclusion of conversation structure or linguistic features significantly reduced performance, with average decreases in f-measures ranging from 13% to 16%, respectively. The results lend credence to the feature set incorporated, which encompasses key system, linguistic, and conversation structure attributes useful for coherence analysis.

Table F1. Impact of Feature Set Combinations on Coherence Analysis Performance

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Feature Set	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Sys-Ling-ConStr*	**77.0**	**85.7**	**81.1**	**80.4**	**95.3**	**87.2**	**87.3**	**95.0**	**91.0**
Sys-Ling	61.5	63.6	62.5	70.5	87.0	77.9	74.7	85.0	79.5
Sys-ConStr	58.9	61.8	60.3	66.2	79.7	72.3	68.5	82.9	75.0
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Feature Set	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Sys-Ling-ConStr*	**72.3**	**86.4**	**78.7**	**71.8**	**90.6**	**80.1**	**84.3**	88.5	**86.4**
Sys-Ling	56.6	64.4	60.2	65.2	79.7	71.7	71.0	81.3	75.8
Sys-ConStr	54.5	61.4	57.7	65.7	80.8	72.5	71.4	82.0	76.3
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Feature Set	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Sys-Ling-ConStr*	**77.6**	**84.8**	**81.0**	**79.5**	**88.3**	**83.7**	**90.1**	**94.9**	**92.5**
Sys-Ling	59.8	62.9	61.3	69.2	80.1	74.2	79.3	87.3	83.1
Sys-ConStr	56.9	59.9	58.4	65.2	74.8	69.6	75.9	85.3	80.3
Manufacturing									
	Chat								
Feature Set	Prec.	Rec.	F-Meas						
Sys-Ling-ConStr*	**79.4**	**91.0**	**84.8**						
Sys-Ling	65.7	77.9	71.3						
Sys-ConStr	59.7	65.5	62.5						

*Significantly outperformed comparison feature set combinations, with all p-values < 0.001.

As mentioned in the section "A LAP-Based Text Analytics System for Sense-Making in Online Discourse" of the main paper, the key output of the conversation disentanglement stage are the primitive message and conversation affiliation *variables*. These variables are at the core of the conversation structure features used for coherence analysis and the speech act classification stage's initial classifier. The performance lift for coherence analysis attributable to these conversation structure variables was demonstrated in the results presented in Table F1. As discussed in the main paper, one important thing to note is that *conversation affiliations are not finalized after the disentanglement stage*. Rather, they are finalized once the conversation tree is constructed as the output of the coherence analysis stage. This is why the coherence analysis method compares all message pairs within the entire thread (not just ones within conversations). The rationale for not finalizing conversation affiliations until the coherence analysis stage is to allow provisions for error correction with respect to inaccurate conversation affiliation classifications. Here we present empirical evidence that by waiting until after the coherence analysis stage to finalize conversation affiliations, both conversation affiliation performance and coherence analysis (i.e., reply-to performance) are enhanced. In order to demonstrate this point, we performed two sets of analyses:

(1) Analysis showing that conversation affiliations resulting from the conversation trees output by the coherence analysis phase are *more accurate* than the conversation affiliation classifier presented in the subsection "Conversation Affiliation Classification" of the paper (which handily outperformed existing methods).

(2) Analysis demonstrating that applying coherence analysis only within conversations identified by the affiliation classification phase *would have hurt* coherence analysis performance.

Table F2 shows the conversation disentanglement results after the coherence analysis phase (i.e., for the generated conversation trees) versus the conversation affiliation classification module of LTAS. By not finalizing conversation affiliations until after the coherence analysis stage, conversation disentanglement f-measures increased considerably (by 7 to 22 percentage points).

Table F2: Conversation Disentanglement Performance for Coherence Analysis Module's Conversation Tree Output Versus Conversation Affiliation Classifier

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Convo-Tree*	**88.6**	**90.8**	**89.7**	**94.4**	**96.3**	**95.3**	**95.6**	98.4	**97.0**
Convo-Affil-Class	68.7	72.5	70.6	75.7	95.0	84.2	79.9	**99.2**	88.5
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Convo-Tree*	**90.0**	**91.6**	**90.8**	**90.9**	**92.2**	**91.6**	**94.5**	97.8	**96.1**
Convo-Affil-Class	63.6	75.4	69.0	66.4	80.1	72.6	77.4	**99.4**	87.0
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Convo-Tree*	**89.6**	**91.5**	**90.5**	**90.4**	**91.9**	**91.1**	**95.9**	98.2	**97.0**
Convo-Affil-Class	69.7	75.6	72.5	76.8	80.5	78.6	82.5	**99.6**	90.3
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
Convo-Tree*	**90.0**	**91.4**	**90.7**						
Convo-Affil-Class	64.0	72.7	68.0						

*Significantly outperformed the conversation affiliation classification module in terms of f-measure, with all p-values < 0.001.

Presently, coherence relations are examined across all messages within a given discussion thread. In this analysis section, we refer to this LTAS approach as "EntireThread" We examined the impact of applying coherence analysis only within hypothetical conversation groups generated by the conversation affiliation classifier. In order to convert binary conversation affiliation classifications into conversation groups, we adopted an overlapping clustering approach where a given message could be affiliated with multiple conversations. For example, if message Z was affiliated with messages X and Y, where both X and Y were in different groups, the resulting groups would be X-Z and Y-Z. We then applied coherence analysis only within each group of messages, by comparing messages appearing later (temporally) within a group against all those appearing earlier. Precision, recall, and f-measures were computed on the resulting reply-to relations. We compared this WithinConvoOnly method against the EntireThread approach adopted in LTAS. Table F3 presents the analysis results. Restricting coherence analysis to WithinConvoOnly (i.e., essentially finalizing conversation affiliations after the disentanglement component of LTAS) caused coherence analysis f-measures to drop by 5-10 percentage points.

Table F3. Impact of Restricting Coherence Analysis to Within Conversations on Performance

Telecom									
Feature Set	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
EntireThread*	**77.0**	**85.7**	**81.1**	**80.4**	**95.3**	**87.2**	**87.3**	**95.0**	**91.0**
WithinConvoOnly	69.6	73.8	71.7	73.5	85.0	78.8	82.0	88.5	85.1
Health									
Feature Set	Web Forum			Social Network (Patients)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
EntireThread*	**72.3**	**86.4**	**78.7**	**71.8**	**90.6**	**80.1**	**84.3**	**88.5**	**86.4**
WithinConvoOnly	64.7	74.5	69.3	67.9	82.6	74.5	78.8	82.0	80.4
Security									
Feature Set	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
EntireThread*	**77.6**	**84.8**	**81.0**	**79.5**	**88.3**	**83.7**	**90.1**	**94.9**	**92.5**
WithinConvoOnly	69.9	73.0	71.4	73.9	85.4	79.2	84.5	90.6	87.5
Manufacturing									
Feature Set	Chat								
	Prec.	Rec.	F-Meas						
EntireThread*	**79.4**	**91.0**	**84.8**						
WithinConvoOnly	59.7	65.5	62.5						

*Significantly outperformed WithinConvoOnly setting, with all p-values < 0.001.

Given that the coherence analysis classification module also employs a binary classification scheme to determine reply-to relations, potential issues could arise when, for example, message Z might be considered to reply to X and Y. This could present challenges for the tree structure (where each child node belongs to a single parent), and for conversation affiliation (if X and Y are in different conversations). However, as noted the last two paragraphs of subsection "Coherence Analysis" in the main paper, multi-reply cases occur only for 1% to 2% of message classifications (and rarely for cases where the parent nodes are in different conversations). Nevertheless, for such cases duplicate child nodes are created under each parent along with their sub-tree (i.e., all child nodes for the duplicated node). Figure F1 illustrates how this is done.

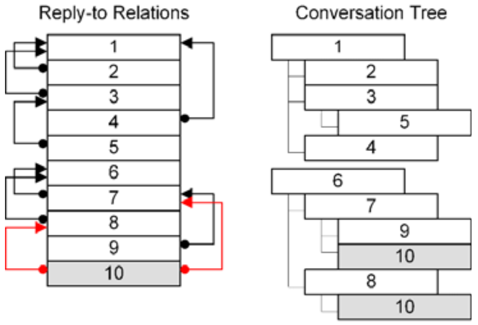


Figure F1. Illustration of How Duplicate Nodes are Created for Child Messages with Multiple Parents

Appendix G

Comparison of LTAS Two-Stage Speech Act Classifier and Initial Classifier

In order to examine the utility of the two-stage lbeled tee classifier utilized in LTAS for speech act identification, we compared its performance against the initial classifier. The results are presented in Table G1. Incorporating the kernel-based labeled tree boosted accuracies by 19% to 24% across the 10 data sets. Additionally, as shown in Figure G1, the enhanced performance of the labeled tree kernel was consistent across the major speech act categories pervasive in our test bed: assertives, suggestions, questions, and commissives. The performance of the initial classifier was slightly better than the n-gram, n-word, and CRF methods. However, the inclusion of the labeled tree kernel facilitated performance gains necessary to significantly outperform the collective classification and joint classification benchmarks. The results are consistent with prior LAP-based studies, which have emphasized the interplay between conversation structure and speech act composition, and how the two are interrelated.

We also believe an interesting future research direction would be to leverage the two-stage classifier as input for itself in an iterative/recursive/adaptive manner as done in prior methods such as tri-training (Zhou et al. 2005).

Table G1. Accuracies for Initial and Labeled Tree Speech Act Classifiers Incorporated in LTAS

Classification Method	Telco			Health			Security			Manu.
	Forum	Social	Twitter	Forum	Social	Twitter	Forum	Social	Twitter	Chat
LTAS – Labeled Tree*	**92.1**	**92.5**	**93.3**	**93.6**	**93.0**	**95.5**	**91.9**	**90.4**	**93.7**	**90.7**
LTAS – Initial Classifier	66.0	69.6	69.8	68.1	68.6	70.1	67.9	71.2	69.5	66.6

*Significantly outperformed initial classifier, with all p-values < 0.001.

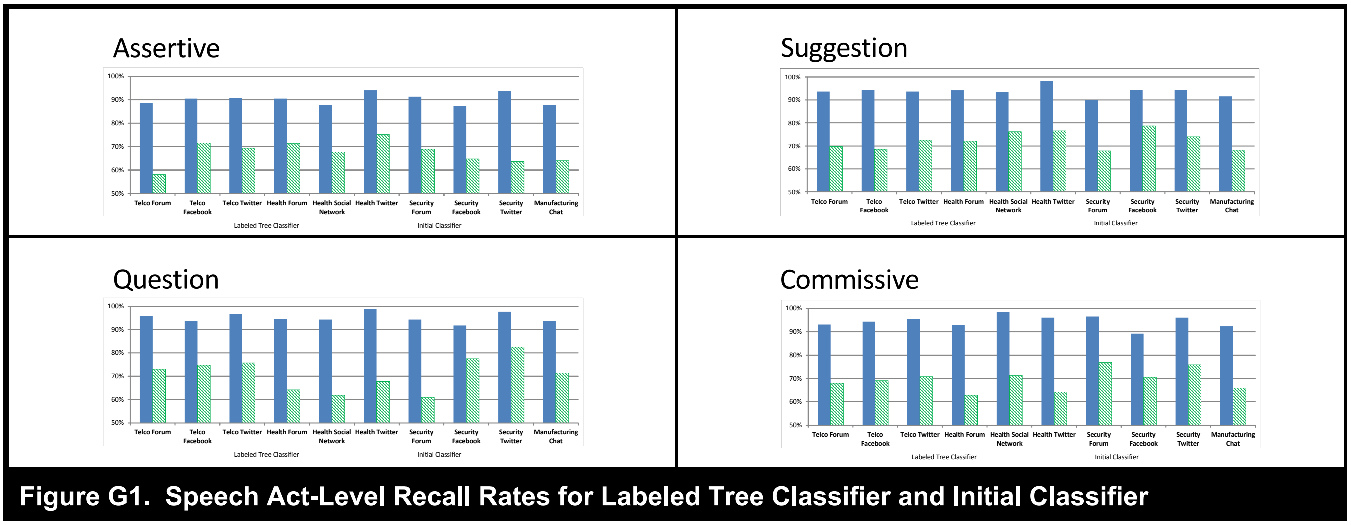


Figure G1. Speech Act-Level Recall Rates for Labeled Tree Classifier and Initial Classifier

Reference

Zhou, Z. H., and Li, M. 2005. "Tri-training: Exploiting Unlabeled Data Using Three Classifiers," *IEEE Transactions on Knowledge and Data Engineering* (17:11), pp. 1529-1541.

Appendix H

Annotation Details and Model Training

When using supervised learning methods for text analytics, the annotation process is incredibly important. Fortunately, the linguistics and discourse analysis communities have developed best practices over the years for speech act labeling and conversation and coherence analysis. These best practices for annotation can be broken down into people, process, and technology.

People

We used two full-time, professional annotators with backgrounds in linguistics. These were not part-time students or individuals hired through an online service. Our industry partners helped fund the positions through their financial contributions to the research project. Coauthors experienced in natural language processing and members of industry social media monitoring teams participated in the candidate screening and interviewing process.

Process

Training is an important component to the annotation process (Kuo and Yin 2011). During a two-month training phase, the annotators learned best practices for annotating conversations, coherence relations, and speech acts from existing literature and standards from the linguistics and discourse analysis community. For instance, speech act annotations were guided by standards laid out in the Dialog Act Markup in Several Layers (DAMSL), developed by the Multiparty Discourse Group. These standards provide concrete prescriptions, including decision trees of annotation rules for how to annotate certain texts. Consequently, they have been used in prior supervised-learning based speech act studies (e.g., Stolcke et al. 2000). Similarly, the coherence relations identification body of knowledge is largely governed by Halliday and Hasan's (1976) seminal text *Cohesion in English*, which provides taxonomies of coherence relations, examples, and identification/classification rules. Over the past 20 years, these rules have been adapted to online discourse through many studies, including several that we cited in the paper. In addition to Halliday and Hasan, conversation identification was guided by the texts from the discourse analysis and pragmatics literature.

During the training phase, the annotators developed and refined guidelines for annotation by examining threads pertaining to the channels and industry contexts employed in our test bed. The guidelines included details on necessary annotation meta-data such as the rationale for the annotation (categorical attribute), reference to specific rule(s) guiding the rationale, and additional notes. Disagreement resolution protocols between annotators incorporate discussion of these annotation notes and meta-data. Additionally, industry experts with domain knowledge and experience analyzing similar types of data in similar contexts were used throughout the annotation process as an additional check. The use of a rigorous process allowed the annotations to be rigorous and consistent, with very high inter-annotator agreement measures (as reported later in this appendix).

Technology

All annotations were performed through a custom software tool developed for this project. The tool allowed annotators to add meta-data and notes, mark/flag items, modify annotations, etc. It also recorded annotation clickstreams as part of logs that derived metrics such as annotation speed and user-system interaction. These summary reports were sent to one of the coauthors on a weekly basis for examination to ensure annotation efforts were consistent and congruent with benchmarked effort levels.

Labeling and Inter-Anotator Agreement

Over an 11-month period, the annotators labeled each test bed message with respect to primitive/non-primitive status, conversation affiliation, reply-to relations, and speech act composition (approximately one data set per month). These annotations formed the gold standard used to evaluate the proposed text analytics systems and comparison methods. Though not shown in the paper, a similar quantity of training data (approximately 25,000 messages) was also labeled during that time period. Accordingly, the annotation process was extensive and rigorous, involving input from domain experts provided by our industry partners and the use of best practices. Initially, the annotators underwent two rounds of training on messages from social media discussions that were not part of the test bed (Kuo and Yin 2011). In each training round, the annotators independently labeled multiple discussion threads, totaling over 500 messages, pertaining to the industries and social media channels employed in the test bed. They then met to discuss and resolve differences. In parallel, the same messages were annotated by a social media analyst from a relevant industry partner firm. Next, the analysts and annotators discussed their annotations and reached a consensus. Such a two-stage discussion approach was utilized because the analysts were employees tasked with monitoring various social media channels on a daily basis, and hence possessed considerable domain knowledge to complement the annotators' linguistic training and discourse analysis expertise. After two rounds of training, the two annotators independently annotated each message in the test bed. They met after every 1,000 messages to resolve disagreements. As a final periodic check, the analysts also annotated approximately 10% of the 1,000 messages per iteration. Table H1 lists the inter-annotator agreements for primitive/non-primitive message status (PM), conversation disentanglement (CD), coherence analysis (CA), and speech acts (SA) for the two training rounds and the test bed. The improvements between training and test bed, as well as the agreement values themselves are on par with prior discourse analysis studies (Kuo and Yin 2011; Twitchell et al. 2012).

Table H1. Inter-Annotator and Annotator-Analyst Agreement for Test Bed (Cohen's Kappa)

	Inter-Annotator Agreement				Annotator-Analyst Agreement			
Stage	PM	CD	CA	SA	PM	CD	CA	SA
Training – Round 1	0.85	0.71	0.74	0.78	0.93	0.82	0.81	0.87
Training – Round 2	0.88	0.79	0.81	0.86	0.95	0.89	0.87	0.95
Test Bed	0.90	0.85	0.87	0.90	0.96	0.92	0.93	0.95

Choice of Speech Act Categories Included

The Stolcke et al. (2000) study, as well as others cited in the paper such as Moldovan et al. (2011), noted that the speech acts proposed by Searle (1969) can be considered a hierarchical taxonomy, with assertives, directives, commissives, expressives, and declaratives being at the top level. Examples of directives (i.e., child nodes/subcategories in the taxonomy) include questions, suggestions, and commands. The presence of different subtypes/categories in the taxonomy largely depends on characteristics of the data set and application domain. Hence, prior studies have often adapted a subset of the taxonomy based on prevalence and key use cases, as deemed appropriate. For instance, Moldovan et al. incorporated special subcategories of directives (questions) and commissives (accept/reject) in their speech act classification of online chat. Similarly, Stolcke et al. incorporated multiple subcategories of questions in their analysis of speech acts in switchboard call transcripts data. Accordingly, in our paper, in addition to commissives, assertives, declarativies, and expressives, we incorporated two sub-categories of directives: questions and suggestions. These were included due to their close connection with our social media use cases (namely identifying issues and suggestions), and prevalence of these types in the various organizational social media data sets examined in the paper.

Model Training Data Set

As noted in the note for Table 5 in the "Evaluation" section of the main document, and earlier in this appendix, a separate set of approximately 25,000 messages was used for training purposes. These messages were completely independent and non-overlapping with the test bed described in the "Evaluation" section and Table 5. LTAS and comparison methods were trained on data from the same domain and channel. Similarly, in the TelCorp field study presented in "Field Experiment" subsection of the main document, all models were trained on data from the same domain and channel. More details about model management and training for the 4-month field study appear in appear in Appendix M.

Following data mining best practices, LTAS parameters were tuned using cross-validation applied on the training set. In order to ensure that all comparison methods employed in experiments 1–3 garnered the best possible results, their parameters were tuned *retrospectively* using a grid (i.e., full combinatorial) search applied on the *test data performance*. For instance, for conversation disentanglement, Wang and Oard (2009)'s method uses a t_{sim} similarity threshold as well as an alpha and three lambda parameters. For each parameter, several different values were tested, resulting in over 3,000 parameter combinations examined during the grid search. For all comparison methods in experiments 1–3, the parameter settings yielding the best results were reported.

References

Halliday, M. A. K., and Hasan, R. 1976. *Cohesion in English*, London: Longman.

Kuo, F. Y., and Yin, C. P. 2011. "A Linguistic Analysis of Group Support Systems Interactions for Uncovering Social Realities of Organizations," *ACM Transactions on MIS* (2:1), Article 3.

Moldovan, C., Rus, V., and Graesser, A. R. 2011. "Automated Speech Act Classification for Online Chat," in *Proceedings of the 22nd Midwest AI and Cognitive Science Conference*, Cincinnati, Ohio.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*, Cambridge, UK: Cambridge University Press.

Stolcke, A., Ries, K., Jurafsky, D., and Meteer, M. 2000. "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistic* (26:3), pp. 339-373.

Twitchell, D., Jensen, M. L., Derrick, D. C., Burgoon, J. K., and Nunamaker Jr., J. F. 2013. "Negotiation Outcome Classification Using Language Features," *Group Decision and Negotiation* (22:1), pp. 135-151.

Wang, L., and Oard, D. 2009. "Context-Based Message Expansion for Disentanglement of Interleaved Text Conversations," in *Proceedings Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Boulder, CO, pp. 200-208.

Appendix I

User Sense-Making Experiment Details

For each industry context, two discussion threads were included in the user experiment. For the telecommunications, health, and security contexts, the two threads were taken from the web forum and social networking data sets in order to demonstrate the user sense-making support utility of the proposed LAP-based system on different types of social media. Tables I1 and I2 provide a brief summary of the threads and questions/tasks for the four industry contexts.

Table I1. Summary of Thread Topics and Social Media Channels in Sense-Making User Experiments

Thread Characteristics	Telecommunications	Health	Security	Manufacturing
Number of Threads	2	2	2	2
Social Media Channels	Web forum, social networking	Web forum, social networking	Web forum, social networking	Chat
Thread Topics	Discussion of recent change to wireless data plan pricing and monthly usage limits	Discussion of side-effects for a specific pain medication	Discussion of a recent update for security software	Discussion of solutions for a tea manufacturer's over-production problem

Table I2. Summary of Types of Tasks/Questions Asked in SenseMmaking User Experiments

Task or Question Type	Use Case(s)	Telecom	Health	Security	Manufacturing
Basic	Identifying issues; identifying ideas and opportunities	List all questions asked in the discussion thread	List all side effects mentioned in the discussion thread	List all questions asked in the discussion thread	List all solutions presented in the discussion thread
Action	Identifying issues; identifying ideas and opportunities	Which questions posed by a particular discussant were answered	Which answer(s) a particular discussant agreed with	Which questions posed by a particular discussant were answered	Which solution(s) a particular discussant supported
Situated action	Identifying issues	Which question caused the greatest confusion in terms of number of diverging answers	Which side-effect resulted in the greatest conflict in terms of dichotomy between agreement and disagreement	Which question caused the greatest confusion in terms of number of diverging answers	Which solution results in the greatest conflict in terms of dichotomy between support and opposition
Symbolic action	Identifying issues; identifying ideas and opportunities	Which discussants seem frustrated about a particular issue	Which discussants seem concerned about a particular side-effect	Which discussants seem concerned about a proposed solution	Which discussants seem enthusiastic about a proposed solution

Appendix J

Survey Items for Field Experiment

The survey items pertaining to system perceived usefulness and ease of use were adapted from Venkatesh et al. (2003), which incorporated some items from Davis (1989), Davis et al. (1989), and Moore and Benbasat (1991). For each construct, we incorporated five items. Each item was on a 1 to 10 scale ranging from strongly disagree to strongly agree. The items are presented in Table J1. In the main paper, for each construct, we present the averages across the items.

Table J1. Field Experiment Survey Items

Construct	Items	Sources
Usefulness of system	1. Using the system enables me to accomplish tasks more quickly.	Davis 1989 Davis et al. 1989 Venkatesh et al. 2003
	2. Using the system improves the quality of the work I do.	
	3. Using the system makes it easier to do my job.	
	4. Using the system enhances my effectiveness on the job.	
	5. Using the system increases my productivity.	
Ease of system use	1. Learning to operate the system is easy for me.	Davis 1989 Moore and Benbasat 1991 Venkatesh et al. 2003
	2. I find it easy to get the system to do what I want it to do.	
	3. My interaction with the system is clear and understandable.	
	4. I find the system to be flexible to interact with.	
	5. I find the system easy to use.	
Usefulness of information for identifying issues	1. Using the system enables me to identify issues more quickly.	Davis 1989 Davis et al. 1989 Venkatesh et al. 2003
	2. Using the system improves the quality of issues I identify.	
	3. Using the system makes it easier to identify issues.	
	4. Using the system enhances my effectiveness at identifying issues.	
	5. Using the system increases my productivity for identifying issues.	
Usefulness of thread browsing capability	1. Using the system enables me to browse threads more quickly.	Davis 1989 Davis et al. 1989 Venkatesh et al. 2003
	2. Using the system improves the quality of threads browsed.	
	3. Using the system makes it easier to browse threads.	
	4. Using the system enhances my effectiveness at browsing threads.	
	5. Using the system increases my productivity for browsing threads.	
Usefulness of participant ranking capability	1. Using the system enables me to rank participants more quickly.	Davis 1989 Davis et al. 1989 Venkatesh et al. 2003
	2. Using the system improves the quality of my participant rankings.	
	3. Using the system makes it easier to rank participants.	
	4. Using the system enhances my effectiveness at ranking participants.	
	5. Using the system increases my productivity for ranking participants.	

Although our *N* was small, with 22 total subjects in the field experiment, we performed exploratory factor analysis and computed Cronbach's alphas as a construct reliability check. The results from subject responses at the 4-month mark appear in Tables J2 and J3. Prior studies such as de Winter et al. (2009) have found factor analysis to be a valid method even with a smaller sample size (i.e., *N* = 24), for 4-8 factors and 24 variables, in situations where the factor loadings are greater than 0.8. The factor loadings in Table J2 suggest the constructs had convergent and discriminant validity. Similarly, the alpha values in table J3 were all above 0.8, which is considered good.

Table J2. Exploratory Factor Analysis of Survey Items

Construct	Items	1	2	3	4	5
Usefulness of system	us1	-0.09	**0.96**	0.18	0.12	0.15
	us2	-0.08	**0.97**	0.14	0.13	0.23
	us3	-0.11	**0.95**	0.15	0.11	0.29
	us4	-0.09	**0.92**	0.20	0.10	0.20
	us5	-0.12	**0.94**	0.22	0.09	0.17
Ease of system use	es1	**0.91**	-0.11	-0.05	-0.04	-0.01
	es2	**0.92**	-0.07	-0.06	-0.02	-0.02
	es3	**0.96**	-0.09	-0.06	-0.03	0.00
	es4	**0.94**	-0.08	-0.03	-0.05	-0.03
	es5	**0.93**	-0.10	-0.04	-0.07	0.01
Usefulness of information for identifying issues	uii1	-0.08	0.10	0.12	**0.94**	0.15
	uii2	-0.10	0.14	0.10	**0.93**	0.21
	uii3	-0.06	0.15	0.07	**0.92**	0.18
	uii4	-0.09	0.24	0.13	**0.94**	0.13
	uii5	-0.07	0.20	0.16	**0.97**	0.19
Usefulness of thread browsing capability	utbc1	-0.03	0.15	0.04	0.08	**0.90**
	utbc2	-0.02	0.10	0.06	0.04	**0.91**
	utbc3	-0.03	0.21	0.03	0.10	**0.92**
	utbc4	-0.01	0.24	0.01	0.12	**0.91**
	utbc5	-0.01	0.19	0.06	0.09	**0.89**
Usefulness of participant ranking capability	uprc1	0.00	0.15	**0.93**	0.18	0.06
	uprc2	-0.02	0.18	**0.92**	0.22	0.05
	uprc3	-0.01	0.16	**0.94**	0.21	0.02
	uprc4	0.00	0.10	**0.91**	0.15	0.10
	uprc5	-0.03	0.17	**0.94**	0.17	0.04
Eigenvalue		6.65	5.78	4.55	3.41	1.67
Cumulative Variance Explained (%)		26.45	49.52	67.80	81.02	87.98

Table J3. Cronbach's Alpha Values for Survey Constructs

Construct	Cronbach's α
Usefulness of system	0.97
Ease of system use	0.85
Usefulness of information for identifying issues	0.93
Usefulness of thread browsing capability	0.86
Usefulness of participant ranking capability	0.94

Longitudinal Perception Results

In the main paper, we reported the analyst perceptions at the 4-month mark to allow users of System B time to get better acquainted with the new system. As noted in the field experiment discussion in the main paper, following prior behavioral IS studies on technology adoption, the surveys were conducted at four points in time: prior to introduction of System B, after one week of training (for System B users), and at the two and four month marks in the field experiment. The "prior to introduction of B" was intended to get everyone's (i.e., all 22 analysts) baseline perceptions regarding System A before System B was ever mentioned to the 10 analysts assigned to the B setting. We examined the perceptions across all four time periods for various usefulness constructs reported in the paper and found that System A's perceptions remained fairly constant while System B's generally started out lower (relative to the System A "prior to introduction of B" baseline) and improved at the 2-month and 4-month marks. Figure J1 depicts this trend in regards to the overall "usefulness of system" construct. This result is consistent with the behavioral IS research on adoption, which has found that user buy-in to "newer is better" is not a given since familiarity with the status-quo and switching costs are often viewed as impediments to adoption of new technologies (motivating some of the research on technology adoption).

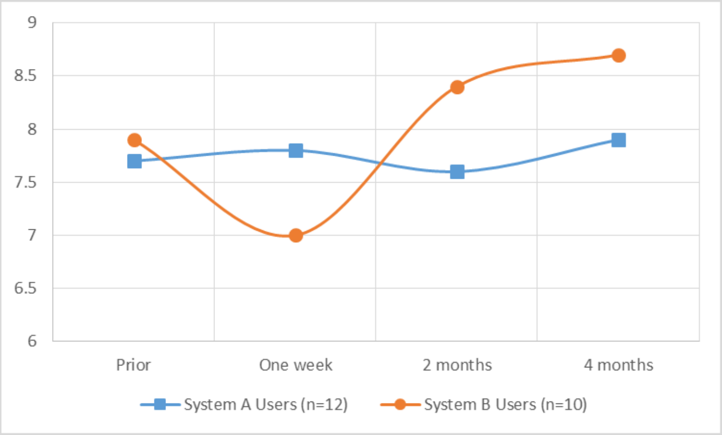


Figure J1. User Perceptions of Overall System Usefulness at Different Points During Field Experiment

References

Davis, F. D. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), pp. 319-340.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. 1989. "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science* (35:8), pp. 982-1003.

de Winter, J. C. F., Dodou, D., and Wieringa, P. A. 2009. "Exploratory Factor Analysis with Small Sample Sizes," *Multivariate Behavioral Research* (44:2), pp. 147-181.

Moore, G. C., and Benbasat, I. 1991. "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation," *Information Systems Research* (2:3), pp. 192-222.

Venkatesh, V., Morris, M., Davis, G. B., and Davis, F. D. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), pp. 425-478.

Appendix K

Contribution of Composite Kernel to Coherence Analysis Performance

In this appendix, we compare the performance of our proposed composite kernel versus a single support vector machine (SVM) classifier. Before evaluating our kernel ensemble, it is important to provide background on kernel-based methods. The objective of our machine learning method is to train a classifier to learn patterns that distinguish positive from negative reply-to relations. Statistical learning theory has prompted the development of highly effective machine learning algorithms for various application domains, including natural language processing, that leverage kernel machines (Vapnik 1999). SVM is a prime example of a kernel-based method (Cristianini and Shawe-Taylor 2000). Kernel machines owe their name to the use of kernel functions which are able to leverage the "kernel trick": the ability to operate in a feature space without explicitly computing its coordinates, by instead computing the similarity between pairs of data points in the feature space (Burges 1998; Muller et al. 2001; Vapnik 1999). This allows kernel-based methods to be highly scalable and robust (Cristianini and Shawe-Taylor 2000), important characteristics for natural language processing such as coherence analysis. Given a number of positive and negative coherence relation instances, the kernel machine would enable the use of a kernel function to compute the similarity between these instances.

Formally, given an input space $\boldsymbol{U}$, in this case the set of all possible reply-to relation pair instances to be examined, the learning problem can be formulated as finding a classifier C: $\boldsymbol{U} \rightarrow \boldsymbol{V}$ where $\boldsymbol{V}$ is a set of possible labels (in this case "reply-to" or "no reply-to") to be assigned to the data points. Finding C relies on a kernel function K that defines a mapping K: $\boldsymbol{U} \times \boldsymbol{U} \rightarrow [0, \infty)$ from the input space $\boldsymbol{U}$ to a similarity score $K(u_i, u_j) = f(u_i) \cdot f(u_j)$ where u_i and u_j represent two data points in $\boldsymbol{U}$, in this case two different message pair instance vectors; $f(u_i)$ is a function that maps $\boldsymbol{U}$ to a higher dimensional space (called a hyperplane) without needing to know its explicit representation. As previously alluded to, this part is often referred to as the "kernel trick" (Cristianini and Shawe-Taylor 2000). It is important to reiterate that here, each instance in the input feature matrix (e.g., u_i) is already a coherence relation pairing between two messages (e.g., reply-to or no reply-to), not an individual message. Hence, in line with our objective or learning patterns that can differentiate positive from negative reply-to relations, the similarity scores $K(u_i, u_j)$ in our case are meant to enable us to create a mapping in some hyperplane that can allow us to separate between positive and negative reply-to pair instances in an accurate and robust manner.

Searching for an optimal C involves evaluating different parameters, where α denotes a specific choice of parameter values for the function $f(u, \alpha)$. These parameters are analogous to the weights and biases incorporated within a trained neural network classifier (Burges 1998). For SVMs, many algorithms have been developed for finding an optimal C, which essentially entails solving a quadratic programming problem in order to create a hyperplane that maximizes the linear separation between instances belonging to the two different classes (often called the "maximum margin" principle).

As mentioned in the main paper, the beauty of kernel-based methods lies in the ability to define a custom kernel function K tailored to a given problem, or to use the standard predefined kernels (e.g., linear, polynomial, radial basis function, sigmoid, etc.). When dealing with classification tasks involving diverse patterns, composite kernels are well-suited to incorporate broad relevant features while reducing the risk of over-fitting (Collins and Duffy 2002; Szafranski et al. 2010). In our case, diversity stems from differences in the occurrence of system, linguistic, and conversation structure features across users, social media channels, and/or industries.

In order to illustrate the efficacy of our composite kernel, we compared its performance against a single SVM classifier on the 10 data sets incorporated in our test bed. The results appear in Table K1. On average, the composite kernel outperformed the single SVM by about 7 percentage points in terms of precision, recall, and f-measure.

Table K1. Comparison of Composite SVM Kernel Versus Single SVM

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Method	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Composite Kernel*	**77.0**	**85.7**	**81.1**	**80.4**	**95.3**	**87.2**	**87.3**	**95.0**	**91.0**
Single SVM	71.2	75.8	73.4	73.2	90.7	81.1	80.9	89.0	84.8
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Method	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Composite Kernel*	**72.3**	**86.4**	**78.7**	**71.8**	**90.6**	**80.1**	**84.3**	88.5	**86.4**
Single SVM	66.3	76.7	71.1	67.9	82.8	74.6	75.8	83.9	79.6
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Method	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
Composite Kernel*	**77.6**	**84.8**	**81.0**	**79.5**	**88.3**	**83.7**	**90.1**	**94.9**	**92.5**
Single SVM	71.6	75.2	73.3	71.4	82.5	76.5	75.8	83.9	79.6
Manufacturing									
	Chat								
Method	Prec.	Rec.	F-Meas						
Composite Kernel*	**79.4**	**91.0**	**84.8**						
Single SVM	71.7	81.2	76.2						

*Significantly outperformed comparison method, with all p-values < 0.001.

References

Burges, C. J. C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery* (2:2), pp. 121-167.

Collins, M. and Duffy, N. 2002. "Convolution Kernels for Natural Language," in *Advances in Neural Information Processing Systems* (Volume 14), T. G. Diettrich, S. Becker, and Z. Ghahramani (eds.), Cambridge, MA: MIT Press, pp. 625-632.

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge, UK: Cambridge University Press.

Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. 2001. "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks* (12:2), pp. 181-201

Szafranski, M., Grandvalet, Y., and Rakotomamonjy, A. 2010. "Composite Kernel Learning," *Machine Learning* (79:1-2), pp. 73-103.

Vapnik, V. 1999. *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.

Appendix L

Impact of WordNet Versus LDA-Based Term Similarity Assessment on Primitive Message Detection

As noted in the paper, the primitive message detection (PMD) method in LTAS uses WordNet to compute similarity between terms. PMD serves as important input for the conversation disentanglement component. In order to empirically examine the effectiveness of WordNet-based similarity, we compared its performance for PMD, and ultimately for conversation disentanglement, against two comparison Latent Dirichlet Allocation (LDA) methods. The first comparison method was standard LDA (Blei et al. 2003). Given a set of documents, it outputs groups of terms, where each group is said to belong to a topic. In LDA, a term may belong to more than one topic/group. Following the approach taken in Zhai et al. (2011), we computed term-similarity by leveraging terms' probabilities across topics. The second comparison method was the use of a Dirichlet Forest prior in a Latent Dirichlet Allocation framework (DF-LDA; Andrzejewski et al. 2009). Consistent with the approach taken with benchmark methods evaluated in experiments 1–3 in the main document, in order to ensure that LDA and DF-LDA garnered the best possible results, their parameters were tuned *retrospectively* using a grid (i.e., full combinatorial) search applied on the *test data performance*. For instance, LDA uses a number of hidden topics *K*, and alpha and beta prior topic distribution/sparsity parameters. For each parameter, several different values were tested, resulting in over 1,000 parameter combinations examined during the grid search. The parameter settings yielding the best results were reported in Tables L1 and L2.

Table L1. Results for Primitive Message Detection Using WordNet Versus LDA Methods

Telecom									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**
WordNet	**60.7**	**74.7**	**67.0**	**68.2**	**93.8**	**79.0**	**62.4**	**94.3**	**75.1**
DF-LDA	58.5	74.1	65.4	67.9	93.8	78.7	62.1	94.3	74.9
LDA	57.4	72.7	64.2	67.6	93.4	78.4	60.7	91.5	73.0
Health									
	Web Forum			**Social Network (Patients)**			**Microblog (Twitter)**		
Technique	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**
WordNet	63.3	69.6	66.3	**57.6**	**89.2**	**70.0**	**63.6**	**98.6**	**77.3**
DF-LDA	**63.8**	**70.1**	**66.8**	56.6	89.2	69.2	63.3	98.5	77.1
LDA	62.0	69.6	65.6	56.6	87.8	68.8	62.8	98.4	76.7
Security									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**
WordNet	**71.6**	**84.4**	**77.5**	62.0	87.6	72.6	**59.9**	**95.5**	**73.6**
DF-LDA	70.1	84.4	76.6	**62.5**	**88.2**	**73.1**	59.4	95.4	73.3
LDA	68.4	83.4	75.1	**60.4**	**87.4**	**71.4**	**58.3**	**92.8**	**71.6**
Manufacturing									
	Chat								
Technique	**Prec.**	**Rec.**	**F-Meas**						
WordNet	**66.9**	**70.6**	**68.7**						
DF-LDA	66.7	69.4	68.5						
LDA	65.5	69.9	67.6						

Table L1 shows the PMD results for WordNet versus LDA and DF-LDA on the 10 data sets in our test bed. WordNet outperformed LDA on all 10 data set sin terms of precision, recall, and f-measure, though the performance deltas were generally small, with an difference in f-measures of about 1.5 percentage points. This result is consistent with some prior studies (e.g., Zhai et al. 2011), where basic LDA has underperformed against WordNet. WordNet also outperformed DF-LDA on 8 out of 10 data sets, but with an average f-measure difference of only 0.3

percentage points. Similarly, on the two data sets where DF-LDA did outperform WordNet, the f-measure improvements were only half a percentage point. By incorporating Dirichlet priors into the LDA process, DF-LDA is better suited for learning domain-specific similarities compared to LDA (Andrzejewski et al. 2009). Furthermore, we examined the impact of the WordNet and two comparison LDA-based methods on conversation disentanglement (where the primitive message information serves as an important input). The results of that comparison appear in Table L2. As expected, given the PMD experiment results, using WordNet-based PMD resulted in conversation disentanglement f-measures that were about 0.2 percentage points better than DF-LDA on average, and 0.7 points better than LDA.

Overall, the results suggest that the WordNet-based method is well-suited for term-similarity assessment in our data sets, relative to the LDA-based techniques examined. Additionally, we believe future research exploring methods that combine WordNet with LDA to balance lexicons with domain-specific learned similarities may constitute a worthwhile direction.

Table L2. Results for Conversation Disentanglement Using PMD with WordNet Versus LDA Methods

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
WordNet	**68.7**	**72.5**	**70.6**	**75.7**	**95.0**	**84.2**	**79.9**	**99.2**	**88.5**
DF-LDA	68.0	72.2	70.1	75.2	94.9	83.9	79.5	99.2	88.3
LDA	67.5	71.3	69.4	74.9	94.8	83.7	78.9	98.1	87.4
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
WordNet	**63.6**	**75.4**	**69.0**	**66.4**	**80.1**	**72.6**	**77.4**	**99.4**	**87.0**
DF-LDA	63.7	75.5	69.1	66.0	80.1	72.4	76.9	99.3	86.7
LDA	63.3	75.0	68.7	65.8	79.7	72.1	76.7	99.3	86.5
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
WordNet	**69.7**	**75.6**	**72.5**	**76.8**	**80.5**	**78.6**	**82.5**	**99.6**	**90.3**
DF-LDA	69.3	75.2	72.1	77.0	80.6	78.8	82.3	98.2	89.5
LDA	68.7	74.6	71.5	76.6	80.2	78.4	81.6	97.6	88.9
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
WordNet	**64.0**	**72.7**	**68.0**						
DF-LDA	63.8	72.0	67.6						
LDA	63.5	69.6	67.3						

References

Andrzejewski, D., Zhu, X., and Craven, M. 2009. "Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors," in *Proceedings of the 26th ACM Annual International Conference on Machine Learning*, New York: ACM Press, pp. 25-32.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993-1022.

Zhai, Z., Liu, B., Xu, H., and Jia, P. 2011. "Clustering Product Features for Opinion Mining," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, New York: ACM Press, pp. 347-354.

Appendix M

Overview of TelCorp Social Media Monitoring Workflow for Field Experiment

In this appendix we offer a brief description of TelCorp's workflow pertaining to social media monitoring (depicted in Figure M1). TelCorp monitors over two dozen online channels including various social networking platforms, blogs, forums, and chat rooms. During the four-month field experiment, over 5.2 million new messages associated with 464,000 threads were examined by the analysis systems (i.e., on average, slightly over 43,000 messages per day, or about 1,806 per hour). During peak message volume periods, more than 5000 messages per hour were received. (i.e., over 83 per minute).

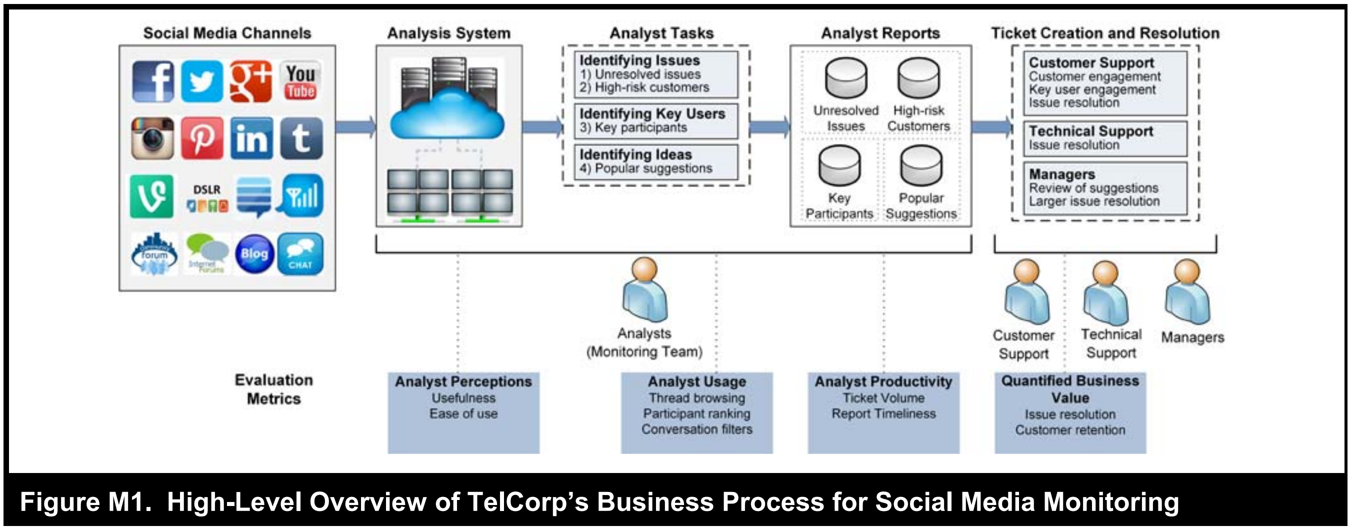


Figure M1. High-Level Overview of TelCorp's Business Process for Social Media Monitoring

A/B testing is a commonly used method to concurrently examine the performance of alternative artifacts or design settings. The key outputs of LTAS are conversation affiliations, coherence relations, and message speech acts. Treating the existing system used by TelCorp as setting A, we worked with the TelCorp's IT folks to develop setting B. For the B system setting, LTAS was embedded into their real-time analysis pipeline (Figure M1), adding conversation affiliation, reply-to relation, and speech act labels to all messages. Furthermore, participant importance rankings were now computed using these revised social network analysis metrics. In the custom dashboards, sequential ordering was complemented with an SATree option and conversation and speech acts were added as additional filters/dimensions for search, browsing, and visualization.

TelCorp's existing analysis system (i.e., System A in the field experiment), encompassing text analytics servers, computing instances, storage, and application servers, all run in the cloud. This is important to enable elastic compute since incoming social media message volume is most certainly not uniformly distributed across 24 hours a day, 7 days a week. The System B leveraging LTAS information was also deployed in the cloud. Four sets of models were trained prior to the field experiment; one for forums, one for micro-blogs (e.g., Twitter), one for social networking sites and blogs (e.g., Facebook, Google+, YouTube, Tumblr, etc.), and one for chat (e.g., Live Chat). These four sets of models covered incoming messages/threads from each of the 24 channels. The training data was not updated at all during the 4-month field experiment. No productivity or business value drops were observed longitudinally with System B in that time period. However, consistent with model management practices adopted regarding other forms of analytics at TelCorp, we suspect periodic model management and updating would be necessary to keep pace with TelCorp's evolving product/service offerings and outreach, changing customer experiences, and as novel new classes of issues emerge.

During the field experiment, TelCorp felt it was important to keep System B's average message processing times within acceptable levels, since identifying potential issues in a timely manner is a key metric. For System B, every time a new message was received, the conversation disentanglement, coherence analysis, and speech act classification modules from LTAS were applied to the entire discussion thread. Table M1 provides the mean processing times per new message for the three main components of LTAS (this includes the total time for processing the entire thread related to the new message). On average, LTAS processed each new message in about 1.5 seconds (with over 99% of messages processed within 3 seconds). As previously noted, during peak message volume periods, typically three or four additional cloud servers were used to ensure that the average message processing times for System B were comparable to those of System A. The additional cloud computing costs for System B were factored into the business value assessment (discussed in a subsequent paragraph).

Table M1. Mean Processing Times per Message During 4-Month Field Experiment

Module	Processing Time (in milliseconds)
Conversation disentanglement	425 (103)
Coherence analysis	304 (136)
Speech act classification	829 (342)
Total for LTAS Components	**1558 (507)**

As previously alluded to in the main paper, TelCorp's monitoring team focused on three key social media monitoring tasks: identifying issues, identifying key users, and identifying suggestions. Identifying issues encompasses (1) unresolved issues and (2) high-risk customers. TelCorp defines unresolved issues as events that adversely impact a set of customers. An example of an unresolved issues that arose during the 4-month field experiment is an error in the billing system which caused customers in three U.S. states to receive excess charges on their monthly statements. Another example is a technical issue with a new integrated router-plus-modem's installation software which caused tens of thousands of customers to experience random Internet outages. It is important to note that TelCorp monitoring analysts generate a separate report instance for each customer impacted by an unresolved issue. For instance, if analysts identify 5,000 customers discussing the billing system error on social media, they would generate 5,000 reports since the expectation is that customer support reps should follow-up individually with many/most of them. High-risk customers are customers that may possibly churn due to what TelCorp considers standard operational issues. Examples include an individual upset about call center wait times, or a customer considering switching to another carrier due to price differences.

While issue identification is the primary use case for TelCorp's monitoring team, they also look to identify key discussion participants based on social network centrality; these include key positive/negative influencers, brand advocates, etc. Additionally, analysts in the monitoring team seek to identify popular suggestions. Examples include ideas about fund-raising events, charities valued by existing and prospective customers, requests for new product and/or service offerings, and suggestions on how to enhance the customer web portal and mobile app. For suggestions reported by the monitoring teams, TelCorp's managers only create tickets for new, unique suggestions.

Analyst submitted reports, with each report including a description, severity level (mostly used for issues), and associated social media discussants, conversations, and/or threads. These reports were routed to customer support representatives, technical support, and/or managers. For a subset of reports, tickets are created indicating cases requiring action. Customer support reps attempt to engage with high-risk customers with the goal of reducing attrition. They also reach out to key users in order to preemptively garner brand advocacy or mitigate negative influence. Customer support reps also reach out to customers impacted by unresolved issues. Tech support reps work to resolve technical issues. Managers review suggestions and may also be involved in resolution of larger issues.

As depicted in Figure M1, four sets of evaluation metrics were used to examine the effectiveness of System B relative to System A. The behavioral IS research has extensively examined the importance of user perceptions (i.e., usefulness and ease of use) as key antecedents for actual system usage. The main paper describes how analyst perceptions were captured longitudinally at the beginning, and then after one week, two months, and four months. Similarly, the main paper discusses how usage of various key system features was measured.

Ultimately tangible value results from observed increases in productivity leading to quantifiable business value. As mentioned earlier in this appendix, and stated in the main paper, during the 4-month experiment, Systems A and B were run in parallel using non-overlapping teams. Reports generated by users' of each system were tracked, resulting in two sets of reports. The Venn diagram in Figure M2 illustrates these two sets. The first of the two productivity measures incorporated by TelCorp was *timeliness* of overlapping reports created by users of both systems. This was the time between once a report was generated and when the data first entered the system, measured in minutes. The timeliness delta between report submission timestamps within $A \cap B$ is an important measure of how quickly analysts can identify items of interest. The second productivity measure was *ticket volume*. Only reports deemed most important are converted to tickets by the customer/technical support reps or managers. For TelCorp, the number of generated tickets attributable to reports submitted by users of System A versus System B constitutes an important productivity measure. If we treat the tickets generated by Systems A and B as two partially overlapping sets tA and tB, the key ticket volume measures are the total number of generate tickets attributable to System A and B's reports (|tA| and |tB|), and the unique/non-overlapping tickets generated by each system, which is the cardinality of their ticket complements: $|tA \cap tB^c|$ and $|tA^c \cap tB|$. For the productivity assessments, in the main paper we focus on unresolved issues and high-risk customers (although System B also garnered higher report/ticket volumes for identification of key participants and suggestions). For the field experiment, TelCorp elected not to quantify the monetary value attributed to identifying key participants or suggestions; however, they did mention the value proposition of system B in regards to these key productivity measures (discussed later in the appendix).

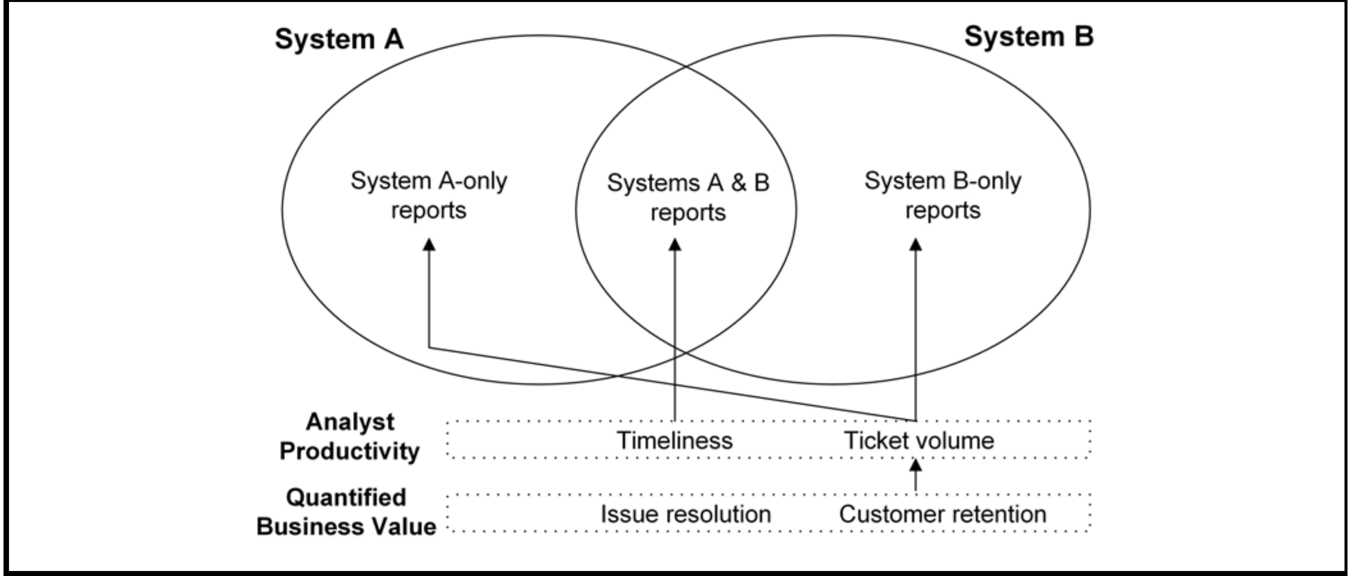


Figure M2. Method for Productivity and Business Value Assessment of Systems A and B

Business value stems from *better* identifying issues, key participants, and ideas in a *timelier* manner. For the field experiment, TelCorp chose to quantify business value primarily in terms of identified issues, including the value of resolving issues on customer churn reduction (i.e., for those customers impacted by the issue), and successfully engaging and retaining high-risk customers. This quantification focused on computing the monetary value of a ticket, and was performed as follows:

- For each *unidentified* high-risk or unresolved issue customer (i.e., ones for which TelCorp failed to generate reports/tickets), TelCorp had derived an estimated customer value (ECV), where ECV = individual customer's one-year mean revenue * mean expected % of year retained.

- For each ticket in the 4-month field experiment, TelCorp was also able to monitor customer churn over the 12-months since the field experiment to compute actual customer value (ACV), where ACV was the sum of the actual 12-month revenue for each ticketed customer that TelCorp sales/tech support reps and/or managers followed-up with.

- The quantified business value for a system was then ACV – (ECV * ticket volume). For system B, the additional cloud computing costs attributable to LTAS were also subtracted from this value.

TelCorp did not provide us with quantifications of the monetary value attributed to identifying key participants or suggestions, although the number of generated reports pertaining to these two use cases was also higher in System B (as presented in Table 14 of the main paper). They also chose not to quantify the value of timelier detection. For obvious reasons, although both systems were allowed to submit a report regarding the same customer or issue, tickets were only generated for one instance (i.e., the earlier received report). This made it difficult to quantify the precise monetary value of the timelier receipt within $A \cap B$: for instance, how much higher was the customer retention rate resulting from the customer service reps' engagement efforts because they were able to reach out to the customer one hour earlier? Although most certainly valuable, the experiment design was less conducive to properly quantifying what would have happened if they had waited longer. Additionally, TelCorp did not experience any major unresolved issues during the 4-month field experiment such as the Fall 2012 premium customer upgrade debacle which cost them an estimated $110 million over a 54-hour period. Hence, TelCorp believes the actual long-term business value of System B may be even higher than what they quantified for the purpose of the 4-month field experiment.

Table M2 includes sample quotes from various employees at TelCorp, including members of the monitoring team, a customer support representative, managers, and the VP for Digital Operations. The quotes, which were captured after the 4-month field experiment, relate to various facets of System B, including the system as a whole, the thread/conversation browsing capability, as well as the system's ability to support issue identification, participant ranking, and identification of suggestions. In the quotes, square brackets indicate insertions/modifications made to preserve anonymity.

Table M2. Sample TelCorp Employee Quotes Related to System B Incorporating LTAS Information

Category	Sample Quotes
System as a Whole	“*The system has been amazing. For the first time in my 5 years here, I don’t feel overwhelmed…We really understand what’s going on, as its happening....It no longer feels like we’re constantly swimming upstream.*” [Analyst #1 on Monitoring Team] “*It took me a few weeks to figure out how to do things in the new environment, but now I can’t imagine life without it....I have a friend that works in a similar role at [a major competitor] and she shook her head in disbelief when I told her what we can do.*” [Analyst #2 on Monitoring Team] “*I don’t work with the system directly as much, but I’ve noticed an uptick in the quality of [reports] produced by [the monitoring team]…we seem to be generating [tickets] for the important stuff, faster.*” [Customer Support Representative #1] “*We are much more diligent and effective across the board…[the system] has made the entire process more efficient and valuable.*” [Manager #1] “*As we shift from being an infrastructure company to one focused on providing premium customer experiences, this project is a microcosm of how data analytics can help us get to where we want to go. We’ve taken an important step towards better understanding voice of the customer.*” [VP, Digital Operations]
Thread or Conversation Browsing	“*The ability to peruse online chatter as actual discussions instead of streams of babble that we used to have to piece together ourselves has been huge.*” [Analyst #2 on Monitoring Team] “*For me it’s all about context. The sooner we can figure out why someone is saying what they’re saying, the better....Viewing threads or messages as conversations has helped us to not miss the forest for the trees.*” [Analyst #3 on Monitoring Team]
Identifying Issues	“*Identifying issues has always been a high-stakes, high-stress aspect of my job. Analyzing threads and messages based on action tags and conversations has allowed us to better detect all sorts of issues such as orphaned questions, [at risk customers], and [matters requiring attention].*” [Analyst #2 on Monitoring Team] “*I can find and get to the crux of the issues more efficiently and faster.*” [Analyst #4 on Monitoring Team] “*It’s been a game-changer for us. Now we’re really tapping into these [online channels] to unearth problems and fix them fast. We have higher satisfaction and retention at lower costs.*” [Manager #2]
Ranking Participants	“*To be perfectly honest, in the past I was so busy trying to look for smoke—to put out fires before they got started —that reporting [key online participants] was hardly on my radar. However, all that has changed now with our ability to identify them more easily and effectively.*” [Analyst #1 on Monitoring Team] “*The network metrics and charts are a pretty cool way to quickly determine which [online community members] are most visible within a given conversation, thread, or channel.*” [Analyst #4 on Monitoring Team]
Identifying Suggestions	“*It’s like someone woke up one day and said, ‘what if we added an easy button?’…Being able to view all the [suggestions] being made with a few clicks is one of my favorite features.*” [Analyst #5 on Monitoring Team] “*The number of quality ideas we’ve uncovered through [new system] has been remarkable. In the past three months alone, online suggestions have spawned a new YouTube campaign, two public service announcements, and a successful charity event.*” [Manager #1]

*Square brackets indicate our insertions into the employees’ quotes.

Appendix N

Detailed Experiment Results for Conversation Disentanglement, Coherence Analysis, and Social Network Analysis

Conversation Disentanglement

Table N1 presents the precision, recall, and f-measure details for LTAS and the comparison methods. LTAS attained markedly better precision, recall, and f-measures values (typically 15%–20% higher). The high recall rates suggest that it was able to identify more of the conversations appearing in the discussion threads than other methods, whereas the high accompanying precision rates are indicative of accurate assignment of messages to their respective conversations. While certain comparison methods also yielded relatively high recall rates on the Twitter data sets, these methods had markedly lower precision. Furthermore, they did not perform as well on the social networking, web forum, and chat data sets.

Table N1. Detailed Results for Conversation Disentanglement Experiment on Various Channels

Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS*	**68.7**	**72.5**	**70.6**	**75.7**	**95.0**	**84.2**	**79.9**	**99.2**	**88.5**
Elsner & Charniak	47.0	44.9	45.9	55.2	72.3	62.6	65.9	83.3	73.6
Adams & Martell	43.2	55.1	48.4	56.8	67.3	61.6	57.0	73.6	64.2
Shen et al.	39.6	35.4	37.3	51.7	67.7	58.7	56.0	69.0	61.8
Choi	28.1	25.6	26.8	47.0	57.9	51.9	49.1	59.1	53.7
Wang & Oard	31.1	30.7	30.9	37.9	42.9	40.3	42.6	49.5	45.8
Health									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS*	**63.6**	**75.4**	**69.0**	**66.4**	**80.1**	**72.6**	**77.4**	**99.4**	**87.0**
Elsner & Charniak	45.9	52.2	48.8	54.3	66.9	59.9	66.8	95.4	78.6
Adams & Martell	38.5	52.2	44.3	44.7	61.8	51.9	58.2	82.1	68.1
Shen et al.	38.7	42.7	40.6	52.5	67.2	58.9	57.3	75.7	65.2
Choi	26.1	22.9	24.4	51.9	62.2	56.6	46.4	60.4	52.5
Wang & Oard	29.8	28.0	28.9	53.7	67.4	59.8	39.0	48.3	43.1
Security									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas	Prec.	Rec.	F-Meas
LTAS*	**69.7**	**75.6**	**72.5**	**76.8**	**80.5**	**78.6**	**82.5**	**99.6**	**90.3**
Elsner & Charniak	47.2	44.8	46.0	55.1	64.0	59.2	64.7	82.9	72.7
Adams & Martell	43.2	54.7	48.3	54.6	59.0	56.7	56.3	73.5	63.7
Shen et al.	39.5	34.9	37.1	51.1	59.4	55.0	57.3	75.7	65.2
Choi	27.5	25.2	26.3	48.3	54.2	51.1	46.4	60.4	52.5
Wang & Oard	30.5	30.3	30.4	41.2	44.2	42.6	39.0	48.3	43.1
Manufacturing									
	Chat								
Technique	Prec.	Rec.	F-Meas						
LTAS*	**64.0**	**72.7**	**68.0**						
Elsner & Charniak	39.0	36.5	37.7						
Adams & Martell	39.5	51.1	44.6						
Shen et al.	30.9	27.1	28.9						
Choi	26.2	22.6	24.3						
Wang & Oard	33.5	32.5	33.0						

*Significantly outperformed comparison methods, with all p-values < 0.001.

Coherence Analysis

The f-measure results were discussed in the main paper. As shown in Table N2, LTAS also attained higher precision and recall on 9 of the 10 data sets. On the health tweets data, it also attained higher precision and f-measure than all comparison methods, though the classification method had slightly higher recall.

Table N2. Detailed Results for Coherence Analysis Technique Comparison Experiment

Telecom									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**
LTAS*	**77.0**	**85.7**	**81.1**	**80.4**	**95.3**	**87.2**	**87.3**	**95.0**	**91.0**
Heuristic	58.1	60.0	59.0	51.8	51.1	51.5	69.7	73.5	71.6
Classification	56.1	60.0	58.0	55.2	59.7	57.4	74.0	84.3	78.8
Linkage-Previous	40.1	37.8	38.9	43.2	46.1	44.6	63.2	81.3	71.1
Linkage-First	35.1	36.6	35.9	31.7	33.5	32.6	47.8	57.5	52.2
Health									
Technique	**Web Forum**			**Social Network (Patients)**			**Microblog (Twitter)**		
	Prec.	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**
LTAS*	**72.3**	**86.4**	**78.7**	**71.8**	**90.6**	**80.1**	**84.3**	88.5	**86.4**
Heuristic	49.2	55.6	52.2	49.6	57.9	53.4	69.6	78.4	73.8
Classification	46.9	55.6	50.9	51.2	63.8	56.8	74.3	**90.5**	81.6
Linkage-Previous	33.4	32.8	33.1	37.1	39.4	38.2	59.3	86.2	70.3
Linkage-First	26.1	26.4	26.2	30.5	33.6	32.0	53.7	73.0	61.9
Security									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**	**Prec.**	**Rec.**	**F-Meas**
LTAS*	**77.6**	**84.8**	**81.0**	**79.5**	**88.3**	**83.7**	**90.1**	**94.9**	**92.5**
Heuristic	54.5	54.3	54.4	59.0	60.3	59.7	73.4	75.7	74.5
Classification	52.4	49.2	50.7	62.4	68.7	65.4	76.1	80.8	78.4
Linkage-Previous	33.4	27.1	29.9	50.9	57.2	53.9	61.5	78.5	69.0
Linkage-First	28.5	26.0	27.2	39.6	44.9	42.1	48.1	54.9	51.3
Manufacturing									
	Chat								
Technique	**Prec.**	**Rec.**	**F-Meas**						
LTAS*	**79.4**	**91.0**	**84.8**						
Heuristic	54.8	57.5	56.1						
Classification	45.2	42.0	43.5						
Linkage-Previous	25.3	19.0	21.7						
Linkage-First	15.6	12.1	13.7						

*Significantly outperformed comparison methods, with all p-values < 0.001.

Speech Act Classification

Figure N1 depicts the class-level recall values for LTAS and the two best comparison methods (Joint Classification and Collective Classification) on four of the highly prominent speech acts: assertive, suggestion, question, and commissive. LTAS's Labeled Tree kernel consistently outperformed both comparison methods for all speech acts across the ten data sets, with class-level recall rates of 86.5% to 98.8%.

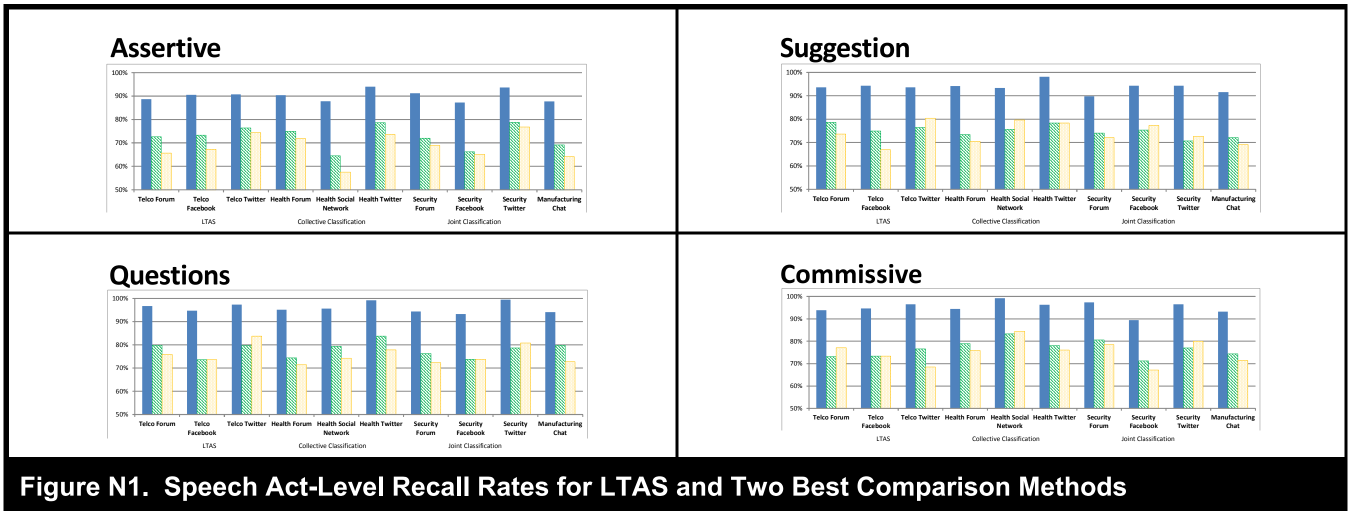


Figure N1. Speech Act-Level Recall Rates for LTAS and Two Best Comparison Methods

Social Network Centrality Measures

Table N3 shows the experiment results for degree, closeness, and betweenness centrality. LTAS had the smallest mean absolute percentage errors across all three metrics, for all data sets in the test bed.

Whereas mean absolute percentage error (MAPE) measures the error percentages relative to gold standard values, examination of differences in rankings is also important since it shed light on how centrality errors could impact assessments of “key participants.” Table N4 shows the Spearman’s rank correlation results for degree, closeness, and betweenness centrality. LTAS had the highest correlations across all three metrics, for all data sets in the test bed. The performance gains were most pronounced on closeness and betweenness centrality. However even for degree centrality, LTAS had rank correlations of 98% or better, which were markedly higher than comparison methods. The results confirm that the coherence analysis module of LTAS enables generation of social networks that are more accurate with respect to percentage error and rank order.

Table N3. Detailed Mean Absolute Percentage Error Results for Social Network Centrality Measures

Telecom

	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**
LTAS*	**4.9**	**4.8**	**17.8**	**4.3**	**2.4**	**12.0**	**2.6**	**1.9**	**9.7**
Heuristic	15.2	22.9	42.2	14.0	20.0	37.9	13.7	19.8	37.3
Classification	18.3	29.0	53.3	15.9	25.4	46.5	14.9	20.2	40.2
Linkage-Previous	25.2	24.2	53.7	29.9	32.4	68.1	23.9	24.9	53.1
Linkage-First	37.0	36.8	64.2	34.8	33.2	59.2	35.8	37.0	63.3

Health

	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Technique	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**
LTAS*	**6.1**	**5.3**	**21.6**	**6.2**	**3.6**	**20.1**	**3.3**	**4.4**	**13.5**
Heuristic	17.2	23.4	46.9	17.1	22.2	45.7	10.3	11.4	25.7
Classification	18.0	21.7	47.7	16.5	17.7	41.9	8.7	4.6	17.8
Linkage-Previous	27.8	29.2	60.8	26.2	26.3	56.2	16.9	6.0	27.0
Linkage-First	37.9	35.1	65.9	35.6	31.6	60.7	23.7	12.9	34.0

Security

	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**
LTAS*	**4.7**	**5.5**	**17.6**	**4.3**	**4.3**	**15.2**	**2.1**	**1.5**	**6.6**
Heuristic	15.2	22.9	42.2	13.7	19.8	37.3	8.9	12.1	23.6
Classification	15.9	25.4	46.5	12.5	15.7	32.7	8.0	9.6	20.4
Linkage-Previous	26.6	29.2	60.9	19.6	17.1	40.1	14.7	7.5	25.1
Linkage-First	42.2	43.9	75.1	30.2	27.6	50.4	26.1	21.3	41.6

Manufacturing

	Chat		
Technique	**Degr.**	**Close.**	**Betwe.**
LTAS*	**7.9**	**4.8**	**18.4**
Heuristic	16.9	13.7	29.7
Classification	17.1	25.6	32.4
Linkage-Previous	41.3	45.6	37.5
Linkage-First	55.7	50.8	50.4

*Significantly outperformed comparison methods, with all p-values < 0.001.

Table N4. Detailed Spearman's Rank Correlation Results for Social Network Centrality Measures

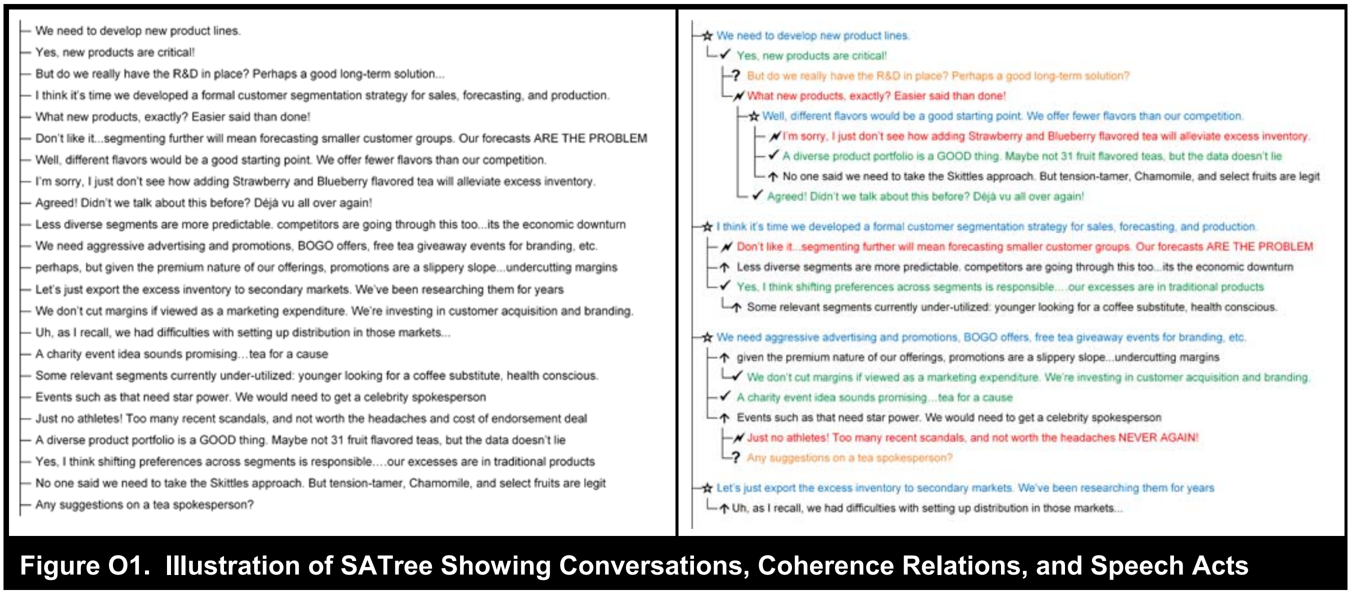
Telecom									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**
LTAS*	**99.0**	**99.1**	**83.4**	**99.3**	**99.8**	**94.5**	**99.7**	**99.9**	**96.5**
Heuristic	90.4	74.1	36.7	91.2	76.0	35.5	93.0	81.5	40.5
Classification	87.5	54.4	31.4	88.8	66.1	34.6	90.9	78.9	38.8
Linkage-Previous	54.9	67.3	29.5	56.9	53.8	24.4	64.3	62.4	27.4
Linkage-First	36.5	49.6	21.8	44.3	48.6	28.3	45.4	48.2	21.5
Health									
	Web Forum			**Social Network (Patients)**			**Microblog (Twitter)**		
Technique	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**
LTAS*	**98.7**	**99.0**	**71.2**	**98.5**	**99.5**	**75.2**	**99.6**	**99.2**	**92.5**
Heuristic	87.3	67.6	29.9	86.7	73.9	26.4	95.7	94.6	59.2
Classification	85.4	76.1	35.0	88.2	85.9	40.9	97.3	99.2	84.1
Linkage-Previous	57.7	56.6	28.3	59.1	66.4	26.0	86.9	98.5	53.3
Linkage-First	32.3	39.7	21.7	40.5	52.2	31.8	68.8	92.7	39.5
Security									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**	**Degr.**	**Close.**	**Betwe.**
LTAS*	**99.2**	**98.8**	**85.9**	**99.3**	**99.2**	**88.6**	**99.8**	**99.9**	**98.1**
Heuristic	90.1	73.5	38.9	91.8	79.3	41.9	97.1	95.0	65.0
Classification	89.0	67.0	25.3	93.5	88.4	49.1	97.4	96.7	78.2
Linkage-Previous	62.3	46.6	24.2	81.3	85.8	33.9	90.4	98.3	68.4
Linkage-First	41.1	42.9	22.8	55.0	51.8	33.1	62.1	75.4	37.1
Manufacturing									
	Chat								
Technique	**Degr.**	**Close.**	**Betwe.**						
LTAS*	**97.1**	**98.7**	**88.5**						
Heuristic	87.3	88.6	47.8						
Classification	87.6	54.5	64.6						
Linkage-Previous	33.2	22.9	31.5						
Linkage-First	15.5	24.9	33.2						

*Significantly outperformed comparison methods, with all p-values < 0.001.

Appendix O

Illustration of how SATrees Can Facilitate Identification of Key Issues, Suggestions, and Participants

As noted in the main paper, the conversation disentanglement, coherence relation, and speech act classification components of LTAS are combined to create an SATree for each group discussion. Figure O1 presents an example of an SATree. In the tree, each branch represents a conversation; nodes under those branches represent messages in the conversations. Symbols to the left of each message are used to indicate speech act composition; for example, assertions ↑, directive-suggestions ☆, directive-questions **?**, commissives ✓, and expressives ↯. Even from this small example, it is apparent that this particular discussion encompasses multiple conversations, some of which have elaborate interaction patterns and diverse message speech act compositions.



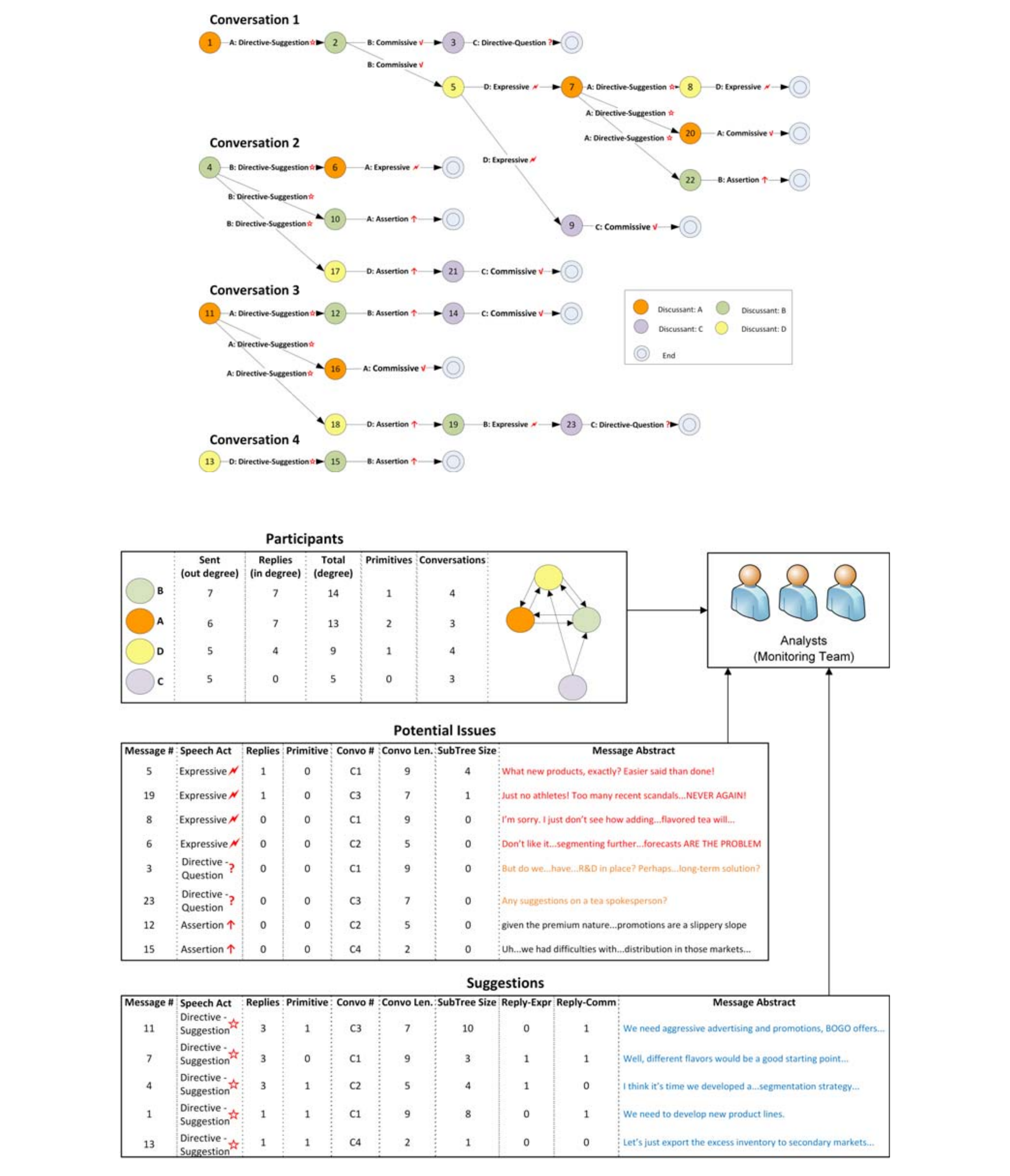
- We need to develop new product lines.
- Yes, new products are critical!
- But do we really have the R&D in place? Perhaps a good long-term solution...
- I think it's time we developed a formal customer segmentation strategy for sales, forecasting, and production.
- What new products, exactly? Easier said than done!
- Don't like it...segmenting further will mean forecasting smaller customer groups. Our forecasts ARE THE PROBLEM
- Well, different flavors would be a good starting point. We offer fewer flavors than our competition.
- I'm sorry, I just don't see how adding Strawberry and Blueberry flavored tea will alleviate excess inventory.
- Agreed! Didn't we talk about this before? Déjà vu all over again!
- Less diverse segments are more predictable. competitors are going through this too...its the economic downturn
- We need aggressive advertising and promotions, BOGO offers, free tea giveaway events for branding, etc.
- perhaps, but given the premium nature of our offerings, promotions are a slippery slope...undercutting margins
- Let's just export the excess inventory to secondary markets. We've been researching them for years
- We don't cut margins if viewed as a marketing expenditure. We're investing in customer acquisition and branding.
- Uh, as I recall, we had difficulties with setting up distribution in those markets...
- A charity event idea sounds promising…tea for a cause
- Some relevant segments currently under-utilized: younger looking for a coffee substitute, health conscious.
- Events such as that need star power. We would need to get a celebrity spokesperson
- Just no athletes! Too many recent scandals, and not worth the headaches and cost of endorsement deal
- A diverse product portfolio is a GOOD thing. Maybe not 31 fruit flavored teas, but the data doesn't lie
- Yes, I think shifting preferences across segments is responsible….our excesses are in traditional products
- No one said we need to take the Skittles approach. But tension-tamer, Chamomile, and select fruits are legit
- Any suggestions on a tea spokesperson?

- ☆ We need to develop new product lines.
  - ✓ Yes, new products are critical!
    - ? But do we really have the R&D in place? Perhaps a good long-term solution?
    - ↯ What new products, exactly? Easier said than done!
      - ☆ Well, different flavors would be a good starting point. We offer fewer flavors than our competition.
        - ↯ I'm sorry, I just don't see how adding Strawberry and Blueberry flavored tea will alleviate excess inventory.
        - ✓ A diverse product portfolio is a GOOD thing. Maybe not 31 fruit flavored teas, but the data doesn't lie
        - ↑ No one said we need to take the Skittles approach. But tension-tamer, Chamomile, and select fruits are legit
      - ✓ Agreed! Didn't we talk about this before? Déjà vu all over again!
- ☆ I think it's time we developed a formal customer segmentation strategy for sales, forecasting, and production.
  - ↯ Don't like it...segmenting further will mean forecasting smaller customer groups. Our forecasts ARE THE PROBLEM
  - ↑ Less diverse segments are more predictable. competitors are going through this too...its the economic downturn
  - ✓ Yes, I think shifting preferences across segments is responsible….our excesses are in traditional products
    - ↑ Some relevant segments currently under-utilized: younger looking for a coffee substitute, health conscious.
- ☆ We need aggressive advertising and promotions, BOGO offers, free tea giveaway events for branding, etc.
  - ↑ given the premium nature of our offerings, promotions are a slippery slope...undercutting margins
    - ✓ We don't cut margins if viewed as a marketing expenditure. We're investing in customer acquisition and branding.
  - ✓ A charity event idea sounds promising…tea for a cause
  - ↑ Events such as that need star power. We would need to get a celebrity spokesperson
    - ↯ Just no athletes! Too many recent scandals, and not worth the headaches NEVER AGAIN!
    - ? Any suggestions on a tea spokesperson?
- ☆ Let's just export the excess inventory to secondary markets. We've been researching them for years
  - ↑ Uh, as I recall, we had difficulties with setting up distribution in those markets...

Figure O1. Illustration of SATree Showing Conversations, Coherence Relations, and Speech Acts

By incorporating coherence relations in conjunction with message speech act composition information, SATree is able to (1) represent conversation structure by depicting interactions between users and their messages and (2) depict user actions in the appropriate conversation context within which they occur. Consequently, the information encompassed in SATrees is well-suited to support analyst social media sense-making use cases, such as identifying key issues, suggestions, and participants. This point is illustrated in Figure O2. The top half of the figure shows the four conversations from the SATree depicted in Figure O1, using a "conversation tree" structure format similar to the one employed by Winograd and Flores (1986). The bottom half shows how conversation structure, reply-to relations, and message speech act labels can support analyst use cases such as participant ranking, issue identification, and discovery of key suggestions. We elaborate further on these items in the ensuing paragraphs.

As noted in the main paper, effective representation of reply-to relations allows more accurate discussant centrality measures and social network representation. The *Participant* box in the bottom half of Figure O2 lists the in/out and total degree centrality for the four discussants in the tea manufacturing chat thread. Although all four discussants posted a roughly even number of messages (between five and seven), Discussants B and A received far more replies, resulting in higher overall degree centrality in the network. B, A, and D appear more central in the network, are responsible for starting all four conversations, and for generating all suggestions, expressives, and assertions in the thread. Conversely, none of Discussant C's messages, which are mostly commissives or unanswered questions, received any replies. The example illustrates how, even with only a single discussion thread comprising 23 messages exchanged between 4 discussants, the language action perspective (LAP) based text analytics system (LTAS) can support analyst sense-making regarding key discussion participants.



Participants

	Sent (out degree)	Replies (in degree)	Total (degree)	Primitives	Conversations
B	7	7	14	1	4
A	6	7	13	2	3
D	5	4	9	1	4
C	5	0	5	0	3

Potential Issues

Message #	Speech Act	Replies	Primitive	Convo #	Convo Len.	SubTree Size	Message Abstract
5	Expressive ⚡	1	0	C1	9	4	What new products, exactly? Easier said than done!
19	Expressive ⚡	1	0	C3	7	1	Just no athletes! Too many recent scandals...NEVER AGAIN!
8	Expressive ⚡	0	0	C1	9	0	I'm sorry. I just don't see how adding...flavored tea will...
6	Expressive ⚡	0	0	C2	5	0	Don't like it...segmenting further...forecasts ARE THE PROBLEM
3	Directive - Question ?	0	0	C1	9	0	But do we...have...R&D in place? Perhaps...long-term solution?
23	Directive - Question ?	0	0	C3	7	0	Any suggestions on a tea spokesperson?
12	Assertion ↑	0	0	C2	5	0	given the premium nature...promotions are a slippery slope
15	Assertion ↑	0	0	C4	2	0	Uh...we had difficulties with...distribution in those markets...

Suggestions

Message #	Speech Act	Replies	Primitive	Convo #	Convo Len.	SubTree Size	Reply-Expr	Reply-Comm	Message Abstract
11	Directive - Suggestion ☆	3	1	C3	7	10	0	1	We need aggressive advertising and promotions, BOGO offers...
7	Directive - Suggestion ☆	3	0	C1	9	3	1	1	Well, different flavors would be a good starting point...
4	Directive - Suggestion ☆	3	1	C2	5	4	1	0	I think it's time we developed a...segmentation strategy...
1	Directive - Suggestion ☆	1	1	C1	9	8	0	1	We need to develop new product lines.
13	Directive - Suggestion ☆	1	1	C4	2	1	0	0	Let's just export the excess inventory to secondary markets...

Figure O2. Illustration of How SATree Information Can Illuminate Conversation Structures and Actions to Support Key Analyst Use Cases

The speech act classification component of LTAS can detect suggestions, one of the major types of directives found in user-generated content. The *Suggestions* box in the bottom half of Figure O2 depicts the five suggestions presented in the discussion thread. The box also depicts other conversation structure and speech act-related dimensions for each suggestions, including number of replies (as well as number of expressive or commissives replies), whether the suggestion is the primitive message in its conversation, the total length of the conversation containing the suggestion, and the number of messages that followed this one in the conversation (i.e., its sub-tree size). These are just examples of the types of variables analysts can use to sort lists of suggestions derived using LAP-based system that produces SATree-type information. In this particular thread, three of the proposed suggestions seem to garner the most attention: advertising and promotions, introducing different flavors, and a formal customer segmentation strategy. All three also have direct replies with commissives and/or expressives, which indicate the suggestions are being evaluated within the conversations. Analysts can use such information to more easily identify suggestions, see which ones are generating discussion, evaluate the level of support/opposition to these suggestions within their respective conversations, and peruse the conversations for greater context.

As mentioned in sections on the need for sense-making and the language-action perspective of the main paper, in the TelCorp example discussion threads, the issue conversations included greater frequencies of questions, assertions of indifference/negligence, negative expressives, and declarations of having switched to other providers. Hence, potential issues could include negative expressives and assertions, or unanswered questions. The *Potential Issues* box in the bottom half of Figure O2 lists all expressives and questions appearing in the discussion thread, as well as select assertions (in this case ones with negative sentiment). Examples include discussants' expressives wondering what new products could help, how such a solution might alleviate the excess inventory problem, current forecasting issues, and questions about current R&D capabilities. Once again, it is important to note that the columns in the box are illustrative, rather than exhaustive. For instance, an analyst may wish to include a count of the speech act composition of all messages in a suggestion's subtree to get a quick broader sense of how that suggestion was received by others in the conversation. The purpose here is to demonstrate the utility of LTAS which advocates consideration of the interplay between conversations, reply-to-relations, and speech acts.

This illustration presents a couple of key takeaways. First, even within a single discussion thread, there is considerable information that systems geared toward LAP can help derive pertaining to participants, suggestions, and issues. Second, many social media monitoring teams at large organizations encounter large volumes of user-generated content every hour. It is conceivable that an analyst at a company such as TelCorp might have a couple of minutes or less to make sense of such a discussion thread to check for problems and/or opportunities, or to identify key contributors. In such contexts, having conversation metrics, reply-to data, and speech act information at one's disposal can be invaluable. Later in the field experiment results section in the main paper, and Appendix M, the 4-month TelCorp field study results shed light on the potential value proposition of such a LAP-based IT artifact for supporting sense-making in organizational settings.

Reference

Winograd, T., and Flores, F. 1986. *Understanding Computers and Cognition*, Norwood, NJ: Abex Publishing.