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.

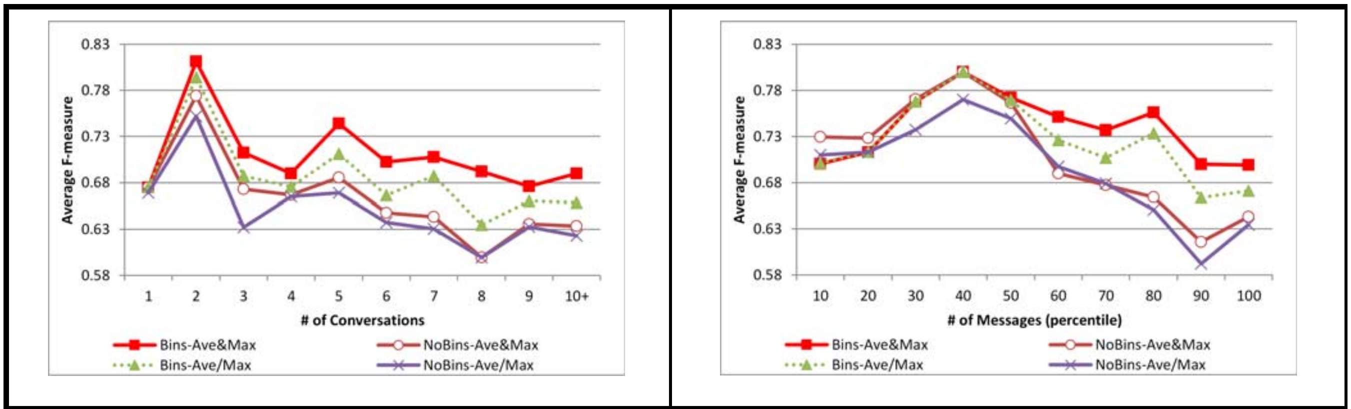
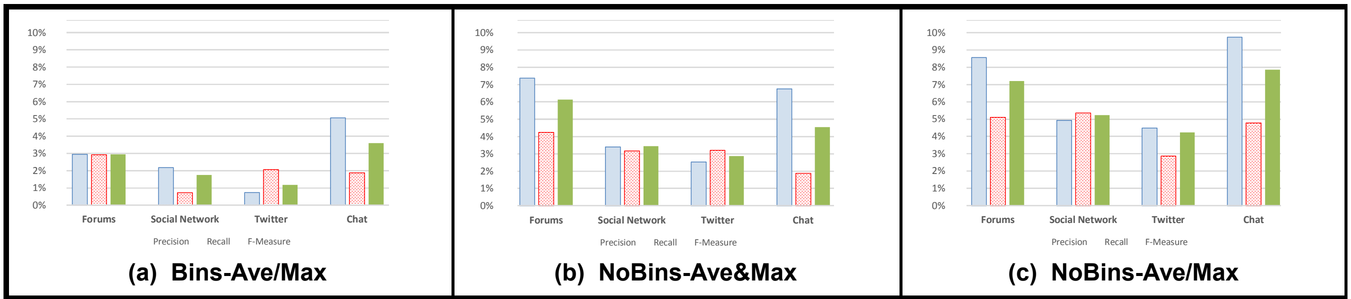


Figure B1. Average f-Measures for Proposed Bins-Ave&Max Method and Alternatives Across Discussion Threads Grouped by Number of Conversations (left) and Number of Messages (right)

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**						

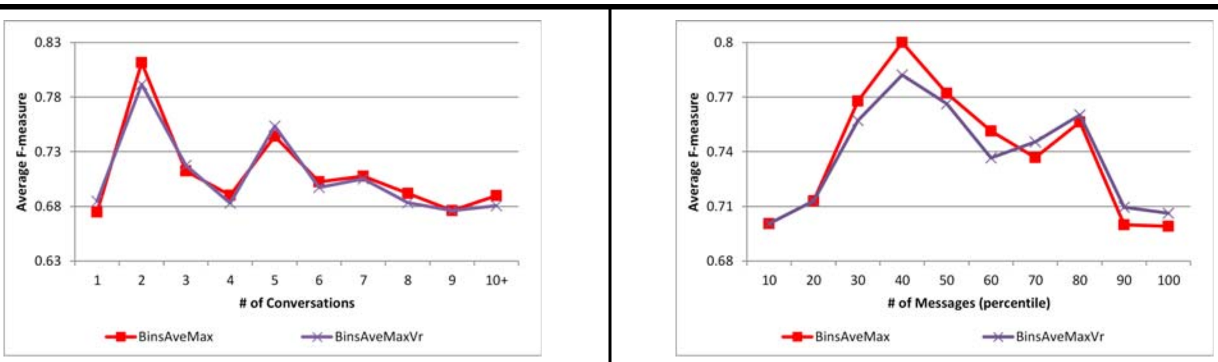


Figure B3. Average f-Measures for BinsAveMax Method and BinsAveMaxVr Alternative Across Discussion Threads Grouped by Number of Conversations (left) and Number of Messages (right)

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.

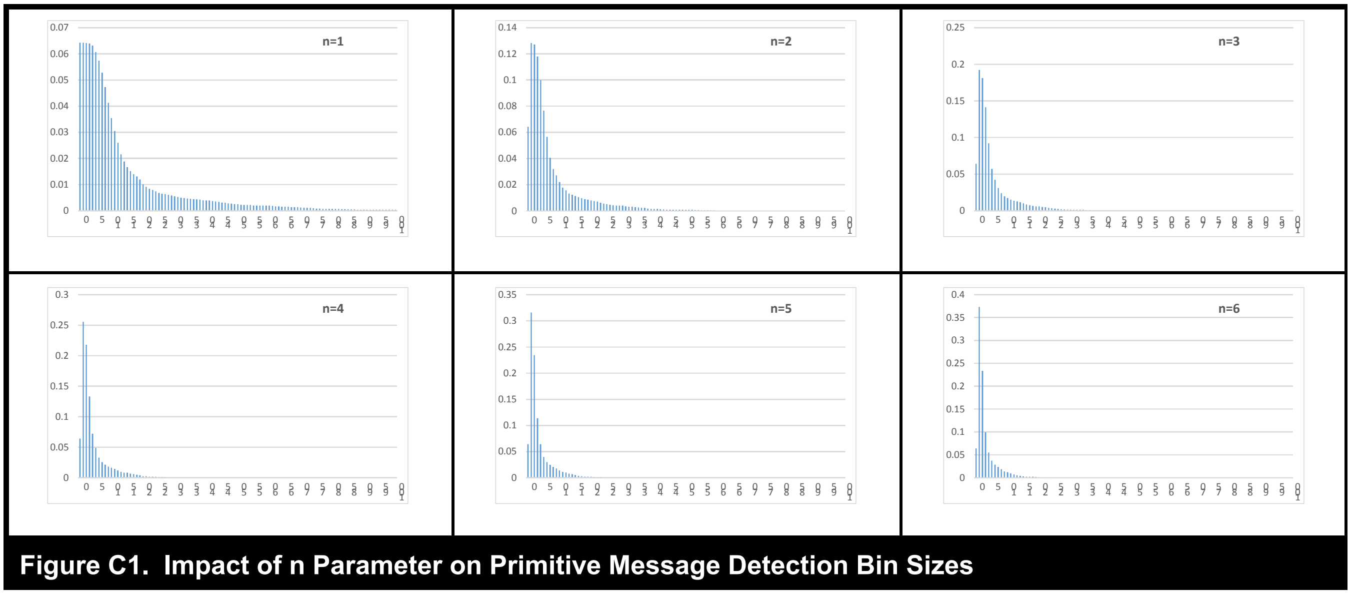


Figure C1. Impact of n Parameter on Primitive Message Detection Bin Sizes

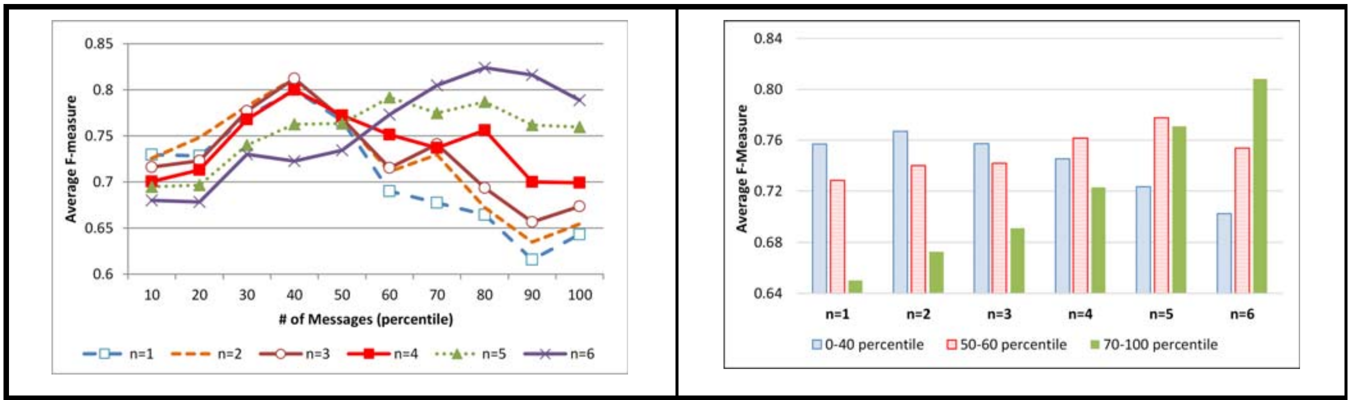


Figure C2. Average f-Measures for Primitive Message Detection Method Using Varying Values for n, Grouped by Thread Length Percentiles (left) and Percentile Range Aggregation (right)

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									
	Web Forum			**Social Network (Patients)**			**Microblog (Twitter)**		
Technique	**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									
	Web Forum			**Social Network (Facebook)**			**Microblog (Twitter)**		
Technique	**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									
	Chat								
Technique	**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.

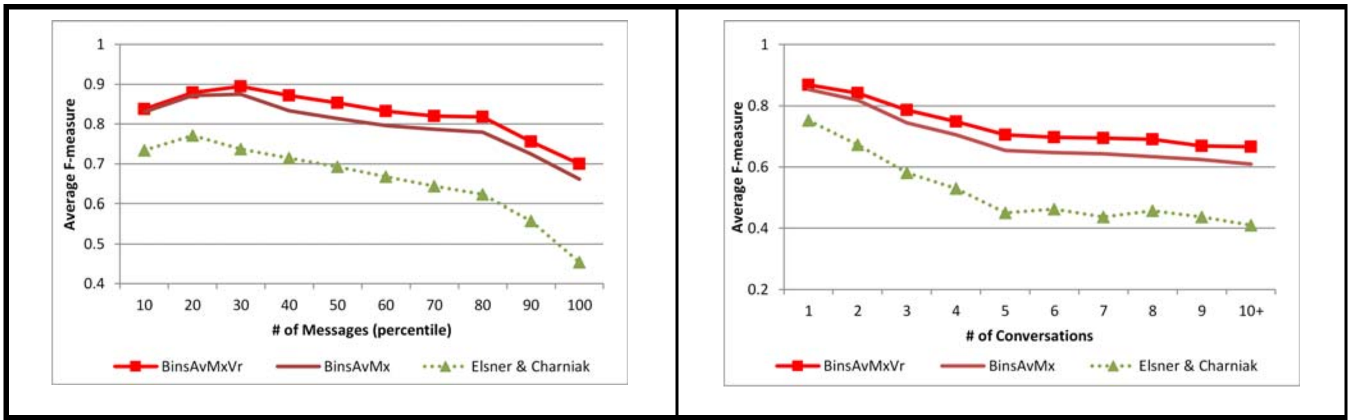


Figure D1. Average f-Measures for Conversation Disentanglement, Grouped by Number of Messages (left) and Number of Conversations (right)

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.

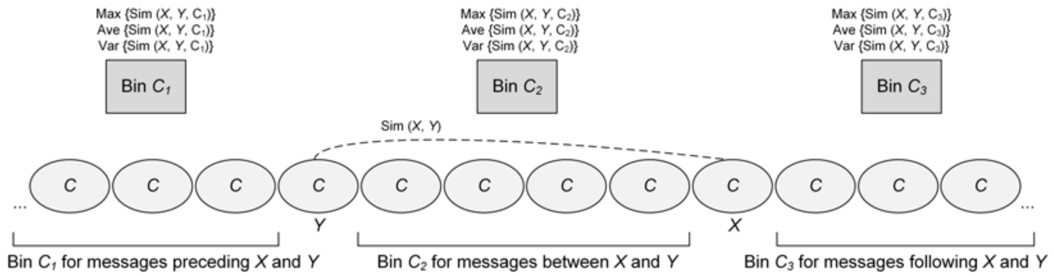


Figure E1. Illustration of Regions, Bins, and Similarity Scores Used in Affiliation Classification Stage Devoid of Primitive Message Detection Component

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									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Feature Set	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									
	Web Forum			Social Network (Patients)			Microblog (Twitter)		
Feature Set	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									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Feature Set	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									
	Chat								
Feature Set	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.

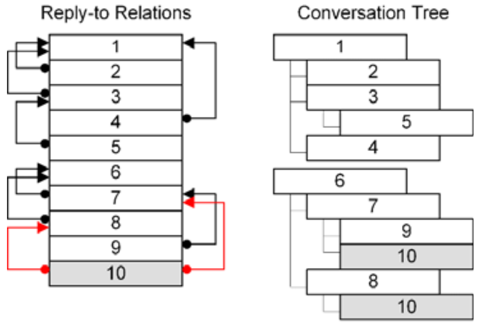


Figure F1. Illustration of How Duplicate Nodes are Created for Child Messages with Multiple Parents

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.

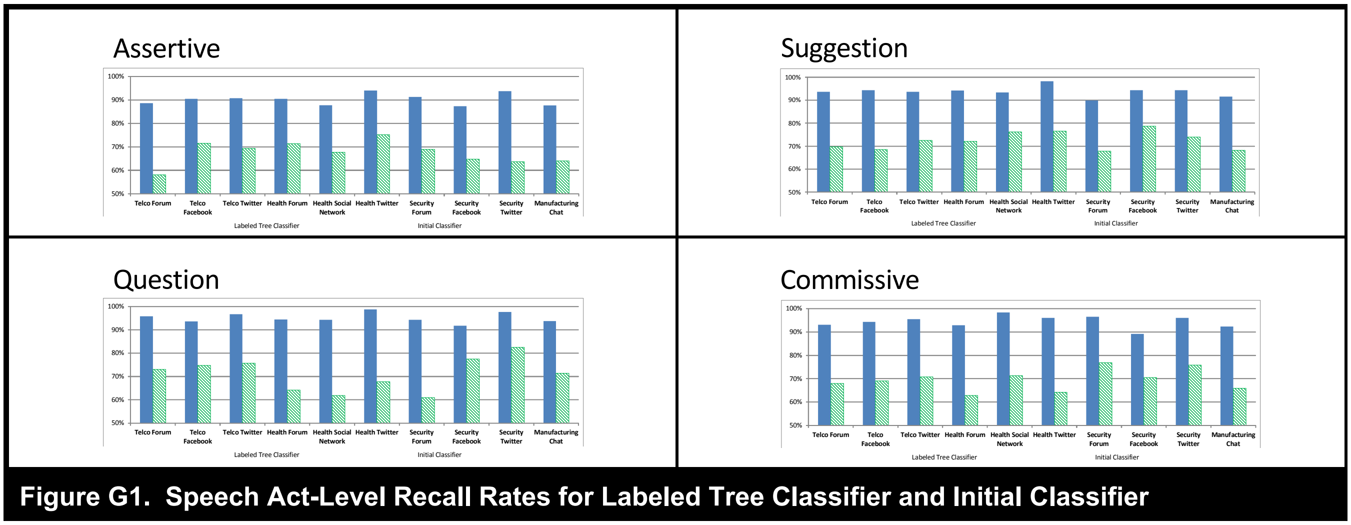


Figure G1. Speech Act-Level Recall Rates for Labeled Tree Classifier and Initial Classifier

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).

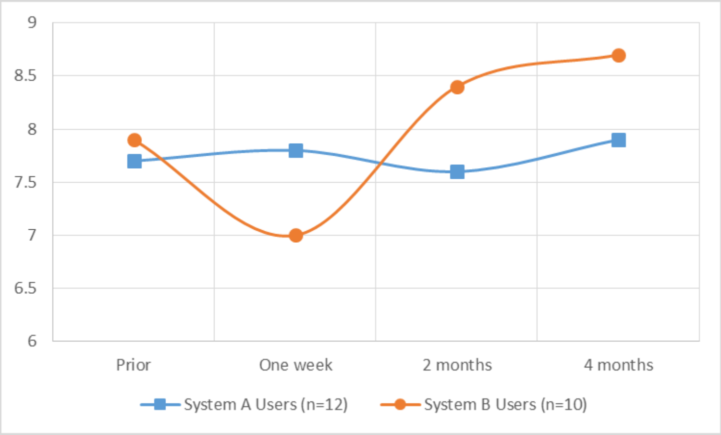


Figure J1. User Perceptions of Overall System Usefulness at Different Points During Field Experiment

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									
Method	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	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									
Method	Web Forum			Social Network (Patients)			Microblog (Twitter)		
	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									
Method	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	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									
Method	Chat								
	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).

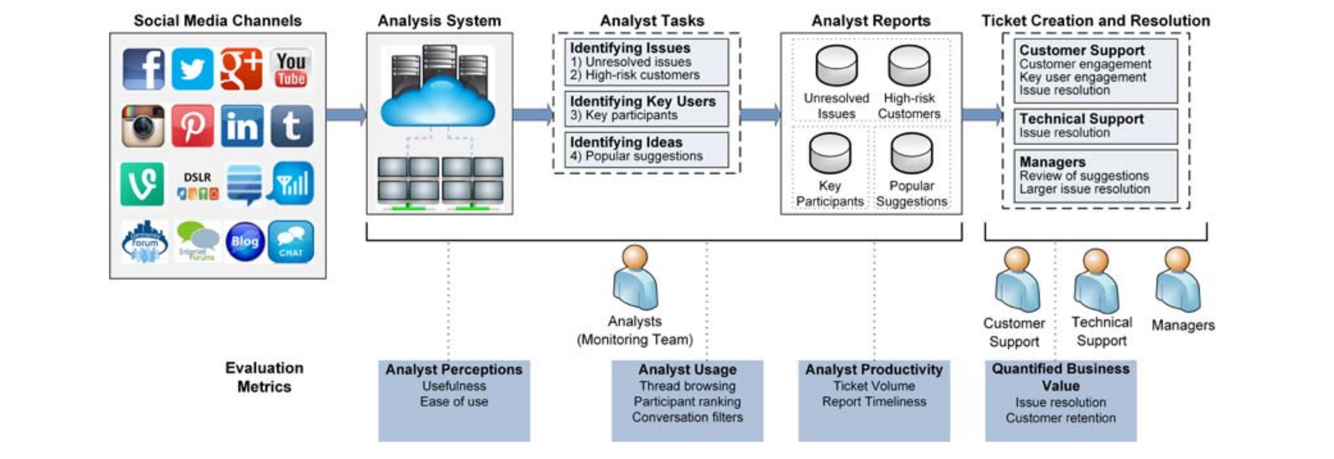


Figure M1. High-Level Overview of TelCorp's Business Process for Social Media Monitoring

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).

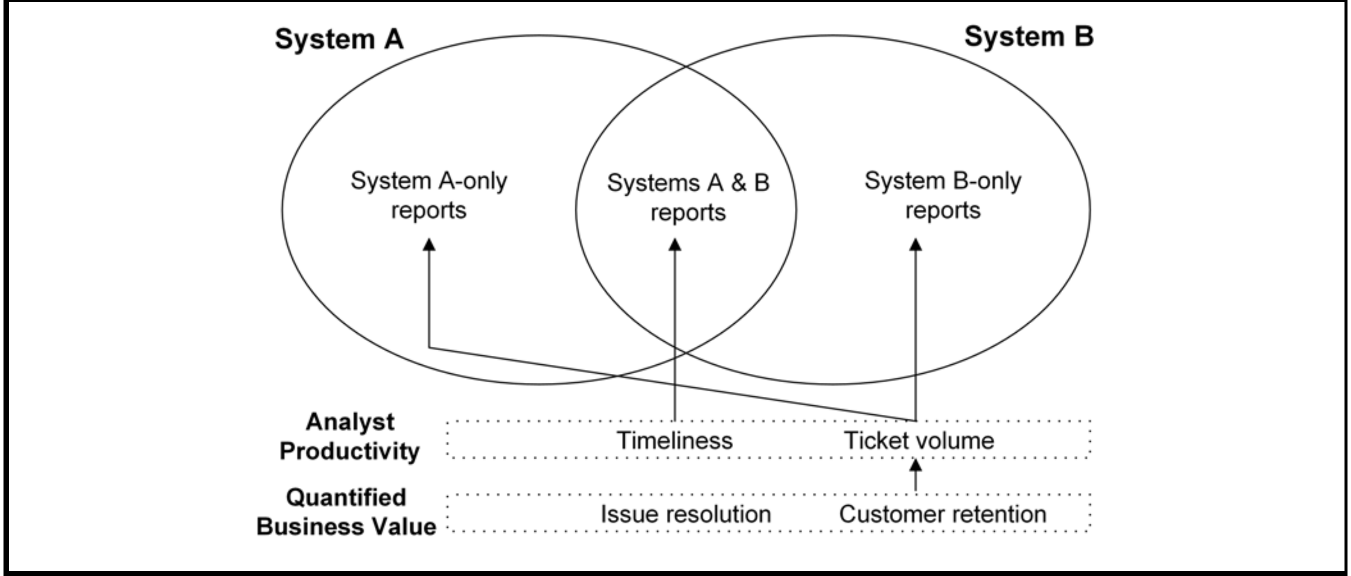


Figure M2. Method for Productivity and Business Value Assessment of Systems A and B

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%.

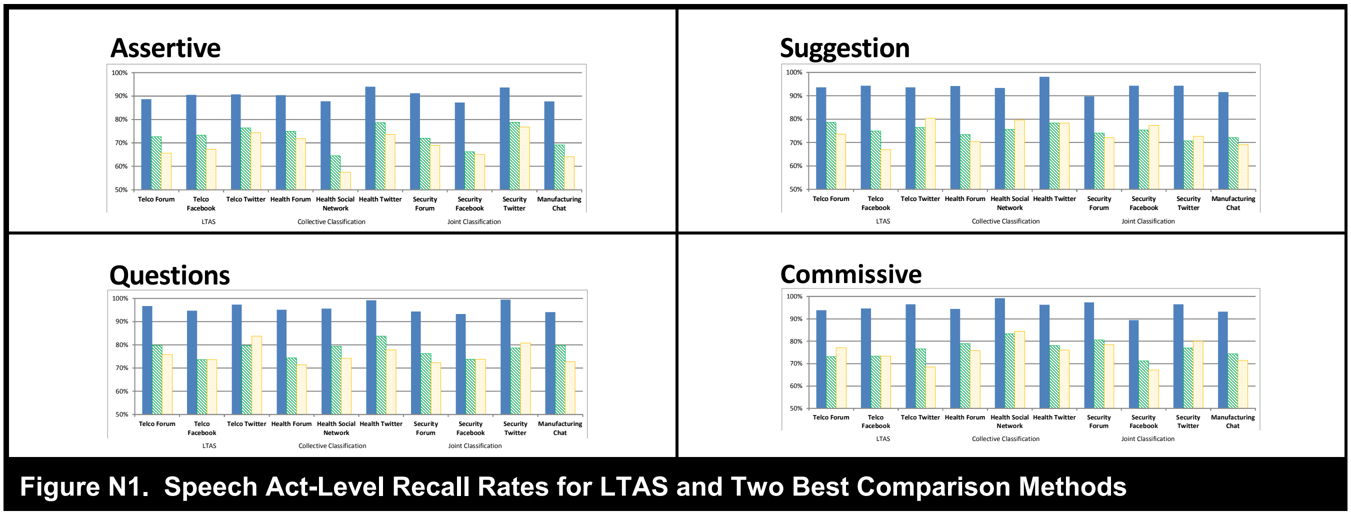


Figure N1. Speech Act-Level Recall Rates for LTAS and Two Best Comparison Methods

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

Technique	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	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

Technique	Web Forum			Social Network (Patients)			Microblog (Twitter)		
	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

Technique	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
	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

Technique	Chat		
	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

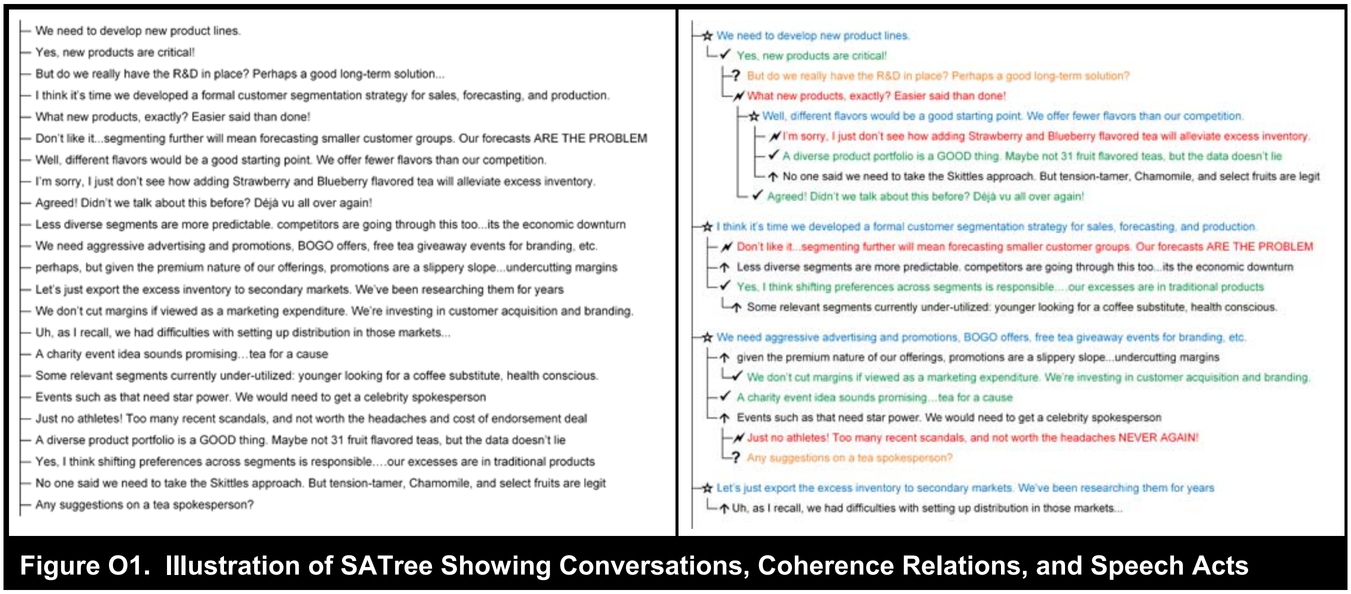
Telecom									
	Web Forum			Social Network (Facebook)			Microblog (Twitter)		
Technique	Degr.	Close.	Betwe.	Degr.	Close.	Betwe.	Degr.	Close.	Betwe.
LTAS*	**99.0**	**99.1**	**83.4**	**99.3**	**99.8**	**94.5**	**99.7**	**99.9**	**96.5**
Heuristic	90.4	74.1	36.7	91.2	76.0	35.5	93.0	81.5	40.5
Classification	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.



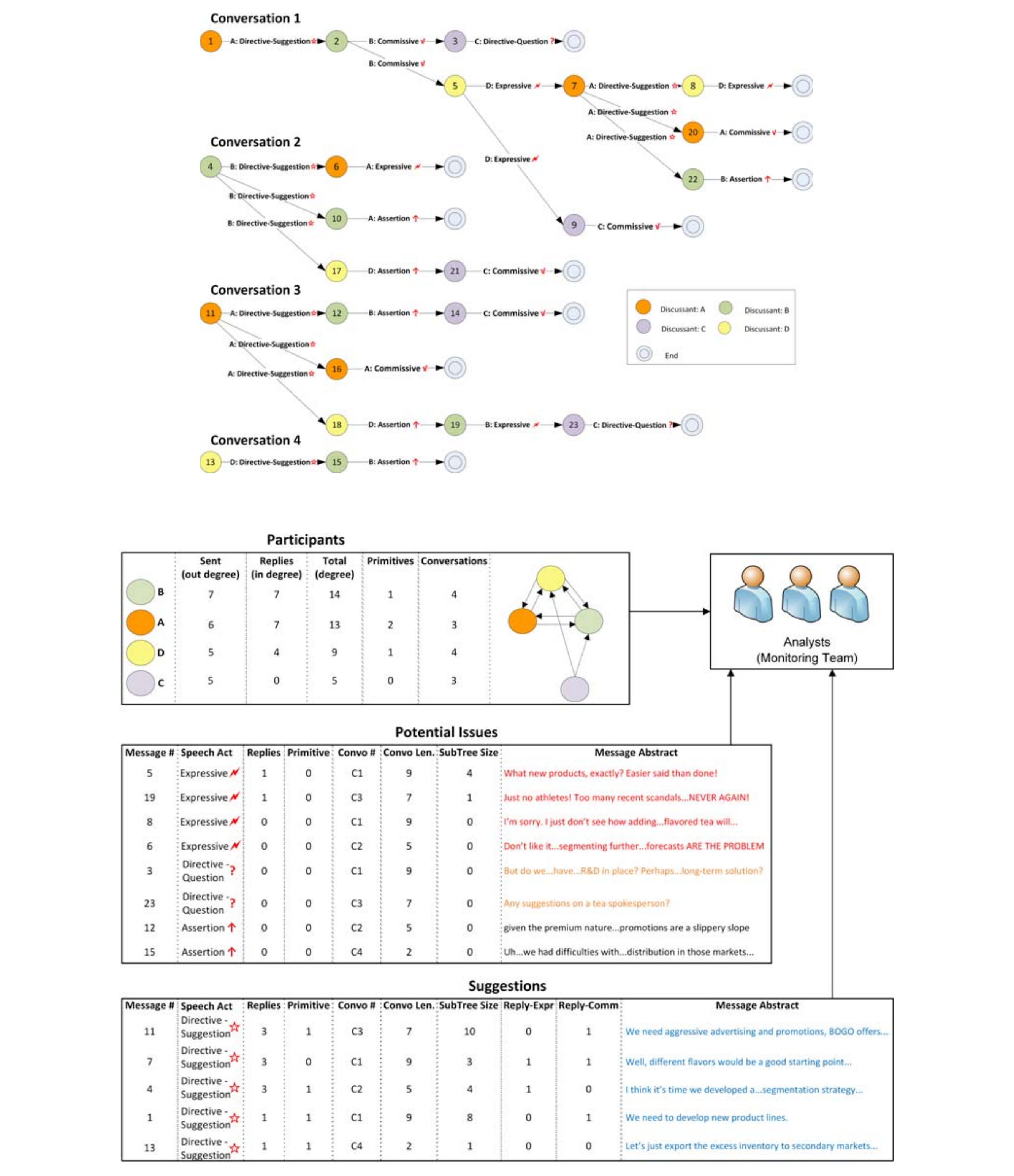
- We need to develop new product lines.
- Yes, new products are critical!
- But do we really have the R&D in place? Perhaps a good long-term solution...
- I think it's time we developed a formal customer segmentation strategy for sales, forecasting, and production.
- What new products, exactly? Easier 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.