Modeling Interactions in Web Forums

Weifeng Li
Artificial Intelligence Lab
University of Arizona
weifengli@email.arizona.edu

Ahmed Abbasi
Predictive Analytics Lab
University of Virginia
abbasi@comm.virginia.edu

Shiyu Hu, Victor Benjamin,
Hsinchun Chen
Artificial Intelligence Lab
University of Arizona

ABSTRACT

The ability to accurately identify "reply-to" relations in online discussions has important implications for various social media analytics applications. However, accurately identifying such interactions remains a challenge, with existing methods providing inadequate performance. In this study, we propose a novel method for modeling social media interactions. The proposed method leverages several empirical insights about online interaction patterns, coupled with a robust machine learning algorithm, for enhanced classification of social media interactions. Furthermore, the proposed method also facilitates the creation of more accurate social media networks. As topological information derived from online communication continues to play an integral role in various social media analytics application areas, the results of our work have important implications.

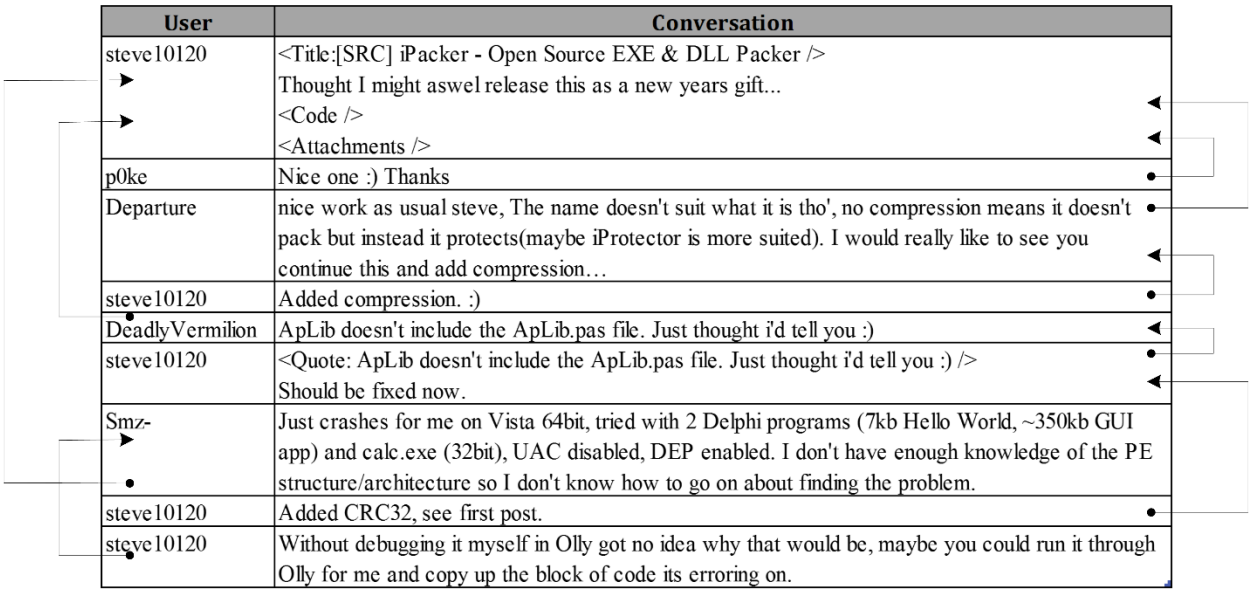
Keywords — interactional coherence analysis; modeling interactions; social media analytics; text mining; machine learning;

I INTRODUCTION

In communication, an interaction occurs when a conversant sends a message to one or more recipients [1]. We often refer to these as "reply-to" relations [2]. Such interactions form the building blocks for various social media analytics use cases with implications for researchers and practitioners. For instance, two questions many researchers and practitioners like to ask are: (1) Which members of the social media network are the most important? (2) Which postings in a discussion thread have garnered the most responses? These questions are the basis for many important types of social media-based analyses, including identifying influential community members, examining network cliques, ranking messages based on importance, identifying unresolved issues, etc. [3][4]. Accurately identifying interaction patterns is essentially the price of admission for performing such advanced analyses. However, interaction identification remains a non-trivial problem. One of the major problems is disrupted turn adjacency: the fact that messages within a discussion thread are often not referring to those that immediately precede them. In order to illustrate the problem, consider the discussion thread sample presented in Figure 1. In the figure, arrows are used to indicate which previous message subsequent messages are responding to. The figure illustrates the concept of disrupted turn adjacency; in the example, only 3 of 8 messages that precede the initial posting respond to the previous message. This situation is highly pervasive in most thread-based communication modes, including web forums, social networking sites, chat, blogs, and micro-blogs [5][6][7][8].

Interactional coherence analysis (ICA) techniques leverage text mining and natural language processing methods to correctly determine "reply-to" interactions in various communication modes [2][5][8]. Simply put, the goal of ICA is to predict the correct interaction arrows depicted in Figure 1. However, prior ICA methods have primarily leveraged manually crafted rules for assigning interaction relations [7]. Consequently, their precision and recall rates for assigning interactions have been sub-par.

In order to address this gap, we propose a novel machine-learning method for ICA that learns robust interaction patterns. Incorporating critical empirical insights from observed interactions in social media, the proposed method attains markedly better classification performance than existing methods. As a result, social networks built based on the proposed method are more accurate in terms of their representation of centrality measures for key community members. As topological information derived from communication channels continues to play an integral role in various social media analytics application areas, the results of our work have important implications.



User	Conversation
steve10120	<Title:[SRC] iPacker - Open Source EXE & DLL Packer /> Thought I might aswel release this as a new years gift... <Code /> <Attachments />
p0ke	Nice one :) Thanks
Departure	nice work as usual steve, The name doesn't suit what it is tho', no compression means it doesn't pack but instead it protects(maybe iProtector is more suited). I would really like to see you continue this and add compression…
steve10120	Added compression. :)
DeadlyVermilion	ApLib doesn't include the ApLib.pas file. Just thought i'd tell you :)
steve10120	<Quote: ApLib doesn't include the ApLib.pas file. Just thought i'd tell you :) /> Should be fixed now.
Smz-	Just crashes for me on Vista 64bit, tried with 2 Delphi programs (7kb Hello World, ~350kb GUI app) and calc.exe (32bit), UAC disabled, DEP enabled. I don't have enough knowledge of the PE structure/architecture so I don't know how to go on about finding the problem.
steve10120	Added CRC32, see first post.
steve10120	Without debugging it myself in Olly got no idea why that would be, maybe you could run it through Olly for me and copy up the block of code its erroring on.

Figure 1. Sample Discussion Thread with Disrupted Turn Adjacency (arrows indicate "reply-to" interactions between messages)

II RELATED WORK

Prior methods for ICA have mostly leveraged manually crafted rules for inferring "reply-to" relations. Three popular rule-based methods that have been commonly adopted are (1) reply to the first message; (2) reply to the previous message; and (3) reply to all previous messages [7][9]. Some such rule-based methods embody reasonable intuitions. For instance, many messages do respond to the prior message, or the first one [8][10]. However, such rules are gross over-simplifications of social media interaction patterns.

Other studies have incorporated system-enabled and linguistic features [7]. Collectively, these are often referred to as interaction cues. System features include the presence of re-quoted content [11]. Linguistic features encompass attributes such as lexical relation, direct address, and co-reference [9]. Direct address is when a message sender mentions the recipient by name [12]. For instance, in Figure 1, in the third message the sender mentions the recipient "Steve" by name. Lexical relations are based on the notion that messages discussing similar topics are likely to contain the same or similar keywords due to repetition and synonymy [7]. Co-references are interaction cues such as pronouns and comparison terms such as "same" and "similar." Some recent methods have developed a manual collection of rules arranged as finite state automata which incorporates several of the aforementioned system and linguistic features.

With respect to state-of-the-art ICA methods, the Hybrid Interactional Coherence method (HIC) has yielded the best performance [7][9]. HIC firstly checks for system features such as re-quoted content, which are considered "lower hanging fruit" from a classification precision perspective. It then sequentially checks for linguistic features such as direct address and lexical relation, and then uses a rule-based residual match for any unclassified messages. In experiments, it outperformed other rule-based methods and ones based on linguistic features.

Nevertheless, existing methods including HIC are still problematic for two reasons. First, they incorporate rules manually crafted based on observed interactions. However, these rules are often susceptible to over-fitting since they are *observed* on small sets of data, as opposed to being *learned* from larger quantities of social media. Second, most prior methods incorporate limited sets of interaction cues. Consequently, reply-to social networks constructed using these methods are highly erroneous. Consider the example presented in Figure 2. The network in the top-left shows the gold standard for a single discussion thread from a technology web forum. The gold standard was manually crafted using multiple domain experts. The chart on the top-right shows the network for the same thread constructed based on reply-to relations classified by HIC. The ones on the bottom-left and bottom-right are the reply-to-first and reply-to-all-previous rule-based

methods, respectively. From the figure, it is apparent that the three ICA methods all misclassify many interactions. In the figure, node sizes are proportional to their degree centrality in the network. Due to misclassifications, the HIC method (top-right) exaggerates the degree centrality for a few of the users, whereas the reply-to-all-previous inflates the degree centrality for virtually all participants (bottom-right). Conversely, the reply-to-first method significantly exaggerates the first poster's centrality while under-attributing degree to many of the other conversation participants. The figure illustrates some of the implications poor ICA performance can have on the quality of resulting social media networks.

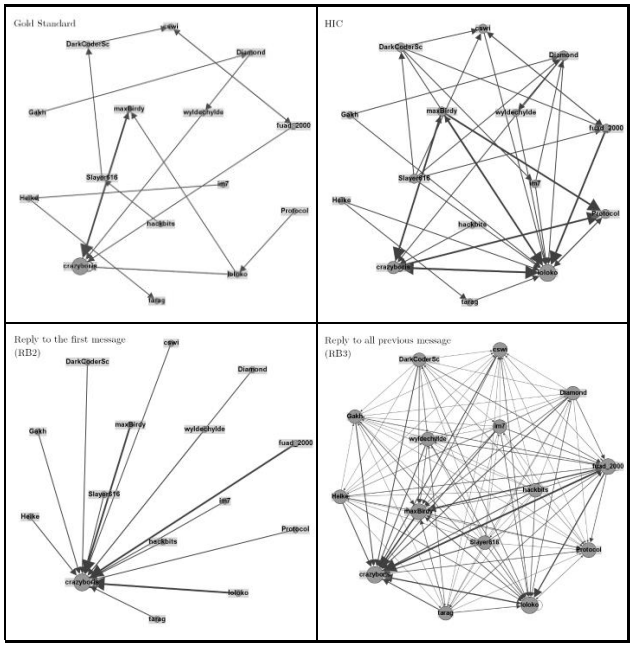


Figure 2. Reply-to Networks: Gold Standard and Comparison Methods

Clearly there is a need for enhanced ICA methods capable of more accurately identifying interaction relations. In the following section, we provide empirical insights regarding three important characteristics of web forum interactions. We use these characteristics to explain the poor performance of many existing ICA methods, and to inform the design of our proposed method.

III UNDERSTANDING WEB FORUM INTERACTIONS: EMPIRICAL INSIGHTS

In order to better understand interactions in web forums, we analyzed over 1,000 postings in 50+ discussion threads spanning 4 web forums. Each message was examined by a pair of annotators familiar with social media and interactional coherence. Based on the analyses, three interesting insights emerged. First, we found that most messages respond to either (1) the first message in the thread, (2) the prior message, or (3) some other message within a short turn proximity window. Second, an array of system and linguistics features are needed to accurately detect reply-to interactions in discussion forums. Third, since most messages respond to only a single previous message, binary machine learning classifiers may be an effective approach for ICA. These insights are elaborate upon in the remainder of this section.

Figure 3 shows the cumulative percentage of reply-to interactions covered (y-axis) using turn-based message proximity windows of different sizes. The analysis was performed in four web forums: elitehack, icode, hackhound, and vctool. The x-axis shows the window size, where 0 means first-message, 1 means the first and/or previous message, 2 means the first and/or a two previous message window. From the figure we can see that in all four forums, between 80% and 95% of messages respond to ones within a turn proximity of 2 (i.e., either the first message in the thread, or one of the two previous messages). This insight suggests that turn proximity may be leveraged as a filtering mechanism for removing less likely distant candidate reply-to relations, thereby reducing false positive rates.

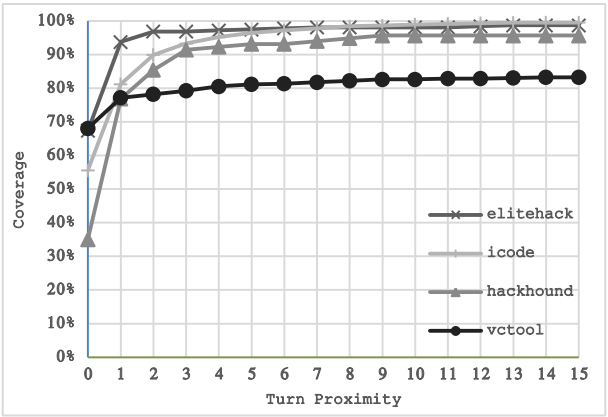


Figure 3. Coverage of Reply-To Relations by Turn Proximity Windows

Figure 4 shows the percentage occurrence distribution of lexical relations, quoted content, co-reference relations, direct address, as well as other categories of relations, for the four aforementioned web forums. From the figure, it is apparent that linguistically-based interaction cues such as lexical

relations are quite pervasive in these forums. Similarly, quoted content is quite common. But the distributions across feature types do vary considerably, as well. For example, lexical relations occur 35% for often in the vctools forum as compared to icode. On the other hand, co-reference relations are quite prevalent in icode, accounting for nearly 25% of interaction cues, but virtually non-existent in the other three forums. These results have two important implications: (1) ICA necessitates the use of a broad set of interaction cues, encompassing various system and linguistic feature categories; (2) manually crafted rule-based approaches are ill-suited to effectively capture the forum-specific intricacies and nuances of communication interactions.

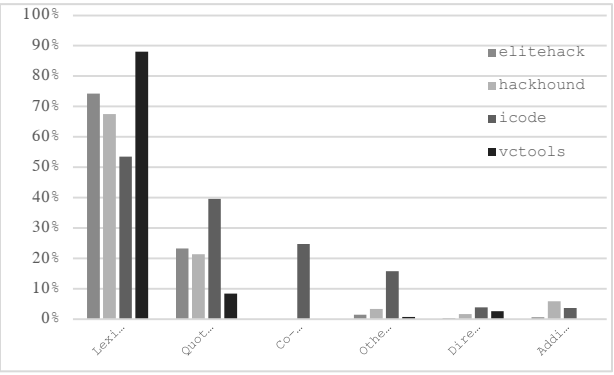


Figure 4. Percentage Interaction Feature-Type Occurrences

Most messages respond to only a single previous message. In our analysis, this was true in over 99% of cases. However, given n messages, the nth message has n-1 previous messages that are potential reply-to candidates. This requires n-1 messages pairs to be evaluated for a single potential interaction [7], resulting in a heavily imbalanced binary classification problem. For example, the 100th message would have to be compared against 99 earlier messages. In order to examine the impact this class imbalance might have on machine learning methods, we developed a basic decision-tree model using select lexical relation, quoted content, direct address, and co-reference features. The model was run using cross-validation on the aforementioned 1,000 message web forum data. We looked at the decision tree classifier's prediction confidence scores on the test set to see where the true positives ranked, given the large number of potential message pairs. Figure 5 shows the analysis results.

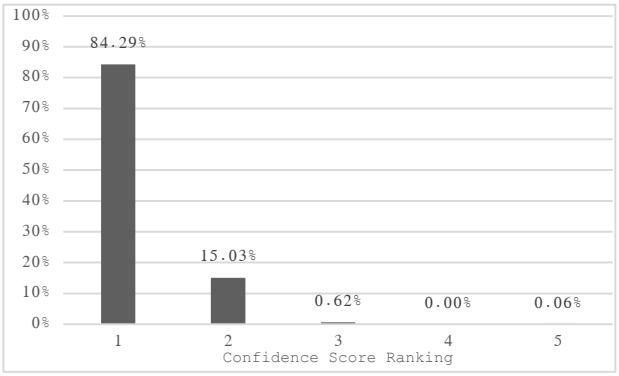


Figure 5. True Positive Distribution Across Machine Learning Model Confidence Score Rankings

From the figure we can see that 84% of true positive classifications are the highest ranked of the n-1 possible message pairs, while another 15% are second. This suggests that when using machine learning classification models, for a given message, only the one or two highest-ranked relations are likely to be true positives. The result provides another mechanism by which to reduce false positive rates by narrowing the reply-to interaction candidate message pair pool. In the following section, we describe our proposed interaction modeling framework which leverages these insights as part of its algorithm intuition.

IV INTERACTION MODELING FRAMEWORK

Figure 6 shows the proposed interaction modeling framework (IMF). IMF encompasses two important facets: a rich set of content and context interaction cues coupled with an optimized decision tree classification model for ICA (called DT-ICA). Descriptions of the two facets of the framework are as presented below.

Most prior studies have at the most used a few linguistic and system features [7][9]. However, Figure 4 illustrated the importance of using a broad set of interaction cues. Accordingly, the IMF feature set encompasses content and context features which include several important attributes. Table 1 provides a description of the feature set. Three types of content features are incorporated: lexical relation, hypernym augmentation, and direct address. Consistent with prior work, the lexical relation focuses on the similarly between message pairs' text using a vector-space model that weights less-common topical keywords higher using TFIDF combined with cosine similarity. Such an attribute facilitates the creation of lexical chains between reply-to relations based on topical similarity [7][13].

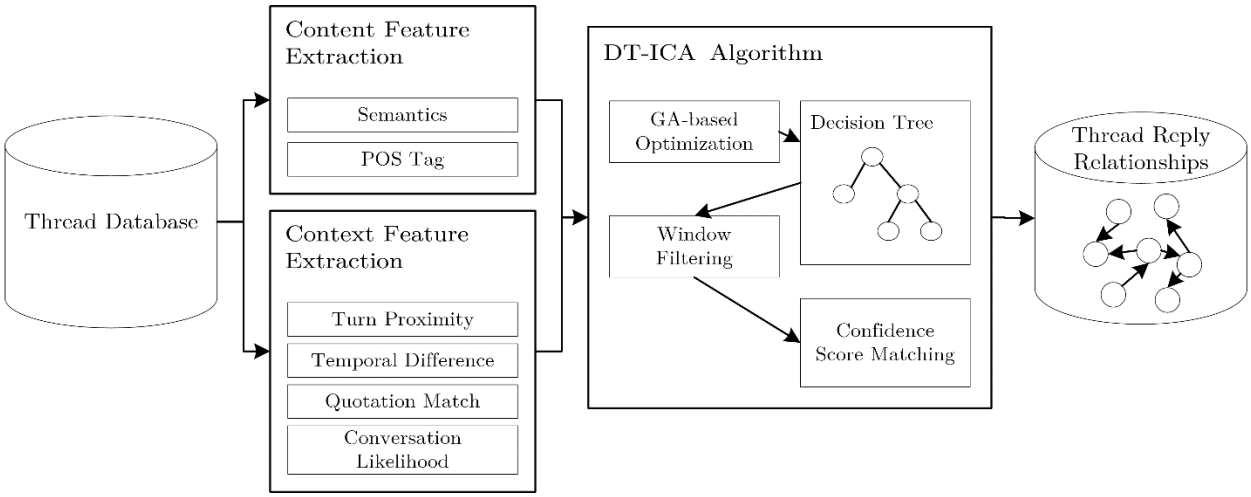


Figure 6. The Interaction Modeling Framework (IMF) for Detecting Reply-To Relations in Social Media

In order to account for lexical synonymy relations, a hypernym augmentation method based on WordNet is adopted [14]. This feature is similar to the lexical relation attribute, but also considers synsets in the similarity calculation. We also included a direct address matching feature which took into account the likelihood of common typos by incorporating a dice-based screen-name similarity matcher.

Feature	Explanation
Content-based Features	
Lexical Relation	The lexical similarities between the message pair
Hypernym Augmentation	The similarity based on the overlapping of hypernyms of nouns between the message pair
Direct Address	The possibility of message y mentioning speaker of x
Context-based Features	
Quotation	Whether message y quoted message x
Temporal Distance	The time interval between the message pair
Length Difference	The length difference between the message pair
Speaker	How active are the two speakers within the thread. 4 features: including both temporal and turn based attributes and both recent window and overall measures
Positional Preference	Derived from rule-based residual match in [7], dedicated to bridging messages that have preferred relative positions
Turn Proximity	Turn difference between two messages.
First Message	Whether one of the message is the first message
Conversation Likelihood	Whether these two speakers are likely conversing in the thread. Given speakers A and B, two variables: average p(A|B) and p(B|A) for each window of length 3 following each message instance of A or B in thread.

Table 1. Feature Set of Interaction Cues used in IMF

The context-based attributes include both system features as well as ones based on the discussion thread history. System-based interaction cues include the presence of quoted content, the temporal distance between the message pairs (based on time stamps), turn proximity, and whether one of the messages in the pair is the first one in the thread. The discussion history-based context cues are the length difference between the two messages [14], how active the two users are in the thread, the likelihood that they are having a conversation based on recent message history, and a positional preference variable based on a commonly observed interaction rule [7].

The features are extracted for each potential message pair in the discussion thread, resulting in a message pair feature matrix X with each message-pair's feature vector as a separate row. In order to incorporate the message turn proximity insight, we filter out message pair rows from X where the turn proximity is greater than w. This feature matrix is then input into the optimized decision tree model for identifying interactions, called DT-ICA. In order to create robust interaction patterns, it is critical to minimize over-fitting with the machine learning models. Accordingly, DT-ICA uses a genetic algorithm to optimize the decision tree parameters across several different training samples. The parameters included as part of the heuristic-based optimization include minimal gain, maximal depth, minimal split-size, and information criterion (e.g., information gain, gain ratio, etc.). We used 5-fold cross-validation on the training set to optimize the tree Θ. Interestingly, this resulted in trees that were smaller, more accurate, more

generalizable, and that outperformed other information theoretic classification methods such as Rule Induction and Random Forest.

DT-ICA Algorithm

Input: Message pair feature matrix $X = \{x_1, x_2, \dots, x_n\}$ where x_i is the feature of ith message pair, $(message1_i, message2_i)$, in the training set, $Z = \{z_1, z_2, \dots, z_n\}$ is the test matrix, and window size, w.
Output: Binary prediction of the linkage Y.
Optimize Decision Tree model using Genetic Algorithm.
Train Decision Tree model with optimized parameters, θ.
for $z_i \in Z$ **do**
  $(y_i, conf_i) = DecisionTree(z_i; \theta)$, where $y_i \in \{true, false\}$ is the prediction and $conf_i$ is the associated confidence score.
  If $z_i.isFirstMessage$ is $false$ and $z_i.turnProximity$ is greater than w **then**
    $y_i \leftarrow false$
  end if
end for
for all $message1_i$ **do**
  Rank $Conf_{message1_i}$, where $Conf_{message1_i} = \{conf_i | z_i.message1 = message1_i\}$
end for
for $z_i \in Z$ **do**
  if The rank of $conf_i$ in $Conf_{message1_i}$ is greater than c **then**
    $y_i \leftarrow false$
  else
    $y_i \leftarrow true$
  end if
end for

Figure 7. Pseudo-Code for DT-ICA Algorithm in IMF

The optimized decision tree is then used to make predictions on the test case feature matrix *Z*. Once again, in light of the message turn proximity insight, we filter out candidate message pair rows from *Z* where the proximity is greater than *w*. In this case, this message pairs are automatically assigned a prediction of *false* (i.e., no reply-to relation). This obviously results in some false negatives, but also alleviates many false positives as previously shown in Figure 3. Once a set of predictions are generated for each remaining row in *Z*, for a given message z_i, we only assign its reply-to = *true* value (i.e., classified as reply-to relation) to the *c* predictions with the highest confidence scores. As previously mentioned in Figure 5, logical values for *c* are 1 or 2.

Overall, IMF incorporates all three of the insights described in the previous section: (1) a rich feature set of content and context attributes that encompass system, linguistic, and discussion history-based cues; (2) the DT-ICA method for classifying interactions, which incorporates turn proximity and reply-to prediction confidence scores. In the following section, we discuss the evaluation of IMF relative to existing methods.

V EVALUATION

Independent annotators were used to develop gold-standard interaction test beds from four web forums: icode.org, hackhound.org, vctools.net, and elitehackforums.com. These forums discuss technology, code, hacking, and related topics that are of interest to the cyber-security community. The threads and messages incorporated in the test bed included ones that were orthogonal to those used to derive the empirical insights in earlier in Section 3. For each forum, we incorporated between 500 and 2,000 forum postings from 20 to 50 discussion threads per forum, featuring between 40 and 200 users per forum. Next, candidate message pairs were generated for each forum. Overall, we utilized slightly under 90,000 candidate message pairs with 3,000+ true interactions and 85,000+ false ones. The inter-annotator coding reliability was high, with a kappa statistic of around 0.8.

IMF was evaluated against four comparison methods: HIC and three rule-based methods. The rule-based methods were reply-to-previous (RB-1), reply-to-first (RB-2), and reply-to-all (RB-3). Both IMF and HIC require training for model development and/or parameter tuning. These methods were run using 5-fold cross validation.

Standard evaluation metrics were incorporated, including class-level precision, recall, and f-measure. For instance, reply-to recall is computed as the percentage of total reply-to relations that were correctly identified as such. Reply-to precision is the percentage of all reply-to classifications that are actually reply-to relations. The f-measure (F1) is the harmonic mean of precision and recall.

The experiment results for each of the four web forums are presented in Tables 2-5. IMF had the highest f-measures for the no reply-to relation category on all four data sets, with higher precision and recall than the best comparison methods for most data sets. IMF also attained the highest reply-to f-measures on three of the four data sets, typically outperforming the best comparison

methods by 5%. The one exception was the hackhound forum, where RB-1 outperformed IMF by less than 1%. IMF typically attained higher precision and recall on the no reply-to class (between 94% and 99%) and ranging from 50% to 77% on the reply-to class.

	Class Reply-To			Class No Reply-To		
	Precision	Recall	F1	Precision	Recall	F1
IMF	70.43%	85.90%	77.40%	99.28%	99.72%	99.49%
RB-1	28.90%	29.18%	29.04%	98.58%	98.56%	98.57%
RB-2	66.88%	67.54%	67.21%	99.35%	99.33%	99.34%
RB-3	1.97%	100.00%	3.86%	0.00%	0.00%	0.00%
HIC	2.87%	40.98%	5.37%	98.38%	72.14%	83.24%

Table 2. Experiment Results on the elitehack Forum

	Class Reply-To			Class No Reply-To		
	Precision	Recall	F1	Precision	Recall	F1
IMF	51.43%	49.32%	50.35%	94.64%	95.05%	94.84%
RB-1	50.00%	52.05%	51.01%	94.88%	94.47%	94.68%
RB-2	36.84%	38.36%	37.58%	93.42%	93.01%	93.22%
RB-3	9.61%	100.00%	17.53%	0.00%	0.00%	0.00%
HIC	17.28%	71.23%	27.81%	95.42%	63.76%	76.44%

Table 3. Experiment Results on hackhound Forum

	Class Reply-To			Class No Reply-To		
	Precision	Recall	F1	Precision	Recall	F1
IMF	61.21%	58.83%	60.00%	97.96%	98.15%	98.05%
RB-1	56.37%	55.50%	55.93%	97.79%	97.87%	97.83%
RB-2	29.52%	28.93%	29.22%	96.47%	96.57%	96.52%
RB-3	4.74%	100.00%	9.05%	0.00%	0.00%	0.00%
HIC	10.71%	13.07%	11.77%	95.60%	94.55%	95.07%

Table 4. Experiment Results on the icode Forum

	Class Reply-To			Class No Reply-To		
	Precision	Recall	F1	Precision	Recall	F1
IMF	73.54%	71.67%	72.59%	99.57%	99.61%	99.59%
RB-1	9.96%	10.15%	10.05%	98.64%	98.61%	98.63%
RB-2	66.80%	68.08%	67.43%	99.52%	99.49%	99.50%
RB-3	1.49%	100.00%	2.94%	0.00%	0.00%	0.00%
HIC	3.83%	28.33%	6.74%	98.80%	89.23%	93.77%

Table 5. Experiment Results on the vctool Forum

With respect to the performance of comparison methods, interestingly, HIC did not perform well at all, despite outperforming rule-based methods such as RB-1 and RB-2 in prior studies [7][9]. Of the comparison methods, RB-1 had the best performance on two of the data sets, while RB-2 was stronger on the other two. Hence, in some forums, messages tended to respond to the previous message. In others, many responded to the first message in the thread. Furthermore, the linguistic and system, and discussion history cues also varied somewhat from forum-to-forum as evidenced by the considerable variation in the performance of HIC. Collectively, these findings underscore the importance of incorporating a broad set of interaction cues in conjunction with methods capable of learning forum-specific idiosyncrasies.

The results have important implications for social network construction. Incorrect reply-to relation classifications can impact the accuracy of centrality measures such as degree (total number of in/out links for a network node) and betweenness (extent to which a node connects others in the network). In order to illustrate this point, Figure 8 shows the mean absolute percentage error for the IMF (labeled DT-ICA in the figure), RB-1, RB-2, and HIC methods relative to the gold standard (y-axes) for the 5 to 25 most central members (x-axes) within the icode web forum. The figure shows the top 25 based on degree centrality. IMF had the lowest degree centrality MAPE across the top 25 most central forum members, with MAPE values ranging between 10% and 20%. RB-1 had the second highest degree centrality MAPE, followed by RB-2 and HIC which had MAPE values of around 40% or higher for the top 25.

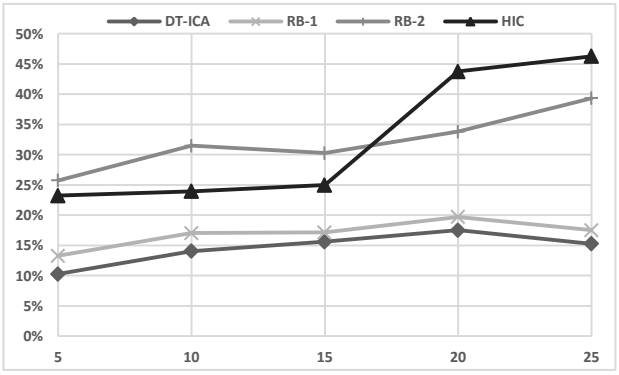


Figure 8. Degree Centrality MAPE for Top 25 Forum Members

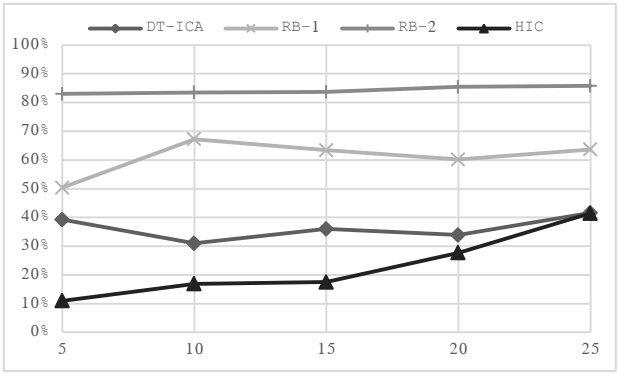


Figure 9. Betweenness Centrality MAPE for Top 25 Forum Members

Figure 9 shows the MAPE values for betweenness centrality for the 25 icode forum members with the highest betweenness centrality. Interestingly, despite its poor overall performance, HIC had lower betweenness MAPE than IMF for the first 20 forum members. However, by the 25 member-mark, the

two methods were comparable. Furthermore, RB-1 and RB-2 had betweenness MAPEs that were in the 60% to 80% range.

Overall, the results presented in Tables 2-5 and Figures 8-9 demonstrate the effectiveness of IMF relative to comparison methods. Enhanced precision, recall, and f-measure values for reply-to and no reply-to cases resulted in generally better MAPE values compared to existing rule-based methods as well as ones leveraging system and linguistic features. Earlier on, in Figure 2, we showed how poor ICA performance can distort social media network linkages and centrality measures. Conversely, IMF is able to produce networks that are far more accurate. In order to illustrate this point, Figure 10 shows gold standard and IMF (labeled DT-ICA in the figure) networks for the top 10 users in the icode forum. The IMF network fairly closely resembles the gold standard in terms of both linkage patterns and degree centrality values for nodes (which are represented by node size). The figure underscores the impact of enhanced ICA on social media network quality.

VI CONCLUSION

In this work we presented IMF, a novel method for ICA that leverages an extended set of interaction cues spanning content (i.e., linguistic cues) and context (i.e., system and discussion history) features in conjunction with an optimized decision tree classifier (DT-ICA). IMF's superior performance relative to comparison methods was largely attributable to the use of a machine learning method that adopted important insights about forum interaction dynamics, facilitating enhanced ICA capabilities. These insights include the pervasiveness of a broad set of interaction cues and various idiosyncrasies pertaining to these cues, impact of turn proximity, and true positive rankings in a heavily imbalanced classification problem.

Given that social network analysis is a critical technique in social media analytics, the results have important implications for several use cases. For instance, researchers and practitioners are increasingly interested in identifying the most important members within online communities [16][17]. Social network centrality measures play a large role in such analyses. Similarly, understanding the propagation of influence in social media is another important topic receiving considerable attention. ICA plays a critical role in such analysis, especially when using links based on reply-to relations such as those found in web forums, social networking sites, chat, blogs, and other forms of social media. Accurate reply-to relation classification is also important for identifying the most important content, such as the messages receiving the greatest number of responses [15][18]. Ultimately, ICA is a critical building block for various types of social media analytics. Hence, future work that expands upon our study constitutes an important endeavor. Some possible extensions include expanding the feature set, improving upon the classification models, and applying such methods to other forms of social media. Nevertheless, this work represents an important first step towards enhanced modeling of social media interactions.

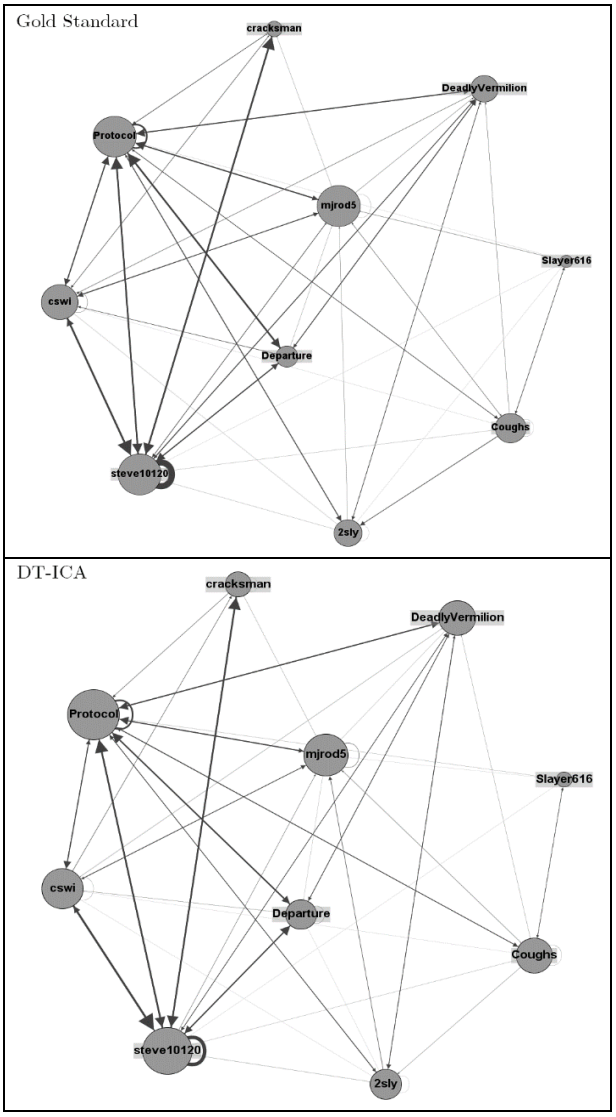


Figure 10. Betweenness Centrality MAPE for Top 25 Forum Members

VII ACKNOWLEDGEMENTS

We would like to thank the US National Science Foundation for their support of our work through the following grants: IIS-1236970, DUE-1303362, SES-1314631.

References

[1] C. Hale, Wired style: Principles of English usage in the Digital Age, San Francisco, CA: HardWired, 1996.

[2] K. S. Eklundh and H. Rodriguez, Coherence and interactivity in text-based group discussions around Web documents. In Proceedings of the 37th Annual Hawaii International Conference on System Sciences, pp. 40108.3. Washington, DC: IEEE Computer Society, 1999.

[3] T. Fu, A. Abbasi and H. Chen, "A Focused Crawler for Dark Web Forums," Journal of the American Society for Information Science and Technology, vol. 61, no. 6, pp. 1213-1231, 2008.

[4] D. Zimbra, A. Abbasi and H. Chen, "A Cyber-Archaeology Approach to Social Movement Research: Framework and Case Study," Journal of Computer-Mediated Communication, vol. 16, pp. 48-70, 2010.

[5] S. C. Herring, “Interactional coherence in CMC,” Journal of Computer-Mediated Communication, vol. 4, no. 4, 1999.

[6] J. Donath, K. Karahalios, and F. B. Viegas, Visualizing Conversation, In Proceedings of the 32nd Annual Hawaii international Conference on System Sciences, vol. 2, pp 2023, Washington, DC: IEEE Computer Society, 1999.

[7] T. Fu, A. Abbasi and H. Chen, "A Hybrid Approach to Web Forum Interactional Coherence Analysis," Journal of the American Society for Information Science and Technology, vol. 59, no. 8, pp. 1195-1209, 2008.

[8] C. Honey and S. C. Herring, “Beyond Microblogging: Conversation and collaboration via Twitter,” In Proceedings of the 37th Annual Hawaii International Conference on System Sciences, pp. 1-10, Washington, DC: IEEE Computer Society, 2009.

[9] A. Abbasi and H. Chen, “CyberGate: A Design Framework and System for Text Analysis of Computer-mediated Communication,” MIS Quarterly, vol. 32, no. 4, pp. 811-837, 2008.

[10] D. Comer and L. Peterson. “Conversation-based mail,” ACM Transactions on Computer Systems, vol. 4, no. 4, pp. 299–319, 1986.

[11] W. Sack, “Conversation map: An interface for very large-scale conversations,” Journal of Management Information Systems, vol. 17, no. 3, pp. 73–92, 2000.

[12] M. C. Nash, “Cohesion and reference in English chatroom discourse,” In Proceedings of the 38th Annual Hawaii International Conference on System Sciences, pp. 108.3, Washington, DC: IEEE Computer Society, 2005.

[13] R. Barzilay and M. Elhadad, “Using lexical chains for text summarization,” In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, pp. 10–17. Morristown, NJ: ACL, 1997.

[14] M. Elsner and E. Charniak “Disentangling chat,” Computational Linguistics, vol. 36, no. 3, pp. 389-409, 2010.

[15] T. Fu, A. Abbasi, D. Zeng and H. Chen. "Sentimental Spidering: Leveraging Opinion Information in Focused Crawlers," ACM Transactions on Information Systems, 30.4, article 24, 2012.

[16] T. Fu, A. Abbasi and H. Chen, "A Focused Crawler for Dark Web Forums," Journal of the American Society for Information Science and Technology, vol. 61, no. 6, pp. 1213-1231, 2008.

[17] D. Zimbra, A. Abbasi and H. Chen, "A Cyber-Archaeology Approach to Social Movement Research: Framework and Case Study," Journal of Computer-Mediated Communication, vol. 16, pp. 48-70, 2010.

[18] A. Abbasi, T. Fu, D. Zeng, and D. Adjeroh. "Crawling Credible Online Medical Sentiments for Social Intelligence." Proceedings of the ASE/IEEE International Conference on Social Computing, pp. 254-263, 2013.