Discovering Topic-Oriented Focal Sets in Cyber-Argumentation Using Link Analysis, Topic Modeling and Social Roles

Najla Althuniyan
The Computer Science Department
Prince Sultan University, Riyadh, Saudi Arabia
Email: nthuniyan@psu.edu.saAbstract

Abstract
Discussions have implicit topics and are often exchanged by users with specific profiles. Users may work cooperatively or collectively to support, attack, or deliver agendas due to similar interests. Much research has been conducted to detect groups with similar interests or specific characteristics using link analysis techniques, text analysis techniques, or a combination of different methods on social media and blogs. However, most of the research that has been published has focused on improving the community, or hidden community, detection algorithms concerning the research challenges. In this research, a framework is proposed to discover groups, or focal sets, with similar topic interests, using the focal structure analysis algorithm and topic modeling in cyber argumentation; then, the social roles of the focal set members are investigated. By combining these techniques, we discover and examine groups and individuals behind specific topics in the discussion. In cyber-argumentation discussions, we can analyze the group's and individuals' structures and profiles. This work can identify groups and individuals behind specific topics and use their characteristics to blend communities and individuals of polarized opinions in an online discussion. This allows for balancing the groups and individuals in a discussion and draws out the crowd's wisdom in cyber-argumentation platforms.

Keywords: topic modeling; focal sets; social roles; cyber-argumentation; community detection; discussion; topic-oriented

1. Introduction

According to Nielsen [1], users' participation in online social networks is unequal. In fact, in all large-scale user-generated content (UGC) platforms, user participation follows the 1-9-90 rule. Whereas 1% of users account for most of the content, 9% of users contribute from time to time, and 90% of users are lurkers who do not participate much in discussions. Therefore, there is a need to investigate users: who is talking, what they are talking about, and how they are connected to extract new knowledge from UGC platforms. Users do not work alone; they are associated with others directly, such as through social connections, or indirectly, such as through user interaction connections like a "reply-to" or "like" on social media platforms. Besides, users may work cooperatively or collectively to support, attack, or deliver agendas due to similar interests. These connections represent some similarities that can be used for community detection or group identification. These types of connections are studied extensively in research to detect communities or hidden communities with similar information or interests.

Community or hidden community detection can be performed using link analysis techniques, such as community detections or graph algorithms [2 - 11]; text analysis techniques, such as topic modeling [12, 13] or sentiment analysis [14]; or other data mining techniques, such as clustering, classification deep learning [15], using game-theoretic modeling [16] or using node attributes and edge structure [17]. Some studies have combined different techniques to discover or detect communities or hidden communities with certain similarities in social media and blogs. However, most of the work that has been conducted focused on improving the community or hidden community detection and structure concerning the research challenges. Furthermore, few researchers have tried to find representative users in the detected communities to analyze these users. To our knowledge, minimal work has been done on cyber-argumentation platforms. Mainly, community detection has been studied extensively in social media platforms such as Facebook and Twitter.

In this research, a new framework of existing work is proposed to discover topic-oriented focal sets using the Focal Structure Analysis Algorithm (FSA) [18] and the topic modeling in the Intelligent Cyber-Argumentation System platform (ICAS). First, topic-modeling techniques are performed at the issue level to identify most topics discussed by users under the parent issue. Second, the users' interaction information applies the FSA algorithm to identify sub-communities and focal structures. A pairwise similarity is then performed to measure the similarities between identified topics and community topics to find the users and focal sets behind specific topics. Finally, further experiments are conducted to study users' roles and the intensity of their opinion in

doi.org/10.31838/hiv23.03.43 Received: 16.10.22, Revised: 19.11.22, Accepted: 19.12.22.

online dissuasions. This work adds more significance to cyber-argumentation platforms by discovering communities or focal structures that are behind specific topics and discovering users' roles and opinion intensity.

This paper is organized as follows: Section II presents the background information used as a foundation for this research. Section III introduces argumentation systems, and the ICAS tool, and the dataset used in this research. Section IV demonstrates the proposed framework. Section V discusses the results and findings of the proposed framework. Finally, Section VI concludes and directs the future work for this research.

2. Background

Community Detection

Community detection [11], influential nodes [19, 20], and topic modeling [20] are well-studied topics in academia. Researchers have been focused on identifying hidden communities using different research methodologies [2, 3, 5, 7, 8, 10, 14] in the last few years. Tang et al. [2] used a novel integration scheme based on structural features to improve conventional community detection methods from one-dimensional to multi-dimensional networks. Hajdu et al. [3] discovered the passenger communities and most frequent trips in the transfer network using graph information and the community detection algorithm. Wang et al. [5] proposed community kernel detection to uncover the hidden community structure in large social networks and discover influential users. He et al. [8] presented a new approach: HIdden COmmunity Detection (HICODE), which identifies hidden communities and dominant community structures by weakening the strength of the dominant structure to uncover the hidden community structure beneath. Prem and Blei [21] used a Bayesian model of networks that allows communities t overlap. A corresponding algorithm naturally interleaves subsampling from the network and updates an estimate of its communities in massive networks. Peng et al. [7] developed an unsupervised learning method to discover implicit communities hidden in tweet datasets. Fortunato [6] has studied well community detection algorithms in graphs. Fortunato and Barthelemy [11] found that modularity optimization might fail to identify modules with smaller communities compared to the massive network size. Lin et al. [22] used a distant-based modularity method for community detection in incomplete network information to discover hierarchical and overlapped communities. Behera et al. [23] used a parallel programming framework to reduce running time for uncovering the hidden communities in a social network.

All the research mentioned above has focused on link-based community or hidden community detection. However, some work has been done through community detection using additional preprocessing or additional techniques such as text analysis. For example, Fu et al. [14] used topic identification to identify the target participants, then applied sentiment analysis and opinion mining for users with similar topic interests. Finally, they applied multi-level community detection to find communities constructed by the users who have a consistent opinion. Zhao et al. [13] addressed the semantics problem shared by people in community detection by introducing topical clustering as an additional step to strengthen or weaken community connections. Dang and Nguyen [24] proposed a new approach for topic modeling using community findings in dynamic networks. They used topic modeling to refine the network based on the topics to reveal the structures and communities in dynamic social networks. Abdelbary [15] proposed multi-layer community detection by applying Gaussian Restricted Boltzmann Machine on users' posts to identify their topics of interest and then construct communities.

Influential nodes are significant in different contexts. Two well-known algorithms are used for influential nodes in graphs. The PageRank Algorithm [25] counts the number and the weight of the links to a node in a graph. The underlying assumption of the PageRank algorithm is the more critical the node in the graph, the more likely to receive more links over time. On the other hand, the HITS Algorithm [26] assigns two scores for each node: authority, which estimates the value of the node's content, and hub, which estimates the value of the links of that node to the other nodes. Both algorithms have been used, optimized, and improved in many contexts. For example, Chen et al. [27] proposed a new metric to identify influential nodes in a network by trading-off between the low-relevant degree centrality and other time-consuming measures. Kempe et al. [28] implemented the Decreasing Cascade Model to choose the active set of nodes for behavior spreading. Wang et al. [29] proposed a new algorithm that detects communities in social networks considering information diffusion and a dynamic programming algorithm for selecting communities to find influential nodes.

Sen et al. [18] developed the FSA algorithm that detects disease-release structure in a network context. It uses the Protein-Protein Interaction (PPI) networks to identify smaller and more relevant focal structures instead of identifying large clusters or communities by applying the modularity algorithm [30] recursively. This algorithm has been used to determine the focal sets in organizing mass protests on social media [31]. Unlike the traditional influential finding algorithms, this algorithm was able to identify a set of influential nodes in a network that forms a compelling power. In this research, we use the community as a group or focal set to reference a collection of individuals who shared links in the argumentation graph.

Topic Modeling

Topic modeling and sentiment analysis attract researchers' attention due to the massive text

generated by UGC, which helps to extract crowd wisdom. LDA [32], Latent Dirichlet Allocation, is a well-known algorithm in the topic modeling field. It is designed as a multi-level Bayesian to model items as a collection of a finite mixture over an underlying set of topics. Xu et al. [33] used non-negative matrix factorization for document clustering in a given corpus. These algorithms are used and improved enormously in academia to fulfill research needs. For example, Jelodar et al. [34] reviewed scholarly articles published from 2003 to 2016 related to LDA-based topic modeling to discover the research development, current trends, and intellectual structure of topic modeling. Debortoli et al. [35] combined LDA with a logistic regression model to explain user satisfaction with an IT artifact by analyzing more than 12,000 online customer reviews.

Social Roles in UGC Platforms

Users in UGC platforms hold different roles; users may switch roles based on their contributions. Other researchers have studied and investigated the social roles of users in various domains. For example, Stuetzer et al. [39] have studied brokering behavior in online learning communities by examining role patterns and information between learners and educators using social network analysis. Chan et al. [40] used nine different role features to profile the user roles in discussion forums. Then, they used two-stage clustering to describe the forums based on their role composition. White et al. [38] used a mixed membership formulation to cluster users with similar egocentric network structures based on the profile models' network statistics. Welser et al. [39] used editing patterns and egocentric network visualizations to develop "structural signatures" as quantitative indicators of role adoption. In another similar domain, Wikipedia talk pages are community-oriented pages. Gleave et al. [40] have standardized the "social role" in online communities as a blend of psychological, social structural, and behavioral properties. They measured and analyzed strategies for identifying social roles in Wikipedia and Usenet. Fisher et al. [41] have used social network analysis or SNA to characterize authors in Usenet newsgroups. They found that second-degree egocentric networks provide apparent differences between different types of authors and newsgroups. On another platform, Reddit.com, Buntain, and Golbeck [42] confirmed the existence of the "Answer-Person" role in Reddit and provided an automated method for identifying this role based on user interaction. Mantzaris and Higham [43] have proposed a new model that quantifies nodes' communication data in social networks, known as dynamic communicators, using standard centrality measures.

Social roles have been studied in the knowledge management domain. For example, Davidson et al. [44] have examined different roles in online communities and developed Reader to Leader framework to utilize online forum users' role evolution. Cranefield et al. [45] have studied lurking behavior in online communities. They found vital key roles in transferring knowledge in two activities: monitoring the knowledge agenda and monitoring or monitoring. Akar et al. [46] identified user roles in an online community using structural role theory, SNA, and community members' contribution behavior. Finally, social roles have been used in Enterprise Social Networks (ESN). Hacker et al. [47] have determined knowledge actions and knowledge worker roles to characterize ESN user behavior.

3. Argumentation Systems

Cyber-argumentation systems are mudiums to deal with contentious debates and deliberation from which they conclude. Most of these platforms are built on formal and informal argumentation frameworks. The formal argumentation framework is an abstract argumentation framework created by Dung [48], known as the Dung Abstract Framework. It defines argumentation systems as a set of arguments and defeasibility relations. The system is viewed as an oriented graph whose nodes are the different arguments, and the edges represent the defeasibility relationship between them. The formal argumentation platform was the fundamental framework for many informal argumentation frameworks, such as logic-based argumentation frameworks, value-based argumentation frameworks, and assumption-based argumentation (ABA) frameworks. However, the Issue-Based Information System by Kunz and Rittel [49], known as IBIS, was invented earlier than Dung's argumentation framework. It has been used widely as an informal argumentation framework to coordinate and solve multiple stakeholders' problems. Many argumentation tools have been built using IBIS.

Cyber-Argumentation platforms are capable of allowing vast discourse between participants and understanding the discussion. They address issues by creating well-defined structures for deliberation. Therefore, online-argumentation platforms are loaded with valuable hidden features worth further study and research due to the extensive discussion they contain. They have exhibited the ability to evaluate the discussion on large-scale platforms and in different contexts. They are capable of identifying groupthink [50], analyzing argument credibility [51], measuring polarization in opinions [52], recommending friendship connections using opinion diversity [53], and predicting the missing collective opinion [54]. Cyber-argumentation platforms enable researchers to extract actionable knowledge from deliberations. This process is known as "Knowledge Discovery" [18] in complex social networks.

Liu et al. [55] implemented an intelligent collaborative system for collaborative engineering design and conflict resolution. This system has been updated and developed over time for many settings and uses. Then, this system has been updated and expanded over time to accommodate different research goals. The current version is the Intelligent Cyber Argumentation System (ICAS).

ICAS

ICAS is an online web-based argumentation tool. The structure of ICAS exhibits the IBIS structure. Each issue and related nodes are modeled into a single tree, as shown in Fig 1.

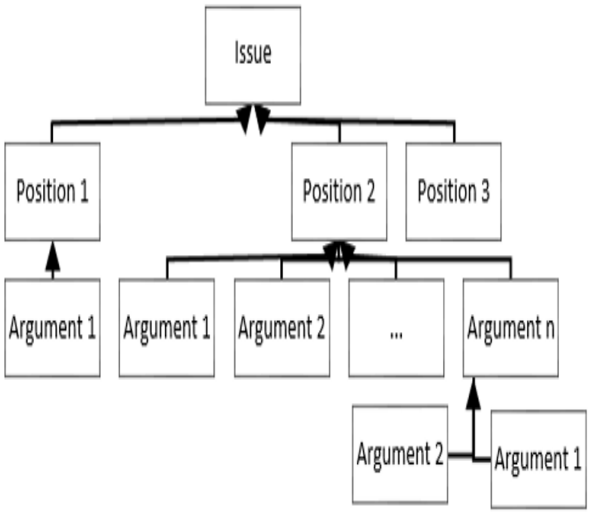

Fig.1. Issue, Position and Argument Tree

To understand the discussion's direction and participants' attitudes, users need to read every argument in the issue tree. This is time-consuming and most likely, users cannot comprehend the whole discussion while it grows massively over time. Therefore, the ICAS has incorporated three analytics models to make the app intelligent and informative. The models are the collective intelligent index by [55], the polarization index by [52], and the prediction of collective opinions on the position level by [54]. These models use analytical techniques to help users understand the discussion, even without participating. For more information about ICAS, please refer to [56].

Empirical Data Collection

For more information about the issues and positions used in this research, please refer to [57].

4. Proposed Framework

Discussions in cyber-argumentation contain more than text and opinions. This proposed framework uses the FSA algorithm to discover topic-oriented hidden communities under a selected issue.

This framework is divided into critical modules, as shown in Figure 2. First, the data is collected from the ICAS platform. Then, topic modeling is performed on the text collected and focal set identifications on the participants using link analysis. Subsequently, a pairwise similarity is conducted between the identified topics, individuals, and focal sets' posts. Finally, the participants' social roles and opinion intensity are used to analyze the discussion dynamic in ICAS discussions. In the following subsections, more information about the technical details of the framework.

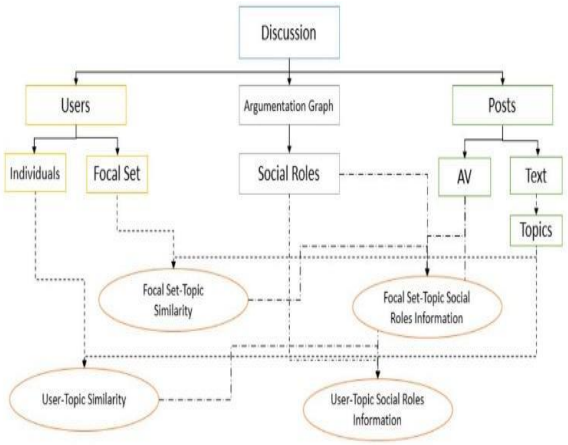

Fig 2. Proposed Framework

User Opinion Vector (UOV)

In ICAS, an issue, as shown in Fig. 1, consists of positions and arguments at different levels of the tree. To derive the user opinion vector, first, any argument or reaction made at any tree level needs to be a direct child to its position parent. The fuzzy reduction engine from Liu et al. [55] is used to reduce all arguments and reactions from any level of the argumentation tree to become direct children to the parent position. Then, all the reduced agreement values (AV) form all the arguments, and the reactions per position for user u are averaged. Suppose there are n positions under issue I; then, the same process is applied for all positions under the selected issue. If user u has not participated in one or more positions, those missing values are imputed with zeros. Finally, the user vector opinion is built and created as:

$$UOV(u, I) = (AV_{p1}^{u}, AV_{p2}^{u}, AV_{p3}^{u}, , , AV_{pn}^{u}) \quad (1)$$

Focal Sets and Participants Partitioning Using Link Analysis

In ICAS, it is assumed that participants do not have any explicit relationships or social connections during the discussion. It is also assumed that users work with each other unintentionally or collectively. With these assumptions, some users share similar interests and behaviors. The FSA algorithm [18] is applied to ICAS discussions to discover sub-communities and focal structures in this section. However, as shown in figure 2-1, the issue tree represents the discussion in ICAS. Therefore, a graph transformation is needed to apply the FSA algorithms. The user interaction graph (UIG) is created from the discussion tree as in [57]. The resulting graph consists of nodes representing users in the discussion and edges representing reply-to and react-to relationships between users. The FSA algorithm is applied to partition the UIG into smaller graphs as the graph is created. As a result, different communities are formed at different levels of the UIG.

Focal Sets Opinion Vector

This section studies the users' opinions within the

focal set. Users who share the same topic do not necessarily share the same opinion about the issue. Therefore, it needs more processing to investigate users' opinions within the same focal set. Two concepts are introduced for this purpose: Average User Opinion and Average Focal Set Opinion—details in the following two subsections.

Average User Opinion (AUO)

On the issue level, to determine the average user's opinion, users' opinions for all positions under the issue are considered, as follows:

$$AUO(I) = \left(\frac{\sum_{i=1}^{n} AV_{i_{p1}}}{m}, \frac{\sum_{i=1}^{n} AV_{i_{p2}}}{m}, \frac{\sum_{i=1}^{n} AV_{i_{p3}}}{m}, , , \underline{\sum_{i=1}^{n}} \right) \quad (2)$$

In (2), m is the number of participants in issue I, j is the total number of positions for I and AVip1 is the average user i opinion on position p1.

The average user's opinion is the same size as the opinion vector. Therefore, we can compare each user with AUO for a particular issue.

Focal Set Average Opinion (FSAO)

On the issue-level, to determine the focal set average opinion for a particular focal set, we need to consider users' opinions for all positions under the issue within the same community.

$$FSAO(C) = \left(\frac{\sum_{i=1}^{n} AV_{i_{p1}}}{n}, \frac{\sum_{i=1}^{n} AV_{i_{p2}}}{n}, \frac{\sum_{i=1}^{n}}{} \right) \quad (3)$$

In (3), n is the number of participants in the focal set C, and j is the total number of positions for I. AVip1 is the average user opinion i in focal set C and on position p1.

The focal set average opinion has the same size as the user opinion vector. Therefore, FSAO with AUO can be compared for a particular issue as the norm.

Issue Topics by Topic Molding

Corpus Preparing

Although an issue has explicit positions, it may have other implicit topics or positions as the discussion goes on between users. Moreover, in the ICAS design, positions do not have to be distinct mutually exclusive positions; they overlap or are alternatives to other positions. Thus, there is a need to identify the main topics discussed under an issue. However, the corpus needs to be prepared and preprocessed for topic modeling. Therefore, each user's arguments are combined under the selected issue, tokenized and stop words, low-frequency words are removed, and finally, stemmed. Each user has a single document of their processed text. Finally, the corpus is made up of all the users' documents.

Topic modeling is done in different ways. A well-known algorithm is Latent Dirichlet Allocation (LDA) by Beli [32]. Another topic modeling method uses non-negative matrix factorization (NMF) [33]. The following subsections introduce both methods in the proposed framework.

Topic Modeling by LDA

LDA is a generative probabilistic model of a corpus [32]. It tells us that each word (w) in each document (d) comes from a topic (t), and the topic is determined from a per-document distribution across topics. It uses two probability values:

The probability distribution of words in topics: Φwt= P(w|d).

The probability distribution of topics in documents: ϴtd = P(t|d) .

The probability of a word in a document is

$$P(w|d) = \sum_{t \in T} p(t,d)\, p(t|d) \quad (4)$$

Where T is the total number of topics. If conditional independence is assumed, then:

$$P(w|t,d) = P(w|t) \quad (5)$$

Therefore,

$$P(w|d) = \sum_{t=1}^{T} p(t)p(t|d) \quad (6)$$

The generative process of the LDA model can be described as the joint probability distribution, and the likelihood of generating the whole corpus (D) is:

$$P(D|\alpha,\beta) = \prod_{d=1}^{M} p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{t_{d,n}} p(t_{d,n}|\theta_d)\, p(w_{d,n}|t_{d,} \right) \quad (7)$$

Where α, β are the parameters of the respective Dirichlet distributions, θd is the topic proportion for document d, td,n is the topic assignment for word n in document d, and wdn is the observed word in document d.

Topic modeling by NMF

Non-negative Matrix Factorization is a dimension reduction method that factors high-dimensional vectors into a low-dimensionality representation. If a corpus D consists of n words and m documents, the D matrix can be factored into matrix W (n words and k topics) and matrix H (k topics and m documents). Topic modeling by NMF usually requires a corpus D normalized by TF-IDF. W and H matrices are optimized over an objective function as

$$\frac{1}{2}\|A - WH\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} (A_{ij} - (WH)_{ij})^2 \quad (8)$$

Using the objective function, the update rules for matrices W and H are:

$$W_{ic} \leftarrow W_{ic} \frac{(AH)_{ic}}{(WHH)_{ic}} \quad (9)$$

$$H_{cj} \leftarrow H_{cj} \frac{(WA)_{cj}}{(WWH)_{cj}} \quad (10)$$

The updated values are calculated in parallel operations. Then, the new matrices, W and H, are used to calculate the reconstruction error. The process is repeated until convergence.

Individuals and Focal Sets and Topics Similarities
A pairwise similarity is used between identified topics and individuals and focal sets' text to measure the contributions from individuals and focal sets' text. The similarity between individual texts and identified topics is a simple calculation task. However, the similarity between the focal set's text and identified topics is not a straightforward calculation. It requires more processing. Thus, algorithm 1determines the focal set contribution to each topic.
Finally, the focal sets are sorted based on their similarity score in decreasing order for each identified topic. Setting a similarity threshold can identify which communities have contributed the most to a particular topic.

Algorithm 1 Focal Set - Topic Similarity

Input: list of focal sets and list of topics
Output: matrix of the similarity between focal sets content and topics
For each focal set:
For each topic:
For each user in the focal set:
User contribution = pairwise similarity between user content and topic
focal set contribution += User contribution
focal set topic similarity = focal set contribution / focal set size
Social Roles in ICAS
Most research on social roles [37 - 47] has used graph information to identify participants' social roles. They have mainly used the node position in a graph where the node in-degree and out-degree information determines the participant roles. The number and description of roles in the research [37 - 47] have varied concerning the associated platform. Since ICAS is a discussion and argumentation platform, the UIG from [57] is used. In addition, the node attributes from the UIG are used to determine the user's role in the discussion.
[40-41] have identified three social role signatures based on the egocentric networks for participants: Answer Person, Discussion Person, and Discussion Catalyst. An "Answer Person" does more posting than receiving in a discussion. Therefore, the out-degree for this user in a graph is much higher than their in-degree. A "Discussion Person" has frequent reciprocal exchanges with other participants. As a result, this user has similar in-degree and out-degree scores. Finally, the "Discussion Catalyst" posts messages that initiate long threads. This role may not have higher out-degree scores. Users with this role do not have more postings than others, but they are likely to attract others to engage in their threads. In this research, the abovementioned roles are used to label users in discussions.

5. Discussion and Results

The FSA Algorithm with ICAS Discussions

The FSA algorithm is used as a link-based algorithm to partition users in each discussion into focal sets. Table II shows the results of the FSA algorithm applied to the UIG for each discussion hosted by ICAS.
Issue 1 has the highest number of participants and connections. The FSA algorithm looped ten times and resulted in 85 focal sets. The UIG was sparse due to the high number of users participating in this discussion. As a result, the UIG had the lowest modularity, cluster coefficient, and transitivity scores. Issue 2 has 79 focal sets from applying the FSA algorithm within nine rounds. However, this issue reported the highest modularity, cluster coefficient, and transitivity scores. This issue has a denser UIG graph than the other issues' UIGs. As shown in Table II, issue 3 and Issue 4 are in between Issue 1 and Issue 2.

Topic Modeling with ICAS Discussion

LDA and NMF models are used for topic modeling on the issue level. Table III reports the evaluation result from applying both models.

Table II. Results From Applying the FSA Algorithm

Issue	# Users	# Connections	# Focal Sets	# Levels	Modularity	Clustering Coefficient	Transitivity
Issue 1	305	4701	83	10	0.609477	0.097111	0.087107
Issue 2	291	3412	79	9	0.681429	0.169482	0.196258
Issue 3	297	3696	84	10	0.633422035	0.139165273	0.130571023
Issue 4	280	3198	88	11	0.635557726	0.097178521	0.17741009

In each issue, the number of topics was set to 4, with ten tokens in each topic. For both models, in all issues, the topics were primarily different. With the limited corpus from each issue discussion, NMF reported better topic results, evaluated by pointwise mutual information, PMI, than LDA. This result is that NMF normalizes the corpus using the TF-IDF. However, it required more run time than the LDA model.
Algorithm 1 is used for the focal sets to find the similarities between topics and focal set documents. Due to the space limitation, the results of the focal sets documents and topic similarities are not included in this dissertation, but they are available upon request. However, one significant finding in this area is that the similarity between topics and focal sets documents increases as focal set size decreases. This correlation is because some users contribute to the discussion but without any text, which negatively affects the topic similarity.

Individual Opinion Intensity and Social Roles

The intensity measure usually is split into three levels: low, medium, and high. In this research, the same intensity levels are used to determine user opinion intensity. Each user opinion vector (UOV) is

calculated as (1) and then is compared with AUO as (2) to label the user. This distance between UOV and AUO determines the user opinion intensity level, as shown in fig. 3. For each user, if the UOV is within 0.33 distance or less from the AUO, then this user opinion is labeled as a low-intensity opinion. If the UOV is within a distance between 0.34 and 0.67 from the AUO, then this user opinion is labeled as a medium-intensity opinion. If the UOV has a distance of more than 0.67 from the AUO, the user's opinion is labeled as high-intensity.

Topic Modeling with ICAS discussions

LDA and NMF models are used for topic modeling on the issue level. Table III reports the evaluation result from both models.

To investigate the user's social role, the egocentric network information is used to label the user as Answer Person, Discussion Person, or Discussion Catalyst, as previously mentioned in section IV

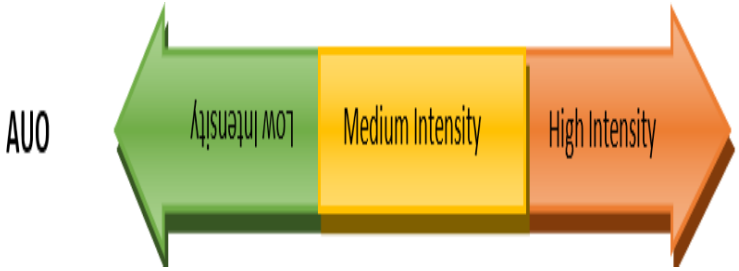

Fig 3. Opinion Intensity Bar

Table III. Topic Modeling Evaluation		
Issues	**PMI - NMF**	**PMI - LDA**
Issue 1	0.05277667	-0.0382568
Issue 2	0.011041	-0.007667
Issue 3	0.1094469	0.021647
Issue 4	0.0495284	-0.005781

Table IV. Statistics of Individual Social Roles and Opinion Intensity in ICAS Discussions				
Issues	**Social Roles**	**Low-Intensity**	**Medium-Intensity**	**High-Intensity**
Issue1	Answer Person	35	56	9
	Discussion Person	29	104	24
	Discussion Catalyst	11	22	11
Issue 2	Answer Person	40	45	8
	Discussion Person	53	101	19
	Discussion Catalyst	6	25	3
Issue 3	Answer Person	21	55	11
	Discussion Person	22	121	34
	Discussion Catalyst	3	23	2
Issue 4	Answer Person	35	47	7
	Discussion Person	49	92	21
	Discussion Catalyst	3	12	10

different pattern. For Issue 1, users with the opinion of "High-Intensity" occur equally to users with the opinion of "Low-Intensity." However, in Issue 4, users with the opinion of "High-Intensity" are more than users with the opinion of "Low-Intensity," while for Issue 2 and Issue 3; users with the opinion of "Low-Intensity" occur more than users with the opinion of "High-Intensity."

The total number of participants across all discussions is 310 users. Of these users, 5% of the participants are users with an "Answer Person" role across all discussions, 25% of the participants have a "Discussion Person" role across all discussions, and only one user found who had the "Discussion Catalyst" across all discussions. On the other hand, users seem to switch social roles among discussions. For example, 72% of the participants alternated between the "Answer person" and "Discussion Person" roles across all discussions, and 38% of the participants had either "Discussion Person" or "Discussion Catalyst" roles across all discussions. However, no user was found with the "Answer person" and "Discussion Catalyst" roles in all discussions. These statistics show that users may switch roles from the "Answer person" role to the "Discussion Person" role or from the "Discussion Person" role or "Discussion Catalyst" role but not from the "Answer person" role to the "Discussion Catalyst" role according to the user behavioral postings.

The PageRank Algorithm [25] has been applied to each issue discussion. The users were sorted in decreasing order based on their PageRank score. Of the top 10%, most of the users were labeled as "Discussion Catalyst," and a few of them were labeled as "Discussion Person." Moreover, according to the PageRank Algorithm scores, two users were influential in all discussions, six were found in three discussions, and twenty-three were found in two ICAS discussions.

Similarly, opinion intensity degree dynamics for users have an identical pattern to the user's social role

dynamics. 30% of the participants have an opinion with a "Low-Intensity" degree, 14% of the participants have an opinion with a "Medium-Intensity" degree, and only one participant has an opinion with a "High-Intensity" degree across all issues. Of the total users, 50% either have opinions of "Low-Intensity" or "Medium-Intensity" degree, 40% have opinions of "High-Intensity" or "Medium-Intensity" degree, and 1% have opinions of "Low-Intensity" or "High-Intensity" degree across all discussions. Like social role dynamics, users may change their opinion intensity degree from "Low-Intensity" to "Medium-Intensity" or "Medium-Intensity" to "High-Intensity" degree but rarely from "Low-Intensity" to "High-Intensity" degree.

Focal Sets Opinion Intensity and Social Roles

Like individual opinion intensity, the focal set opinion intensity is calculated using (3) and then is compared with AUO as (2) to label the focal sets as in the previous section. Table IV shows us the focal set labels in all ICAS discussions. Most of the focal sets in all issue discussions have an opinion with a "Low-Intensity" degree; fewer focal sets have an opinion with a "Medium-Intensity" degree and no focal sets with a "High-Intensity degree." The opinion intensity degree reduces with the users being grouped in focal sets.

The focal sets found in all ICAS discussions have different dynamics. Users with the "Answer Person" role appeared in 80% of the focal set members and made up 34%. Participants with the "Discussion Person" role occurred in 92% of the focal sets and made up 53% of the focal set members. However, users with the "Discussion Catalyst" role appeared in 55% of the focal sets and made up 12% of the focal set members. At least one user with a "Discussion Person" or "Discussion Catalyst" role exists in each identified focal set. We found no focal set where all the members were labeled with the "Answer Person" role.

Table V. StatiSTICs of Focal Set Opinion Intensity in ICAS DisCUssionS

Issues	Low-Intensity	Medium-Intensity	High-Intensity
Issue1	71	12	0
Issue 2	44	35	0
Issue 3	66	17	0
Issue 4	73	13	0

As per the opinion intensity in focal sets, users with "High intensity" opinions appeared in 39% of the focal set and made up 8% of the focal set members. Participants with "Medium-intensity" opinions occurred in 99% of the focal set and made up 75% of the focal set members. Finally, users with "Low-intensity" opinions appeared in 58% of the focal set and made up 17% of the focal set members.

FSA algorithm and topic modeling techniques are valuable tools to extract knowledge within their domain. Combining these techniques lead to more helpful information from UGC platforms. They help discover and profile hidden communities and users based on their discussions and contribution. These algorithms help identify the social roles and pinion intensity of participants in cyber-argumentation platforms.

6. Conclusion

Identifying topic-oriented communities is an essential task in online discussion platforms. Moreover, it is necessary to measure the opinion intensity and identify the social roles of users who make up the focal set to understand discussions and participants better. A new framework is proposed in this research. This framework can measure the users' opinion in a discussion and compare it to the average user's opinion vector and identify focal sets using the FSA algorithm and measure their opinion intensity degree. Additionally, the framework applies a pairwise similarity between topics discussed by individuals and focal sets to leverage topic-oriented individuals and focal sets in cyber-argumentation. Finally, it analyzed the individuals and focal set dynamics using users' opinions and social roles.

This framework can identify similar focal sets and individuals who are behind specific topics but not connected. It can also blend communities and individuals of polarized opinions in an online discussion. This balances the focal sets and individuals in a discussion and draws out the crowd's wisdom in the cyber-argumentation platform.

There are many options to expand this model. For example, since the individual and the focal set opinion intensity degree are calculated, is it possible to calculate the topic intensity degree according to the participants' attitudes? If yes, how do we increase or decrease the topic intensity degree? Another task that could be done is to investigate non-textual users who appeared in multiple focal sets but have not contributed to the discussions with replies. This is left for future work.

References

Nielsen, J., 2014. Participation Inequality: Lurkers vs. Contributors in Internet Communities, (URL: http://www.nngroup.com/articles/participation-inequality/. Accessed: 2014-06-05. (Archived by WebCite® at http://www.webcitation.org/6Q7EwEncA).

Tang, L., Wang, X., & Liu, H. (2012). Community detection via heterogeneous interaction analysis. Data mining and knowledge discovery, 25(1), 1-33.

Hajdu, L., Bóta, A., Krész, M., Khani, A., & Gardner, L. M. (2019). Discovering the hidden community structure of public transportation networks.

Networks and Spatial Economics, 1-23.
Gopalan, P. K., & Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences, 110(36), 14534-14539.
Wang, L., Lou, T., Tang, J., & Hopcroft, J. E. (2011, December). Detecting community kernels in large social networks. In 2011 IEEE 11th International Conference on Data Mining (pp. 784-793). IEEE.
Fortunato, S. (2010). Community detection in graphs. Physics reports, 486(3-5), 75-174.
Peng, D., Lei, X., & Huang, T. (2015). DICH: A framework for discovering implicit communities hidden in tweets. World Wide Web, 18(4), 795-818.
He, K., Li, Y., Soundarajan, S., & Hopcroft, J. E. (2018). Hidden community detection in social networks. Information Sciences, 425, 92-106.
Behera, R. K., Rath, S. K., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2017). Large scale community detection using a small world model. Applied Sciences, 7(11), 1173.
He, K., Soundarajan, S., Cao, X., Hopcroft, J., & Huang, M. (2015). Revealing multiple layers of hidden community structure in networks. arXiv preprint arXiv:1501.05700.
Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. Proceedings of the national academy of sciences, 104(1), 36-41.
Dang, T., & Vinh the Nguyen. (2018, June). ComModeler: Topic Modeling Using Community Detection. In EuroVA@ EuroVis (pp. 1-5).
Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., & Fan, J. (2012). Topic oriented community detection through social objects and link analysis in social networks. Knowledge-Based Systems, 26, 164-173.
Fu, M. H., Peng, C. H., Kuo, Y. H., & Lee, K. R. (2012, June). Hidden community detection based on microblog by opinion-consistent analysis. In International Conference on Information Society (i-Society 2012) (pp. 83-88). IEEE.
Abdelbary, H. A., ElKorany, A. M., & Bahgat, R. (2014, August). Utilizing deep learning for content-based community detection. In 2014 Science and Information Conference (pp. 777-784). IEEE.
Chopade, P., & Zhan, J. (2016). A framework for community detection in large networks using game-theoretic modeling. IEEE Transactions on Big Data, 3(3), 276-288.
Chopade, P., Zhan, J., & Bikdash, M. (2015, April). Node attributes and edge structure for large-scale big data network analytics and community detection. In 2015 IEEE International Symposium on Technologies for Homeland Security (HST) (pp. 1-8). IEEE.
Şen, F., Wigand, R. T., Agarwal, N., Mete, M., & Kasprzyk, R. (2014, June). Focal structure analysis in large biological networks. In 3rd International Conference on Environment, Energy and Biotechnology (ICEEB 2014).
Chen, D., Lü, L., Shang, M. S., Zhang, Y. C., & Zhou, T. (2012). Identifying influential nodes in complex networks. Physica a: Statistical mechanics and its applications, 391(4), 1777-1787.
Kempe, D., Kleinberg, J., & Tardos, É. (2005, July). Influential nodes in a diffusion model for social networks. In International Colloquium on Automata, Languages, and Programming (pp. 1127-1138). Springer, Berlin, Heidelberg.
Gopalan, P. K., & Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. Proceedings of the National Academy of Sciences, 110(36), 14534-14539.
Lin, W., Kong, X., Yu, P. S., Wu, Q., Jia, Y., & Li, C. (2012, April). Community detection in incomplete information networks. In Proceedings of the 21st international conference on World Wide Web (pp. 341-350).
Behera, R. K., Rath, S. K., Misra, S., Damaševičius, R., & Maskeliūnas, R. (2017). Large scale community detection using a small world model. Applied Sciences, 7(11), 1173.
Dang, T., & Vinh the Nguyen. (2018, June). ComModeler: Topic Modeling Using Community Detection. In EuroVA@ EuroVis (pp. 1-5).
age, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Stanford InfoLab.
PKleinberg, J. M. (1999). Hubs, authorities, and communities. ACM computing surveys (CSUR), 31(4es), 5-es.
Chen, D., Lü, L., Shang, M. S., Zhang, Y. C., & Zhou, T. (2012). Identifying influential nodes in complex networks. Physica a: Statistical mechanics and its applications, 391(4), 1777-1787.
Kempe, D., Kleinberg, J., & Tardos, É. (2005, July). Influential nodes in a diffusion model for social networks. In International Colloquium on Automata, Languages, and Programming (pp. 1127-1138). Springer, Berlin, Heidelberg.
Wang, Y., Cong, G., Song, G., & Xie, K. (2010, July). Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1039-1048).
Newman, M. E. (2006). Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23), 8577-8582.
Şen, F., Wigand, R., Agarwal, N., Tokdemir, S., & Kasprzyk, R. (2016). Focal structures analysis: Identifying influential sets of individuals in a social network. Social Network Analysis and Mining, 6(1), 17.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.
Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (pp. 267-273).
Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation

(LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.
Debortoli, S., Müller, O., Junglas, I., & vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. Communications of the Association for Information Systems, 39(1), 7.
Stuetzer, C. M., Koehler, T., Carley, K. M., & Thiem, G. (2013). "Brokering" behavior in collaborative learning systems. Procedia-Social and Behavioral Sciences, 100, 94-107.
Chan, J., Hayes, C., & Daly, E. M. (2010, May). Decomposing discussion forums and boards using user roles. In Fourth International AAAI Conference on Weblogs and social media.
White, A. J., Chan, J., Hayes, C., & Murphy, B. (2012, May). Mixed membership models for exploring user roles in online fora. In Sixth International AAAI Conference on Weblogs and social media.
Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In Proceedings of the 2011 iConference (pp. 122-129).
Gleave, E., Welser, H. T., Lento, T. M., & Smith, M. A. (2009, January). A conceptual and operational definition of 'social role' in online community. In 2009 42nd Hawaii International Conference on System Sciences (pp. 1-11). IEEE.
Fisher, D., Smith, M., & Welser, H. T. (2006, January). You are who you talk to: Detecting roles in usenet newsgroups. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06) (Vol. 3, pp. 59b-59b). IEEE.
Buntain, C., & Golbeck, J. (2014, April). Identifying social roles in reddit using network structure. In Proceedings of the 23rd international conference on world wide web (pp. 615-620).
Mantzaris, A. V., & Higham, D. J. (2012). A model for dynamic communicators. European Journal of Applied Mathematics, 23(6), 659-668.
Davidson, B. I., Jones, S. L., Joinson, A. N., & Hinds, J. (2019). The evolution of online ideological communities. PloS one, 14(5).
Cranefield, J., Yoong, P., & Huff, S. L. (2015). Rethinking lurking: Invisible leading and following in a knowledge transfer ecosystem. Journal of the Association for Information Systems, 16(4), 213.
Akar, E., & Mardikyan, S., & Dalgic, T. (2018). "User Roles in Online Communities and Their Moderating Effect on Online Community Usage Intention: An Integrated Approach", International Journal of Human-Computer Interaction, 45(6), 495-509.
Hacker, J., Bodendorf, F., & Lorenz, P. (2017). Helper, Sharer or Seeker?-A Concept to Determine Knowledge Worker Roles in Enterprise Social Networks.
Dung, P.M., 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial intelligence 77, 321–357.
Kunz, W., Rittel, H., 1970. Issues as elements of information systems (vol. 131). Berkeley, California: Institute of Urban and Regional Development, University of California.
M. Klein, "The CATALYST Deliberation Analytics Server," 2015.
R. Arvapally, and XF. Liu, "Analyzing credibility of arguments in a web-based intelligent argumentation system for collective decision support based on K-means clustering algorithm," in Knowledge Management Research & Practice, vol. 10, no. 4, pp. 326-241, December 2012.
Sirrianni, J., Liu, X., Adams, D., 2018. Quantitative modeling of polarization in online intelligent argumentation and deliberation for capturing collective intelligence, in: 2018 IEEE International Conference on Cognitive Computing (ICCC), IEEE. pp. 57–64.
J. Sirrianni, X. F. Liu, M. M. Rahman, and D. Adams, "An Opinion Diversity Enhanced Social Connection Recommendation Re-ranking Method based on Opinion Distance in Cyber Argumentation with Social Networking," in 2019 IEEE International Conference on Cognitive Computing (ICCC), Milan, Italy, 2019.
Rahman, M.M., Sirrianni, J. W.and Liu, X.F., Adams, D., 2019. Predicting opinions across multiple issues in large-scale cyber argumentation using collaborative filtering and viewpoint correlation. The Ninth International Conference on social media Technologies, Communication, and Informatics.
Liu, X.F., Raorane, S., Leu, M.C., 2007. A web-based intelligent collaborative system for engineering design, in: Collaborative product design and manufacturing methodologies and applications. Springer, pp. 37–58.
Althuniyan, N., Sirrianni, J. W., & Rahman, M. M. (2019, June). Design of Mobile Service of Intelligent Large-Scale Cyber Argumentation for Analysis and Prediction of Collective Opinions. In International Conference on AI and Mobile Services (pp. 135-149). Springer, Cham.
Najla Althuniya, Xiaoqing Liu, Joseph W. Sirrianni, and Douglas Adams, "Opinion Discovery Framework: Toward a Quality Opinion-Centric Platform," Journal of Advances in Information Technology, Vol. 11, No. 2, pp. 48-57, May 2020