Facilitating Efficient Information Seeking in Social Media

by

Suhas Ranganath

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved on October 2017 by the
Graduate Supervisory Committee:

Huan Liu, Co-Chair
Ying-Cheng Lai, Co-Chair
Hanghang Tong
Roman Vaculin

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

Online social media is popular due to its real-time nature, extensive connectivity and a large user base. This motivates users to employ social media for seeking information by reaching out to their large number of social connections. Information seeking can manifest in the form of requests for personal and time-critical information or gathering perspectives on important issues. Social media platforms are not designed for resource seeking and experience large volumes of messages, leading to requests not being fulfilled satisfactorily. Designing frameworks to facilitate efficient information seeking in social media will help users to obtain appropriate assistance for their needs and help platforms to increase user satisfaction.

Several challenges exist in the way of facilitating information seeking in social media. First, the characteristics affecting the user's response time for a question are not known, making it hard to identify prompt responders. Second, the social context in which the user has asked the question has to be determined to find personalized responders. Third, users employ rhetorical requests, which are statements having the syntax of questions, and systems assisting information seeking might be hindered from focusing on genuine questions. Fouth, social media advocates of political campaigns employ nuanced strategies to prevent users from obtaining balanced perspectives on issues of public importance.

Sociological and linguistic studies on user behavior while making or responding to information seeking requests provides concepts drawing from which we can address these challenges. We propose methods to estimate the response time of the user for a given question to identify prompt responders. We compute the question specific social context an asker shares with his social connections to identify personalized responders. We draw from theories of political mobilization to model the behaviors arising from

the strategies of people trying to skew perspectives. We identify rhetorical questions by modeling user motivations to post them.

DEDICATION

To my family and almighty.

ACKNOWLEDGMENTS

I would like to thank Dr. Huan Liu for his guidance and support during my research. He is a great mentor granting freedom in charting our course and willing to work with different ideas. He is also very regular and disciplined in his work, something which I tried to imbibe to the best of my ability. I would like to thank my dissertation committee members, my co-advisor Dr. Ying-Cheng Lai, Dr. Hanghang Tong, and Dr. Roman Vaculin for their valuable interactions and feedback. I would also like to thank Dr. H. Russel Bernard for discussions on sociological aspects of my research and providing insightful suggestions.

I would next like to thank the members of Data Mining, and Machine Learning Lab have inspired and supported me over the years. I would like to thank Xia Hu, Jiliang Tang, Pritam Gundecha, Suhang Wang, Fred Morstatter, Huiji Gao, Ashwin Rajadesingan and Ghazeleh Beigi for discussing ideas and working with me. I would also like to thank my lab members over the years. I would like to thank Ali Abbasi, Lu Cheng, Christophe Faucon, Ruocheng Guo, Isaac Jones, Shamanth Kumar, Jundong Li, Tahora Nazer, Justin Sampson, Kai Shu, Robert Trevino, Liang Wu, and Reza Zafarani.

I would also like to thank mentors, colleagues, and friends from various places who have provided valuable suggestions on my research. The initial idea for the dissertation came from a class project, and I would like to thank my partner Anindita Dey for the brainstorming sessions. I would like to thank my lab members from SenSIP: Jayaraman Thiagarajan, Kartikeyan Ramamurthy, Deepta Rajan, Prasanna Settigeri, Mahesh Banavar, and my advisor Andreas Spanias. I would also like to thank my other collaborators: Hari Sundaram, Ross Maciejewski, Matthew Riemer, Yu-Ru Lin, Vinod Gupta Tankala, and Ed Mancebo.

I would also like to thank all my friends at ASU and beyond who have been with me through these years. Special thanks to Karthik, Rushil, Kuldeep, Jyothi, Hyma, Shashank, Vaibhav, Prasad, Parag, Vinay, Anindita, Parminder, Archana, Vinitha, Megha, Niranjan, Nithin, Vinith, Ashwin, Malvika, Mouna, Nikhil, Rashmi, Sandeep, Shibani, Rohit, Sridhar, Anagha, Sharath, Ritesh and Soma.

This material is based upon work supported by, or in part by, the Office of Naval Research (ONR) under grant number N000141010091 and N00014-16-2257. My Ph.D. study was also sponsored, in part, by the Teaching Assistantship in the Electrical Engineering Department under Dr. Joseph Palais and Clayton Javurek.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INFORMATION SEEKING IN SOCIAL MEDIA

Information seeking is defined as "A conscious effort to acquire resources in response to a need or gap in knowledge" (Case, 2012). Social media makes it easier for users to reach out to a large number of people in real time, leading them to post to their online social network to seek information. Considerable interest on information seeking in social media is shown in recent literature (Morris *et al.*, 2010; Lampe *et al.*, 2014; Zhao and Mei, 2013).

Examples of information seeking in social media are illustrated in Fig 1. Fig 1 features three requests seeking information from the asker's social network. The first request in the image is for assistance in the user's mathematics homework and is looking for a person who is familiar with the asker's math abilities. The second request is for time-critical information regarding a possible tsunami in Bangladesh and a prompt response containing the relevant information is expected. The third request is looking for a broad range of perspectives about the recent 2016 election campaign in the United States.

Designing algorithmic frameworks to facilitate information seeking in social media has several practical applications. Identifying answerers to personal questions will help to bridge the unique information gap of users and increase user satisfaction for personal and time-critical needs. Learning concepts that can satisfy personal needs can also help enhance search in social media by making it personalized to the asker. Facilitating users to get a broad range of perspectives can ensure that they make an informed opinion on the important issues facing society.

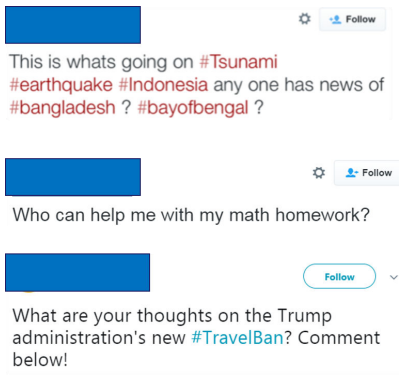


Figure 1: Examples of Requests in Social Media

Generic social media sites are not designed for information seeking (Paul *et al.*, 2011). Questions are buried among other content produced by the social connections of a potential answerer. Thus the responder might not see questions he is suitable to answer. Questions and statements are mixed with each other, and this leads to confusion for any possible system that aims to aid information seeking in social media. Users who seek perspectives are hindered by people who are trying to shape their opinion by pushing their agenda on their social media profiles. Information seeking can be facilitated by identifying responders and identifying posts and users hindering information seeking. I next present a few challenges associated with facilitating information seeking in social media.

First, the information need of social media users is subjective or personal in many cases, unlike traditional Q&A platforms like StackOverflow, and his social context is used to find appropriate people able to satisfy it (Hecht *et al.*, 2012). For example, to assist a person looking to get a new hairstyle, finding people from his social connections who share related context with him can be more useful to him than finding web pages related to hair salons. It is challenging to determine the social context of the asker

related to the question and appropriately utilized to connect the responders to the questions they are suitable to answer.

Second, the characteristics affecting the time taken to respond to a social media question are not precisely known. This makes it hard to estimate the time taken for a user to respond to a time-critical question posted in social media. Parallel research on information retrieval states that the characteristics of urgent queries are yet to be explored (Mishra *et al.*, 2014a), leading to people in need not getting prompt help. Millions of posts are published during emergencies by a large number of users, leading to a large candidate responder set. It is difficult to estimate the availability and inclination of the candidate responders to reply to a given question promptly. Moreover, timeliness is a distinct entity than relevance, and these entities have to be integrated to identify responders who can give timely and relevant responses.

Third, advocates in social media try and push their agenda on popular issues preventing users from getting a broad range of perspectives. They employ nuanced message construction and propagation strategies to shape user opinion and increase the spread of their messages, making it difficult to distinguish them from random users posting on issues related to the campaign. These strategies are very diverse, manifesting both in the activity patterns restricted to individual advocates like constructing persuasive messages, and multiple relational patterns like shared language and interactions, making it a challenge to study collectively in a unified model.

Fourth, rhetorical questions, an example of which is illustrated Fig 1 has the syntactic structure of a question and cannot be easily differentiated from information seeking questions. Previous research on question identification using syntax proposes that rhetorical questions are shown to be most prone to get misclassified (Li *et al.*, 2011). Rhetorical questions have the function of a statement, and determining its

function might lead us to distinguish them for other questions better. However, the purpose served by a standalone post of a social media user is not always apparent. In the question in Fig 1, the reason behind the user posting is not clear just by looking at the statement.

We draw from sociological and linguistic theories which provide concepts using which we can address these challenges. Social foci theory (Feld, 1981) postulates that interactions between people organized around relevant entities, known as foci. Inspired by the social foci theory, we propose that, people sharing question-specific social foci with the asker are suitable to answer personal, informational questions. The response time of a user for a time-critical question given in 1 can be determined by models of future availability(Morris *et al.*, 2010) and self-consistent (Korman, 1970) behavior of the candidate responders.

Sociological studies of political mobilization provide attributes of advocates using which I can build models of their characteristics. Behavioral theories record persuasive language and high degrees of emotion in the messages of advocates in their attempts to shape the opinion of people (McCarthy and Zald, 1977). Campaign communications are (Farrell and Webb, 2006) studies the widespread use of focused messaging and shared language patterns (Philipsen *et al.*, 1997). To increase the reach of messages during political campaigns, the utility of popular users for widespread propagation (González-Bailón *et al.*, 2013) and coordination through social connections have been studied (McCarthy and Zald, 1977).

Rhetorical questions can be filtered out by drawing concepts from linguistic literature to model user motivations to post them (Schmidt-Radefeldt, 1977). The motivation of posters of rhetorical questions, like subtly conveying a message (Schmidt-Radefeldt, 1977) and strengthen or mitigate a previous statement (Frank, 1990),

can be modeled and integrated with textual information of the question to help in differentiating between rhetorical and not rhetorical questions in social media.

Drawing concepts from these theories, we design algorithms to facilitate resource seeking in social media. We specifically answer the following questions.

- How to leverage future availability and self-consistency for identifying prompt and relevant responders to time-critical questions in social media?
- How to leverage question specific social foci to select responders who can satisfy personal requests?
- How to identify accounts of users who are trying to advocate their agenda on a given issue?
- How to model user motivations to distinguish between rhetorical and non-rhetorical questions posted in social media?

Chapter 2

RELATED WORK

While addressing the research questions on facilitating information seeking in social media, I drew concepts from a wide body of literature on social information seeking, information retrieval, linguistics, community Q&A platforms and political mobilization. I next enumerate the wide body of related literature and place my work in context.

2.1 Resource Seeking in Social Media

Resource seeking in social media which has received considerable attention in research communities (Ellison *et al.*, 2013; Lee *et al.*, 2012; Yang *et al.*, 2011). An analytical study of the primary motivations for information seeking and responding in Twitter is presented in (Morris *et al.*, 2010; Paul *et al.*, 2011). They indicated that subjective questions were the most prevalent, trust users have with their friends and the real-time information was the primary factor for asking questions A study of requests and responses received in Facebook (Gray *et al.*, 2013; Ellison *et al.*, 2013) presented information seeking as a tool for resource mobilization in social media. The factors affecting the quantity (Liu and Jansen, 2013) and speed of the responses (Teevan *et al.*, 2011) are studied, and these mainly correlate of question characteristics such as phrasing and posting time with the number and speed of responses. The prediction of response time for social media questions is studied in (Mahmud *et al.*, 2013) by modeling response time taken for the previous questions of the asker to estimate the

reply time. These papers give interesting insights to the question answering process in social media, but here we focus on identifying responders to these questions.

2.2 Responder Identification

Systems have been proposed to identify responders for social media questions to match question content with profile information (Hecht *et al.*, 2012) and use crowd-sourced technology (Jeong *et al.*, 2013). Search architectures with empirical models to route questions to responders using social information are discussed in (Horowitz and Kamvar, 2010; Nandi *et al.*, 2013). These papers are meant to demonstrate architectures of social search systems and hence do not contain any experimental evaluations. A method for recommending users who can answer questions in social media (Mahmud *et al.*, 2014) models temporal, behavioral and content related factors to identify suitable users. It identifies users capable of answering questions in general, and the users not optimized for a particular question.

Expertise finding methodologies in social media have received considerable attention in recent literature. Social expertise systems have been proposed for different social media platforms like Twitter (Weng *et al.*, 2010; Pal and Counts, 2011), enterprise social networks (Bozzon *et al.*, 2013) and image-based social networks like Instagram (Pal *et al.*, 2016) and identify subject matter experts in social media. These papers focus on finding experts for a given information need and do not consider possible response times when choosing an appropriate responder. Social media questions are subjective and personal might require answerers who share social context with the asker rather than subject matter experts.

The convergence of models for search and recommendation is another similar

line of research. A theoretical discussion on the advantages and opportunities of the fusion of search and recommendation algorithms is presented in (Garcia-Molina *et al.*, 2011). Models fusing search and recommendation to provide search results considering the interests of the searcher has been studied in (Weston *et al.*, 2012). The authors of (Mishra *et al.*, 2014b) build upon this to model the effects of the social network between searchers to rank the search results. Features determining the urgency of search queries, taking health-related searches as a case study have been proposed in (Hsiao *et al.*, 2014). These papers do not focus on social media questions and do not consider the timeliness of responses while ranking candidate responders.

2.3 Community Q&A Systems

A related line of research is timely information seeking in search and community Q&A systems like Yahoo! Answers (Adamic *et al.*, 2008) and Quora (Wang *et al.*, 2013a). Content from existing Q&A sessions is used to rank answerers by NLP techniques. (Jurczyk and Agichtein, 2007) uses link structure to find authoritative answerers for a question category. The authors in (Zhou *et al.*, 2012) and (Yang *et al.*, 2013b) combine network and content information to identify authoritative users as answerers. The environment for social media questions is different as the candidate answerers are themselves connected via social relations. Systems utilizing question categories (Zhu *et al.*, 2013) cannot be applied as they are not explicitly known in generic social media.

The temporal behavior of users of community Q&A platforms, such as factors affecting response time (Liu and Agichtein, 2011) and variation of response times for different categories of questions (Chua and Banerjee, 2013) have been recently

studied. The temporal dynamics between experts in a community Q&A platform have been studied in (Pal *et al.*, 2012), and it shows that modeling the evolution of experts improves the performance of expert identification community Q&A. However, these papers do not explicitly identify users who can provide timely and relevant responses to a given question.

A set of categories present in the questions along with the most popular question types has been studied in (Adamic *et al.*, 2008; Wang *et al.*, 2013a). These papers analyze categories already provided by the platform and do not address automatic question categorization. A taxonomy of question types in different Q&A platforms is manually constructed in (Harper *et al.*, 2010), and this gives insight into the variations of response quality and quantity across various categories. Identification of unresolved questions in community Q&A platforms have been addressed by exploiting conversation dynamics (Anderson *et al.*, 2012) and structure (Kim and Kang, 2014). Various methods of automatic question classification to facilitate easier search are evaluated in (Qu *et al.*, 2012; Chan *et al.*, 2013), but do not consider rhetorical questions as a relevant category. Users post in community Q&A platforms to get answers to their queries, and rhetorical questions are posted to make statements and not look for answers. Hence, rhetorical questions are not likely to be prevalent here, and the literature does not focus on them.

2.4 Question Identification

Automatic identification of questions posted in social media has been addressed in recent literature (Zhao and Mei, 2013; Wen and Lin, 2015; Hasanain *et al.*, 2014). The problem of identifying poorly phrased questions have been addressed in (Podgorny

et al., 2015) and the authors use grammatical structures of questions to identify them. Rhetorical questions are different from information seeking, or poorly phrased questions and characteristics unique to rhetorical questions have to be modeled to identify them. The authors in (Bhattasali *et al.*, 2015) address the problem of identifying rhetorical questions by directly combining contextual information. We model the motivations of the user to post rhetorical questions by utilizing particular relations between the question and its context.

2.5 Linguistic Theories of Rhetorical Questions

The characteristics of rhetorical questions have extensively been studied in linguistic literature (Gass and Seiter, 2015; Blankenship and Craig, 2006; Ilie, 1994). The use of rhetorical questions in public discourse as well as arguments between people has been studied in (Ilie, 1994). The authors survey different papers and study motivations of users to employ rhetorical questions in various conversations. The use of rhetorical questions to imply a message from its context instead of directly conveying it is studied in (Schmidt-Radefeldt, 1977). The utility of rhetorical questions in strengthening or mitigating the degree of the statement previously made in a conversation has been studied (Frank, 1990). We model the behaviors arising from these motivations to design a framework to identify rhetorical questions. The use of rhetorical questions for persuasion in social and political campaigns is also studied in linguistics (Gass and Seiter, 2015). The utility of rhetorical questions for persuasion (Petty *et al.*, 1981) as well as resistance to persuasive tactics (Blankenship and Craig, 2006) have been documented in the previous linguistic literature.

2.6 Social Foci Theory

Another related field to our work is the application of social foci theory in social media. Social foci theory has received attention in several domains such as relational learning (Tang and Liu, 2009) and structural hole theory (Burt, 2009). Recently, social foci theory has been used to derive community memberships using both node and edge attributes (Yang *et al.*, 2013a). To the best of our knowledge, this is the first work that has utilized concepts from social foci theory to identify answerers for social media questions.

2.7 Inferring Message Context

Considerable attention has been given to identifying characteristics of a user's status messages from his history (Yang *et al.*, 2012; Zangerle *et al.*, 2011; Liang *et al.*, 2012). The authors in (Godin *et al.*, 2013) predict the characteristics of the user's future messages by modeling the topics from his previous posts. The authors in (Kywe *et al.*, 2012) use collaborative filtering methods by incorporating the content of similar users and posts to compute characteristics of future messages. The authors in (Ma *et al.*, 2014) predict the characteristics of future messages by integrating the past content of the user, temporal information with the effect of interactions between candidate users. These works are recommended hashtags for a given post and are not specific to determining whether a question is rhetorical. Hence, the methodologies to model the user history is not explicitly designed for identifying rhetorical questions.

2.8 Social Media in Emergencies

The use of social media during emergencies has been extensively studied in literature (Gao *et al.*, 2011; Starbird and Palen, 2013, 2011). The utility of social media as a source of real-time information during emergencies has been evaluated in (Palen *et al.*, 2010) and it establishes a set of guidelines to assess information credibility and helpfulness. Identifying users present in the location of the disaster and comparing with those outside has been studied in (Kumar *et al.*, 2013b). These papers do not concentrate on information seeking during emergencies. The trends in information seeking patterns in Twitter is investigated in (Zhao and Mei, 2013) and it observes a high occurrence of bursts in questions during emergencies. A case study of microblogging behavior during the Yushu Earthquake has been conducted in (Qu *et al.*, 2011) and information seeking posts was found to be prevalent during emergencies. A system to match social media posts containing requests with individual posts containing offers during emergencies is proposed in (Purohit *et al.*, 2013). Our work differs in that we match questions with users who can provide timely and relevant responses to social media questions.

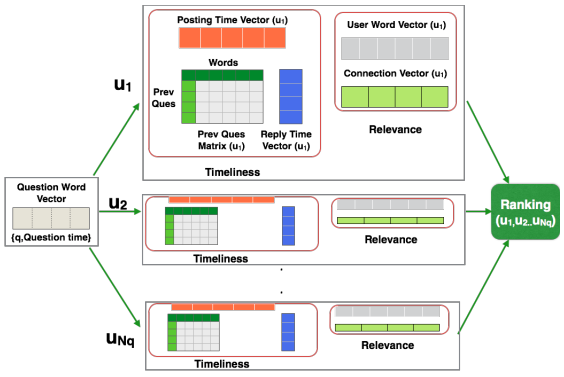
Chapter 3

FACILITATING TIME-CRITICAL INFORMATION SEEKING

Social media has emerged as a popular source of real-time information during natural disasters, social unrest, and political emergencies, where timeliness of information is a critical requirement. Considerable interest has been shown in recent literature (Purohit *et al.*, 2013; Qu *et al.*, 2011; Zhao and Mei, 2013) regarding the use of social media for information seeking and providing responses during natural calamities like Hurricane Sandy, Typhoon Haiyan, and the Haiti Hurricane. Fig 2a illustrates a few examples of requests for information, help and volunteering published emergency situations on the social media platform Twitter.

However, social media platforms are not equipped to facilitate timely information seeking, and this makes it difficult for users to obtain prompt responses (Paul *et al.*, 2011). The responders have to sift through many tweets, thus delaying the

(a) Examples questions during emergencies.



(b) The proposed framework

Figure 2: Identifying responders who provide timely and relevant responses to social media questions

response. Existing frameworks (Horowitz and Kamvar, 2010; Ranganath *et al.*, 2015a) only identify users who can provide relevant responses, not considering timeliness. Frameworks to identify responders who can provide timely and relevant answers faces many challenges.

I identify characteristics affecting the timeliness of responses to social media questions taking inspiration from sociological studies on information seeking and organizational behavior. Free time during the posting time of the question is an important motivation (Morris *et al.*, 2010) and the response time can be related to his future availability. The self-consistency theory (Korman, 1970) states that people perform tasks consistent with their previous instances of performing related tasks. Therefore, the response time can be related to his response times to similar questions. I propose a framework to identify automatically responders who can provide timely and relevant responses to social media questions. Specifically, I answer the following questions: How to model the temporal patterns of the candidates to rank them according to the timeliness of response? How to integrate temporal patterns with interests to identify users who can provide timely and relevant responses to a given question?

The major contributions made in the chapter are

- Formally defining the problem of identifying users who can provide timely and relevant responses to social media questions;
- Proposing an algorithmic framework to integrate timeliness and relevance for identifying responders to social media questions;
- Utilizing the framework to estimate the time taken for a question to obtain a response; and

- Presenting experimental evaluations on two real-world datasets of social media questions.

3.1 Problem Statement

In this section, we present notations used, describe a few relevant terms and formally present the problems we are addressing. Boldface uppercase letters (e.g. $\mathbf{X}$) denote matrices, boldface lowercase letters (e.g. $\mathbf{x}$) denote vectors, and calligraphic uppercase letters (e.g. $\mathcal{X}$) denote a set. The notation $\frac{1}{\mathbf{x}}$ indicates a vector whose elements are the reciprocal of each element of the vector $\mathbf{x}$. $\mathbf{X}_{\mathrm{ij}}$ signifies the element in the i^{th} row and j^{th} column of matrix $\mathbf{X}$. The i^{th} row of matrix $\mathbf{X}$ is denoted by $\mathbf{X}(\mathrm{i},:)$. We denote the Frebonius norm of a matrix $\mathbf{X}$ as $||\mathbf{X}||_{\mathrm{F}} = \sqrt{\sum_{\mathrm{i,j}} \mathbf{X}_{\mathrm{ij}}^2}$.

The terms related to the proposed framework, focusing on a single question, are illustrated in Fig 2b. Let the candidate question be denoted as q, and $\mathcal{U}$ be the set of candidate responders for the set of candidate questions $\mathcal{C}$. Let the set of previous questions answered by the users in $\mathcal{U}$ be denoted as $\mathcal{P}$. From Fig 2b, let $\mathbf{q} \in \mathbb{R}^{1 \times \mathrm{w}_{\mathcal{Q}}}$ denote the word frequency vector of question q, where $\mathrm{w}_{\mathcal{Q}}$ is the total number of words in the candidate question set and the set of previous questions.

We next define the terms related to the response timeliness. We denote t_q as the posting time of question q. For each user in the candidate set of question $\mathrm{u} \in \mathcal{U}_q$, we define the posting time vector $\mathbf{t}$. This vector contains the time in seconds of his previous postings with length equal to the total number of posts he made. The previous question matrix of the user u is represented as $\mathbf{P} \in \mathbb{R}^{o \times w_{\mathcal{Q}}}$, where o is the number of questions he answered previously. We denote the time taken in seconds by user u to reply to the previous questions by the reply time vector $\mathbf{rt}$ of length o.

We next define the terms related to the relevance of the candidate user u to the given question q. Let the user-word vector of the user u be denoted as $\mathbf{k} \in \mathbb{R}^{1 \times \mathrm{w}_{\mathcal{U}}}$, where $\mathrm{w}_{\mathcal{U}}$ is the total number of words used by users in $\mathcal{U}$. The connection vector of each user is obtained from the corresponding row of the network adjacency matrix $\mathbf{N}$. Finally, the relevance of the answered is denoted by a positive acknowledgment from the asker, like a "favorite" or a reply with "thanks".

Given these notations, we formally present two problems we address to facilitate time-critical information seeking in social media. The problem we address is to identify responders who can provide timely and relevant answers to a given question in social media. The problems are formally stated as follows: *"Given a question q, the question word vector $\mathbf{q}$, a set of candidate responders $\mathcal{U}_q$ along with the previous question matrix $\mathbf{P}$, the posting time vectors $\mathbf{t}$, the reply time vectors $\mathbf{rt}$, the user word vectors $\mathbf{k}$, and the social connection matrix $\mathbf{N}$ for all the users in $\mathcal{U}_q$, identify people in $\mathcal{U}_q$ who provide timely and relevant answers".*

3.2 Data Analysis

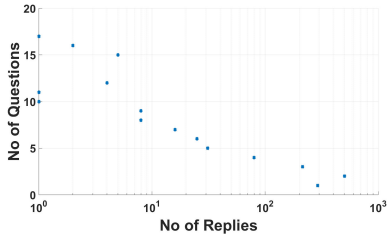
We are inspired by the sociological studies in information seeking (Morris *et al.*, 2010) and organizational behavior (Korman, 1970) to address the two problems of identifying suitable responders and estimating the response time for a given question. We are motivated by (Morris *et al.*, 2010) to postulate that the sooner a candidate responder is active on the platform after a question is posted, the faster can be his response to the question. We postulate from response consistency (Korman, 1970) that the response time of the user for the given question is proportional to his response time to questions related to it. We develop characteristics based on these theories

Parameter	Sandy	Rains
Candidate Questions	1,191	1,863
Askers	1,158	1,481
Positive Examples	2,877	3,905
Negative Examples	40,177	72,746
Candidate Respondents	43,064	75,921
Tweets by Candidate Respondents	26,911,778	24,301,721
Network Connections	812,817	914,044
Previous Questions Answered	572,202	914,488

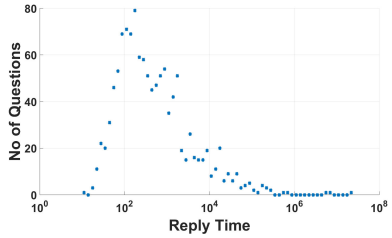
Table 1: Statistics of the two datasets.The first dataset consists of questions collected during Hurricane Sandy while the second dataset consists of questions collected during the recent Chennai Floods

that are capable of distinguishing between users who can provide timely responses and those who have not provided timely responses. To evaluate these characteristics, we collect questions posted on the social media platform Twitter to construct two datasets. We first describe the datasets and then use the datasets in evaluating the effectiveness of the characteristics to distinguish between users who can provide timely responses and those who have not provided timely responses.

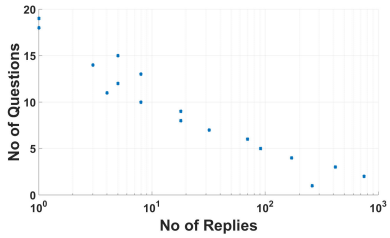
We have collected two datasets having the questions posted on the social media platform Twitter. The first dataset is collected during Hurricane Sandy using keywords and hashtags related to the events collected using (Kumar *et al.*, 2011). The earliest question in the dataset is posted on October 24th, 2012 and the latest question was posted on November 27th, 2012. The second dataset is collected during the Chennai Floods using similar techniques. The first question in the dataset is posted on Dec 1, 2015, and the last question was posted on December 15, 2015. For each question, we collected the text, user information and the timestamps of its replies and assigned the users who replied as the positive examples. We assigned the users who are posting on the same keywords and hashtags related to Hurricane Sandy within a day of the



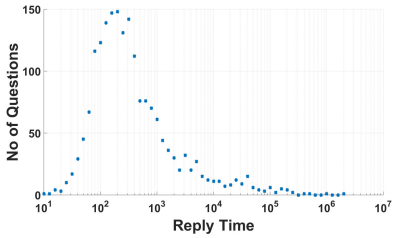
(a) Replies per question: Hurricane Sandy



(b) Reply time per question : Hurricane Sandy



(c) Replies per question : Chennai Rains



(d) Reply time per question : Chennai Rains

Figure 3: Dataset Statistics (a) No. of replies received (b) The reply time per question for Hurricane Sandy Dataset and (c) No. of replies received (d) The reply time per question for Chennai Rains dataset.The reply time per question follows a power law distribution in both the datasets, with a significant number of questions receiving less that 10 replies (Power law coefficient $\rho = -2.15$ for Hurricane Sandy, $\rho = -2.58$ for Chennai Rains). The reply time follows a bell shaped curve with the mode of the reply time is around 500 seconds in both the datasets.

question being posted but have not replied to the questions as negative examples. We use up to 1000 negative examples per question, and each negative example can be used for multiple questions. The positive and negative examples for each question q are jointly considered as the set of candidate responders $\mathcal{U}_{\mathrm{q}}$. We collected the tweets, times of tweets posted, network connections and previous questions answered by the candidate responders to construct $\mathbf{q}$, $\mathbf{k}$, $\mathbf{t}$, $\mathbf{P}$, $\mathbf{rt}$ and $\mathbf{N}$ as defined in Section 3.1.

The Hurricane Sandy dataset has 1,191 questions by 1,158 askers which are responded by 2,877 users. It has a pool of 40,177 respondents from which the negative

examples are drawn, up to 1000 for each question, making it a total of 43,064 candidate respondents. The candidate respondents have posted around 27 million tweets and answered 572,202 questions previously and have 812, 817 network connections. The Chennai Rains dataset has 1,1863 questions by 1,481 askers which are responded by 3,905 users. It has a pool of 72,746 respondents from which the negative examples are drawn, up to 1000 for each question, making it a total of 75,921 candidate respondents. The candidate respondents have posted around 25 million tweets and answered 914,488 questions previously and have 914,044 network connections. The statistics of the datasets are listed in Table 1.

Fig 3 illustrates some salient aspects of the dataset that are related to response quantity and times. Fig 3a and 3c illustrates the number of responses received per question for Hurricane Sandy and Chennai Rains datasets respectively. The number of replies is on the x-axis, and the number of questions having received the reply is on the y-axis. The distribution is power-law on both the datasets with the power law coefficient $\rho = -2.15$ for Hurricane Sandy dataset and $\rho = -2.58$ for the Chennai Rains dataset. This shows very few questions get a large number of replies in both the datasets. Fig 3b and 3d illustrates the characteristics of the response time for Hurricane Sandy and Chennai Rains datasets respectively. The number of replies is on the x-axis and the number of questions having with the response time is on the y-axis. The figures show a bell-shaped distribution for both the datasets with a majority of the questions receiving after 20 minutes. We first evaluate the effectiveness of the characteristics in distinguishing between users who can provide timely responses and those who have not provided timely responses with the Hurricane Sandy dataset. We use the characteristics to build the algorithm to identify timely and relevant responders and later evaluate the algorithm using both the datasets.

3.2.1 Evaluating Future Availability

The response time of a user to a question is dependent on the interval between the time the question is posted and the time he is next available on the platform (Morris *et al.*, 2010). To verify this, we propose the following postulate "The shorter the interval between the question time and the time the user is available on the platform, the faster he is likely to respond to the candidate question.". Let us consider a question $q \in \mathcal{Q}$ with a candidate set U_q. We take the posting time of a user as an indicator of his availability on the platform. We take the past posting times of a user $u \in \mathcal{U}_q$ who has responded to the question. To obtain the time he is available after the question is posted, we use time series forecasting methods proposed in (Zhang and Qi, 2005) to predict his next availability from his past posting times. We compute the reciprocal of the interval between the question time and the predicted future availability and assign it to an element of vector $\mathbf{a}$. We then obtain the reciprocal of his reply time and assign it to the vector $\mathbf{r}$. We then randomly pick a user in $\mathcal{U}_q$ who has not responded to the question and repeated the procedure. We consider the reciprocal of the reply time of the non-responders as 0. We repeat this procedure to all the responders in U_q of question $\mathbf{q}$ and then for the questions in the dataset. We postulate the null hypothesis $\mathcal{H}_0 : \mathbf{a} \nsim \mathbf{r}$ to show that reply time of the user is not correlated with the interval between the future availability and the question time and the alternate hypothesis $\mathcal{H}_1 : \mathbf{a} \sim \mathbf{r}$ to indicate that they are. Here $x \sim y$ denotes that x and y are correlated, and $x \nsim y$ indicates they are not. Computing the Pearson's correlation coefficient between the two vectors, with t-test to assess the significance, show that they are positively correlated with $p < 0.05$, thus verifying the postulate.

3.2.2 Evaluating Self Consistency

The response time of a user to similar questions in the past can also predict his response time to the given question (Korman, 1970). To verify this, we propose the following postulate "The shorter the response time of the user to similar questions in the past, the faster he is likely to respond to the candidate question". We first compute the topic distributions of the questions, and the previous questions responded by the users in their candidate sets using (Blei *et al.*, 2003). Let us consider a question $q \in \mathcal{Q}$ with a candidate set U_q. We first obtain the topic distributions of the past questions answered by user $u \in \mathcal{U}_q$ who has responded to the question q. We then compute the Euclidean similarities between the topic distributions of the candidate question and the set of past questions responded by the user and compute vector $\mathbf{e}$. Here, the Euclidean similarity between two vectors $\mathbf{x}$ and $\mathbf{y}$ as $\mathbf{e}(\mathbf{x}, \mathbf{y}) = \frac{1}{1+euc(\mathbf{x},\mathbf{y})}$, where $euc(\mathbf{x}, \mathbf{y})$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. We then compute the weighted average of the reciprocal of the response times of the user to the questions he answered in the past, using the similarity scores in vector $\mathbf{e}$ as weights. We assign the weighted average of the response times to vector $\mathbf{p}$. We then randomly pick a user in $\mathcal{U}_q$ who has not responded to the question and repeat the procedure appending the result to vector $\mathbf{f}$. We repeat this procedure to all the responders in U_q of question $\mathbf{q}$ and then for the questions in the dataset. We postulate the null hypothesis $\mathcal{H}_1 : \mathbf{p} \nsim \mathbf{r}$ to show that reply time of the user is not correlated with the response time to similar questions he has answered in the past and the $\mathcal{H}_0 : \mathbf{p} \sim \mathbf{r}$ alternate hypothesis to indicate that they are. Computing the Pearson's correlation coefficient between the two vectors, with t-test to assess the significance, show that they are positively correlated with $p < 0.0001$, thus verifying the postulate.

In this section, we evaluated the characteristics inspired by sociological characteristics in their ability to distinguish between users who can provide timely responses and those who have not provided timely responses. We next present a framework to model these characteristics and integrate them to solve the proposed problems aimed at facilitating time-critical information seeking in social media.

3.3 The Proposed Framework

In this section, we present a framework to facilitate time-critical information seeking in social media by addressing two problems: identifying responders who can provide timely and relevant responses to questions, and estimate the response time for a given question in social media. We first describe the ranking criterion and present quantitative models for identifying users who provide timely and relevant responders. A learning algorithm is then proposed to learn the parameters of the ranking criterion along with the time complexity analysis optimally. We then use the learned parameters to design a model to estimate the response time for a given question in social media.

3.3.1 Modeling Timeliness

We first present a model to rank the candidate responders according to their future availability. For each question q posted at time t_q, we take the posting time vector $\mathbf{t}$ of each candidate user up to t_q, taking posting as a measure of his activity on the platform. The rank of a user is inversely proportional to the estimated time after t_q

at which he is active on the platform. Therefore,

$$\mathrm{f_a}(q,u) = \frac{1}{|\mathrm{t}^{\mathrm{est}} - \mathrm{t_q}|}, \tag{3.1}$$

where $\mathrm{f_a}(q,u)$ is the ranking score of candidate user u for question q and $\mathrm{t}^{\mathrm{est}}$ is the time at which he posts in the platform after $\mathrm{t_q}$ as estimated from his posting time vector $\mathbf{t}$. We predict $\mathrm{t}^{\mathrm{est}}$ with a nonlinear autoregressive neural network with a single hidden layer (Zhang and Qi, 2005) on the posting time vector $\mathbf{t}$. The lower the estimated difference between estimated future availability and the question time, the higher the ranking score of candidate user u is.

We next rank the candidate responders according to their past response behavior to related questions. To represent the relationship between the given question and the previous questions answered by the user u, we transform the corresponding question word vectors into a common latent dimension space using $\mathbf{S} \in \mathbb{R}^{\mathrm{n} \times \mathrm{w_Q}}$. Here n is the number of dimensions of the space ($\mathrm{n} \ll \mathrm{w_Q}$). The representation of the given question q in the low dimensional space is then given by $\mathbf{qS}^{\mathrm{T}}$ and the representation of the previous questions answered by the user u is given $\mathbf{PS}^{\mathrm{T}}$. We represent its relationship to the previous question answered by the user incorporating domain correlation with $\mathbf{T} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$ as $\mathbf{qS}^{\mathrm{T}}\mathbf{TSP}^{\mathrm{T}}$. The ranking function can then be computed as

$$\mathrm{f_p}(q,u) = \mathbf{qS}^{\mathrm{T}}\mathbf{TSP}^{\mathrm{T}}\frac{1}{\mathbf{rt}}, \tag{3.2}$$

where $\mathbf{rt}$ is the time taken by user u to answer the previous questions. The ranking score $\mathrm{f_p}(q,u)$ is higher if the user u has promptly answered questions having a close relationship with question q in the past. The overall ranking criterion according to timeliness of reply is given by $\mathrm{f_t}(q,u) = \mathrm{f_a}(q,u) + \alpha \mathrm{f_p}(q,u)$, where α controls the amount of contribution from the past response behavior of the user to related questions.

3.3.2 Modeling Relevance

To model the relevance of a user to a given question, we compute the relationship between the content of the given question and the interests of the user. The nearer the question with the interest of the user, the greater can be his relevance to the question. We obtain the interests of the user from his user word vector $\mathbf{k}$ and represent it in a shared low dimensional space with the question word vector $\mathbf{q}$. Let $\mathbf{V} \in \mathbb{R}^{\mathrm{n} \times \mathrm{w}_{\mathcal{U}}}$ be the latent dimension representation of the user content, where n is the number of dimensions of the space and $\mathrm{w}_{\mathcal{U}}$ is the total number of words used by the set of candidate responders ($n \ll \mathrm{w}_{\mathcal{U}}$). The representation of the question q in the low dimensional space is then given by $\mathbf{qS}^{\mathrm{T}}$ and the representation of user u is given by $\mathbf{Vk}^{\mathrm{T}}$. We compute the relevance of the user u to the question q incorporating domain correlation with $\mathbf{T} \in \mathbb{R}^{\mathrm{n} \times \mathrm{n}}$ as

$$\mathrm{f_r}(q, u) = \mathbf{qS}^{\mathrm{T}}\mathbf{TVk}^{\mathrm{T}}. \tag{3.3}$$

The overall ranking criterion can be obtained by integrating the ranking scores $\mathrm{f_a}(q, u)$, $\mathrm{f_p}(q, u)$ and $\mathrm{f_r}(q, u)$. It is therefore computed as

$$\begin{aligned} \mathrm{f}(q, u) &= \mathrm{f_a}(q, u) + \alpha \mathrm{f_p}(q, u) + \beta \mathrm{f_r}(q, u) \\ &= \frac{1}{|\mathrm{t}^{\mathrm{est}} - \mathrm{t_q}|} + \alpha \mathbf{qS}^{\mathrm{T}}\mathbf{TSP}^{\mathrm{T}}\frac{1}{\mathbf{rt}} + \beta \mathbf{qS}^{\mathrm{T}}\mathbf{TVk}^{\mathrm{T}}, \end{aligned} \tag{3.4}$$

where α controls the amount of contribution from the past response behavior to related questions and β controls the amount of contribution from the relevance to the overall ranking criterion. The higher the candidate responders' estimated timeliness of response and relevance to the question, the higher is the score computed by $\mathrm{f}(q, u)$.

3.3.3 Learning Latent Parameters

We now present a learning algorithm to compute the latent matrices $\mathbf{S}$, $\mathbf{T}$ and $\mathbf{V}$ for optimal ordering of the candidate responders. Given a question q in the training set, we define the vector $\mathbf{f}$ containing the predicted scores for all the candidate responders for question q. The element of $\mathbf{f}$ related to the i^{th} candidate user is denoted by $\mathbf{f}_i$. In order to obtain an optimal ranking order for the candidate responders, we need to penalize the function when the users who have responded to the question are ranked low. The Weighted Approximate-Rank Pairwise (WARP) (Weston *et al.*, 2012) loss is defined as $\mathrm{err_{WARP}} = \sum_{\mathrm{i=1}}^{\mathrm{K}} \mathcal{L}(\mathrm{rank}(\mathbf{f}_i))$. Here, $\mathrm{rank}(\mathbf{f}_i)$ is a marginal ranking criterion which is computed as $\mathrm{rank}(\mathbf{f}_i) = \sum_{b \neq \mathrm{U_i}} \mathbb{I}[1 + \mathbf{f}_b \geq \mathbf{f}_i]$ where $\mathbb{I}(\mathrm{x})$ is the indicator function which is 1 if x is true or 0 if it is false, U_i is the i^{th} candidate responder and U_b is a member of the set of candidate responders of q who have not responded.

The pair $\{U_i, U_b\}$ is known as the violating pair if $1 + \mathbf{f}_b \geq \mathbf{f}_i$. The ranking function assigns to each pair a cost if the ranking score of U_b is larger or within a margin of 1 from the ranking score of U_i. The WARP loss function is therefore the penalty imposed when U_i is ranked within a certain margin or below a negative example U_b. $\mathcal{L}$ transforms the rank into a loss and is defined as $\mathcal{L}(\mathrm{k}) = \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{a_i}$. Here $\mathrm{a_1} \geq \mathrm{a_2} \geq \mathrm{a_3} \geq ...\mathrm{a_k} \geq 0$, with the values of $\mathrm{a_i}$ determining the additional penalty for each successive reduction in rank. A choice of $\mathrm{a_r} = 1/\mathrm{r}$, gives a larger penalty to the top position and provides a smooth weighting over positions (Usunier *et al.*, 2009).

We weigh the WARP penalty in proportion to the timeliness and relevance of the response given by U_i in the violating pair. We weigh the WARP loss as

$$\mathrm{err_{weighted}} = \sum_{\mathrm{i=1}}^{\mathrm{N}} (1 + \frac{1}{\mathbf{rt}_i})(\mathbf{rel}_i)\mathcal{L}(\mathrm{rank}(\mathbf{f}_i)), \tag{3.5}$$

where $\mathbf{rt}_i$ is the response time of U_i and $\mathbf{rel}_i$ is 1 if the response of U_i is accepted as relevant by the asker of q and 0 otherwise.

Calculating the exact rank is computationally expensive (Weston *et al.*, 2010) and we therefore approximate by sampling. We compute the stochastic gradient approach to minimize the error, choosing at each iteration a single training instance randomly from the training set $\mathcal{X}$. We compute the ranking score $\mathbf{f}_i$ of positive example U_i. We then randomly select users from the candidate set who have not replied to the question q and compute the ranking score for each of them until we find a violating pair i.e. $1 + \mathbf{f}_b \geq \mathbf{f}_i$. If L steps are required to find a pairwise violation, then the approximate value of the term $\text{rank}(\mathbf{f}_i)$ is given by

$$\text{rank}(\mathbf{f}_i) = \lfloor \frac{|\mathcal{U}_q| - 1}{\text{L}} \rfloor, \tag{3.6}$$

where $|\mathcal{U}_\text{q}|$ indicates the size of the candidate set and $\lfloor . \rfloor$ denotes the floor function. Following (Weston *et al.*, 2010), the single instance objective becomes

$$\text{f} = (1 + \frac{1}{\mathbf{rt}_i})(\mathbf{rel}_i)\mathcal{L}(\lfloor \frac{|\mathcal{U}_q| - 1}{\text{L}} \rfloor).|1 - \mathbf{f}_i + \mathbf{f}_b|. \tag{3.7}$$

Letting $C_\text{i} = (1 + \frac{1}{\mathbf{rt}_i})(\mathbf{rel}_i)\mathcal{L}(\lfloor \frac{|\mathcal{U}_q|-1}{\text{L}} \rfloor)$, we get

$$\begin{aligned} \text{f}_\text{r} =& C_\text{i}.(1 - (\frac{1}{\text{t}^\text{est} - \text{t}_\text{q}} + \alpha \mathbf{q}\mathbf{S}^\text{T}\mathbf{T}\mathbf{S}\mathbf{P}^\text{T}\frac{1}{\mathbf{rt}} + \beta \mathbf{q}\mathbf{S}^\text{T}\mathbf{T}\mathbf{V}\mathbf{k}^\text{T})_i + \\ & (\frac{1}{\text{t}^\text{est} - \text{t}_\text{q}} + \alpha \mathbf{q}\mathbf{S}^\text{T}\mathbf{T}\mathbf{S}\mathbf{P}^\text{T}\frac{1}{\mathbf{rt}} + \beta \mathbf{q}\mathbf{S}^\text{T}\mathbf{T}\mathbf{V}\mathbf{k}^\text{T})_b), \end{aligned} \tag{3.8}$$

We constrain the magnitude of the elements of matrices $\mathbf{S}$,$\mathbf{V}$,and $\mathbf{T}$ to reduce overfitting. The final objective function is then defined as follows

$$f = \text{f}_\text{r} + \gamma(||\mathbf{S}||^2_\text{F} + ||\mathbf{V}||^2_\text{F} + ||\mathbf{T}||^2_\text{F}). \tag{3.9}$$

We next optimize the objective function through gradient descent to obtain the updated values for the latent matrices $\mathbf{S}$, $\mathbf{V}$, and $\mathbf{T}$.

3.3.4 Computation of $\mathbf{S}$

Solving the optimization function on $\mathbf{S}$ is equivalent to minimizing the following objective function

$$\min_{\mathbf{S}} C_i.(1 - \mathbf{f}_i + \mathbf{f}_b) + \gamma||\mathbf{S}||_{\mathrm{F}}^2 \tag{3.10}$$

We solve this by gradient descent. The gradient of the function with respect to $\mathbf{S}$ is given by

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{S}} =& \mathrm{C_i}(\alpha((\mathbf{T}^{\mathrm{T}}\mathbf{S}\mathbf{K}^{\mathrm{T}} + \mathbf{T}\mathbf{S}\mathbf{K})_b - (\mathbf{T}^{\mathrm{T}}\mathbf{S}\mathbf{K}^{\mathrm{T}} + \mathbf{T}\mathbf{S}\mathbf{K})_i) \\ &+\beta((\mathbf{T}\mathbf{V}\mathbf{L})_b - (\mathbf{T}\mathbf{V}\mathbf{L})_i) + 2\gamma\mathbf{S},\end{aligned} \tag{3.11}$$

where $\mathbf{L} = \mathbf{k}^{\mathrm{T}}\mathbf{q}, \mathbf{K} = \mathbf{P}^{\mathrm{T}}(\frac{1}{\mathbf{rt}}\mathbf{q})$. At each iteration, the matrix $\mathbf{S}$ is updated by

$$\mathbf{S} \leftarrow \mathbf{S} - \eta\frac{\partial f}{\partial \mathbf{S}}, \tag{3.12}$$

where η is the weight given to the gradient.

3.3.5 Computation of $\mathbf{V}$

Solving the optimization function on $\mathbf{V}$ is equivalent to minimizing the following objective function

$$\min_{\mathbf{V}} C_i.(1 - \mathbf{f}_i + \mathbf{f}_b) + \gamma||\mathbf{V}||_{\mathrm{F}}^2 \tag{3.13}$$

The gradient of the function with respect to $\mathbf{V}$ is

$$\frac{\partial f}{\partial \mathbf{V}} = \mathrm{C_i}\beta((\mathbf{T}^{\mathrm{T}}\mathbf{S}\mathbf{L}^{\mathrm{T}})_b - (\mathbf{T}^{\mathrm{T}}\mathbf{S}\mathbf{L}^{\mathrm{T}})_i) + 2\gamma_2\mathbf{V} \tag{3.14}$$

where $\mathbf{L} = \mathbf{k}^{\mathrm{T}}\mathbf{q}, \mathbf{K} = \mathbf{P}^{\mathrm{T}}(\frac{1}{\mathbf{rt}}\mathbf{q}), \mathbf{M}_{\mathrm{ij}} = \mathbf{k}_i\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{k}_j^{\mathrm{T}}$. At each iteration, the matrix $\mathbf{V}$ is updated by

$$\mathbf{V} \leftarrow \mathbf{V} - \eta\frac{\partial f}{\partial \mathbf{V}}, \tag{3.15}$$

where η is the weight given to the gradient.

3.3.6 Computation of $\mathbf{T}$

Solving the optimization function with respect to the user specific matrix $\mathbf{T}$ is equivalent to minimizing the following objective function

$$\min_{\mathbf{T}} C_i.(1 - \mathbf{f}_i + \mathbf{f}_b) + \gamma||\mathbf{T}||_{\mathrm{F}}^2 \tag{3.16}$$

The gradient of the function with respect to $\mathbf{T}$ is given by

$$\frac{\partial f}{\partial \mathbf{T}} = \mathrm{C_i}(-(\alpha\mathbf{SKS}^{\mathrm{T}} + \beta\mathbf{SL}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}})_i + (\alpha\mathbf{SKS}^{\mathrm{T}} + \beta\mathbf{SL}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}})_b) + 2\gamma\mathbf{T} \tag{3.17}$$

where $\mathbf{L} = \mathbf{k}^{\mathrm{T}}\mathbf{q}, \mathbf{K} = \mathbf{P}^{\mathrm{T}}(\frac{1}{\mathbf{rt}}\mathbf{q})$. At each iteration, the matrix $\mathbf{V}$ is updated by

$$\mathbf{T} \leftarrow \mathbf{T} - \eta\frac{\partial f}{\partial \mathbf{T}}, \tag{3.18}$$

where η is the weight given to the gradient during the update. The update equations are summarized as below

$$\begin{aligned}
\mathbf{S} \leftarrow & \mathbf{S} - \eta(\mathrm{C_i}(\alpha((\mathbf{T}^{\mathrm{T}}\mathbf{SK}^{\mathrm{T}} + \mathbf{TSK})_b - (\mathbf{T}^{\mathrm{T}}\mathbf{SK}^{\mathrm{T}} \\
& + \mathbf{TSK})_i) + \beta((\mathbf{TVL})_b - (\mathbf{TVL})_i)) + 2\gamma\mathbf{S}) \\
\mathbf{V} \leftarrow & \mathbf{V} - \eta(\mathrm{C_i}\beta((\mathbf{T}^{\mathrm{T}}\mathbf{SL}^{\mathrm{T}})_b - (\mathbf{T}^{\mathrm{T}}\mathbf{SL}^{\mathrm{T}})_i) + 2\gamma\mathbf{V}) \\
\mathbf{T} \leftarrow & \mathbf{T} - \eta(\mathrm{C_i}(-(\alpha\mathbf{SKS}^{\mathrm{T}} + \beta\mathbf{SL}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}})_i + \\
& (\alpha\mathbf{SKS}^{\mathrm{T}} + \beta\mathbf{SL}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}})_b) + 2\gamma\mathbf{T})
\end{aligned} \tag{3.19}$$

where $\mathbf{L} = \mathbf{k}^{\mathrm{T}}\mathbf{q}, \mathbf{K} = \mathbf{P}^{\mathrm{T}}(\frac{1}{\mathbf{rt}}\mathbf{q})$ Here $\mathbf{L} \in \mathbb{R}^{\mathrm{w}_{\mathcal{U}} \times \mathrm{w}_{\mathcal{Q}}}$ and $\mathbf{K} \in \mathbb{R}^{\mathrm{w}_{\mathcal{Q}} \times \mathrm{w}_{\mathcal{Q}}}$ where $\mathrm{w}_{\mathcal{Q}}$ is the number of words in the question set Q, and $\mathrm{w}_{\mathcal{U}}$ is the number of words used by the set of candidate responders $\mathcal{U}$.

We repeat the procedure by randomly selecting a training instance until the error converges which we test using a validation set. We summarize the learning algorithm in **Algorithm 1**. We substitute the values of the latent matrices and compute the scores for the questions in the test set. For each question q, we order the set of candidate responders $\mathcal{U}_{\mathrm{q}}$ according to the scores and return the ranked list.

ALGORITHM 1: Finding Time Critical Responders in Social Media

Data: Training set with q, U_q , **P**,**t**, **rt**, **k** for each example
Result: Trained values of latent matrices **S**, **V** and **T**
Initialize $\mathbf{S}, \mathbf{V}, \mathbf{T}$ randomly;
do
  Pick a random labeled example i ;
  Compute $\mathbf{f}_i$;
  k=0 ;
  do
    Randomly pick a negative example $U_b \in \mathcal{U}_{\mathrm{q}}$
    Compute $f(q, b)$;
    k=k+1;
  while $1 + f(q, b) < f(q, u)$ *or* $k \leq size(\mathcal{U}_q) - 1$;
  Minimize f by updating $\mathbf{S}, \mathbf{V}, \mathbf{T}$ as in Eq 3.19;
  Substitute update matrices in f;
while $err_{weighted}$ *does not converge*;

To adapt to the real world settings where there are events with rapidly changing large-scale data, online updating methods to obtain the latent parameter values needs can be used. Online updating algorithms for learning to rank in (Schuth *et al.*, 2013) has been proposed and has been shown to work in real-world settings. In the future, we will plan to adopt the methods proposed to our algorithm and deploy it in real-world

settings. Our framework can then be deployed and evaluated using systems employing social media to assist first responders (Kumar *et al.*, 2014).

3.3.7 Estimating Response Time

There has been literature on estimating the response time for a given social media question (Mahmud *et al.*, 2013). Estimating response time can help the asker during emergencies and also guide him in deciding the right time to ask a given question. The response time for a given social media question is estimated from the response times of the previous questions of the asker in (Mahmud *et al.*, 2013). We examine if the response time can be better estimated by the properties of the candidate responders modeled by our framework rather than the properties of the asker. To do this, we first set $\beta = 0$ in Eq 3.4 to consider only the timeliness components and rank the candidate responders of each question accordingly. We then estimate the time taken by the top-ranked user to reply and return this as the estimated response time for the given question.

We estimate the response time taken by the top user by computing a weighted average of his previous reply times. In order to compute the weights, we transform the word vector of the given question $\mathbf{q}$ and word feature matrix of the previous questions he has answered $\mathbf{P}$ to a latent dimension and measure the similarity between the transformed vectors. The estimated time of the top user is computed as

$$\mathrm{t_{est}} = \sum_{\mathrm{i}=1}^{o} \frac{1}{1 + edist(\mathbf{q}\mathbf{S}^{\mathrm{T}}\mathbf{T}, \mathbf{S}\mathbf{P}^{\mathrm{T}}(\mathrm{i}, :))} \mathbf{rt}(\mathrm{i})$$

where $edist(x, y)$ is the Euclidean distance between x and y, $\mathrm{q}\mathbf{S}^{\mathrm{T}}\mathbf{T}$ is the latent dimension representation of the question transformed to the domain of user u, $\mathbf{SP}(\mathrm{i}, :)^{\mathrm{T}}$ is the latent dimension representation of the i^{th} question answered by the user u and

$\mathbf{rt}$(i) is the time taken to answer the i$^{\text{th}}$ question and o is the number of questions answered by him. The weight for a given question is higher if the distance between the latent dimension of the question transformed to the user domain and the latent dimension representation of the previous questions answered by the user is lesser. The time taken by the top user to respond is computed as a weighted average of his past response time to previous questions, with the similarity of the candidate question and the previous questions acting as the weight.

The estimated response time for the given question is computed as the time taken to respond to the top-ranked user. We repeat this procedure for all the candidate questions and compare the estimated response times with actual response times using different error measures. The method is summarized in **Algorithm 2**. We present the results of the experiments to evaluate the framework in next section.

ALGORITHM 2: Estimating Response Time for Questions

Data: Question set with q, U_q , $\mathbf{P}$,$\mathbf{t}$, $\mathbf{rt}$ for each example, and $\mathbf{S}$
Result: Estimated response times $\forall \mathrm{q_i} \in \mathcal{Q}$
Initialize vector $\mathbf{et} \in \{0\}^{1\times Q}$;
for *i=1:Q* **do**
  Compute latent representations of q_i $\mathbf{a} = \mathbf{qS}^{\mathrm{T}}\mathbf{T}$;
  for *k=1:length(U_q)* **do**
    Compute latent representations of questions k, $\mathbf{b} = \mathbf{SP}(\mathrm{k}, :)$;
    Compute similarity scores between $\mathbf{a}$ and $\mathbf{b}$ as $\frac{1}{1+euc(\mathbf{a},\mathbf{b})}$;
    Weigh the response time as $c = \frac{1}{1+euc(\mathbf{a},\mathbf{b})}\mathbf{rt_u}(\mathrm{k})$;
    $\mathbf{et}_{\mathrm{i}} = \mathbf{et}_{\mathrm{i}} + c$;
  end
end

3.3.8 Time Complexity

We now present the time-complexity to demonstrate the scalability of the framework for large datasets prevalent in social media. The majority of the time complexity comes from the update equations in Eq 3.19. The complexity of $\mathbf{T}^{\mathrm{T}}\mathbf{S}\mathbf{K}^{\mathrm{T}}$ and $\mathbf{T}\mathbf{S}\mathbf{K}$ is $O(\mathrm{w_Q n^2})$, due to the sparsity of $\mathbf{K}$. Similarly the complexity of $\mathbf{T}\mathbf{V}\mathbf{L}$, and $\mathbf{T}^{\mathrm{T}}\mathbf{S}\mathbf{L}^{\mathrm{T}}$ is $O(\mathrm{w_U n^2})$ as the matrix $\mathbf{L}$ is sparse.

The complexity of the terms $\mathbf{S}\mathbf{K}\mathbf{S}^{\mathrm{T}}$ is $O(\mathrm{w_Q n})$ due to sparsity of $\mathbf{K}$. Similarly the complexity of the terms $\mathbf{S}\mathbf{L}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}$ is $O((\mathrm{w}_{\mathcal{Q}} + w_{\mathcal{U}})\mathrm{n})$ when computed as shown in the brackets as $\mathbf{L}$ is sparse. Combining all the terms, the total complexity in each iteration is therefore $O((\mathrm{w}_{\mathcal{Q}} + \mathrm{w}_{\mathcal{U}})(\mathrm{n}^2 + \mathrm{n}))$. This is low owing to the low number of latent dimension n, indicating the scalability of the algorithm to a large dataset.

In this section, we propose a set of ranking criterion integrating information from future activity, past response behavior to related questions and user interests. We then present a learning algorithm to obtain an optimal ranking to identify responders who can provide timely and relevant answers. In the next section, we present the collected dataset and experiments designed to evaluate our framework.

Method	Hurricane Sandy			Chennai Rains		
	MRR	**MAP**	**NDCG**	**MRR**	**MAP**	**NDCG**
Random	0.14%	0.76%	0.30%	0.87%	0.11%	0.22%
Nandi et al	3.67%	2.26%	2.29%	1.80%	0.82%	1.35%
Mahmud et al	4.86%	7.32%	3.75%	3.98%	6.95%	4.89%
Topics	2.77%	5.12%	3.52%	7.39%	10.36%	8.82%
Future Availability	3.21%	8.46%	3.32%	10.13%	13.63%	12.49%
Past Response	7.19%	9.38%	8.49%	10.16%	13.77%	12.58%
Relevance	7.24%	9.56%	8.52%	11.70%	15.02%	14.45%
Our Model - **T**	9.47%	9.27%	11.73%	12.56%	16.01%	15.60%
Our Model	11.85%	10.84%	13.10%	13.36%	17.25%	16.49%

Table 2: Performance of the framework in ranking responders providing timely and relevant responses.

3.4 Experiments

In this section, we describe the experiments designed to evaluate our algorithm. We use the datasets to answer the following questions: How effective is the framework in integrating timeliness and relevance in identifying responders for social media questions? How does the framework perform for variations of parameter values and training data size? How does the framework perform in estimating the response time of the question posted in social media? We now use the datasets to proceed to answer these questions.

3.4.1 Experiment Settings

We evaluate the proposed framework and the baselines with Mean Reciprocal Rank (MRR), Mean Average of Precision (MAP), Non-Discounted Cumulative Gain (NDCG). We present some alternative baselines to compare our framework with related methods.

- **Random Selection**: We randomly order the candidate responders for each question and aggregate the rankings obtained by repeating over 100 iterations.
- **Future Availability**: This calculates ranking scores considering only the future availability of the responder ($\alpha = 0$ and $\beta = 0$).
- **Nandi et al. (Nandi *et al.*, 2013)**: The authors built a probabilistic model to combine temporal features and content metrics to rank candidate responders.
- **Mahmud et al. (Mahmud *et al.*, 2014)**: It proposes a supervised learning approach to learning features on the users' posting times and replying time to previous questions.

- **Topic Similarity**: This baseline substitutes latent parameters in the model with topic distributions of the questions, previous questions, and the status messages obtained from LDA (Blei *et al.*, 2003). This baseline is employed to demonstrate the utility of learning latent parameters specifically to rank timely and relevant respondents.
- **Past Response**: This baseline calculates ranking scores only considering information related only to the past timeliness to the previous questions.
- **Relevance**: This baseline calculates ranking scores from information related to the relevance of the user to the candidate questions.
- **Our Model - Dim Corr**: This baseline calculates ranking scores removing the dimension correlation matrix **T**.

We evaluate the proposed framework and the baselines with the following metrics

Mean Reciprocal Rank (MRR): The MRR of the mean of the reciprocal of the rank of the first positive example in the ranked list returned by the algorithm (Radev *et al.*, 2002). The MRR is calculated as

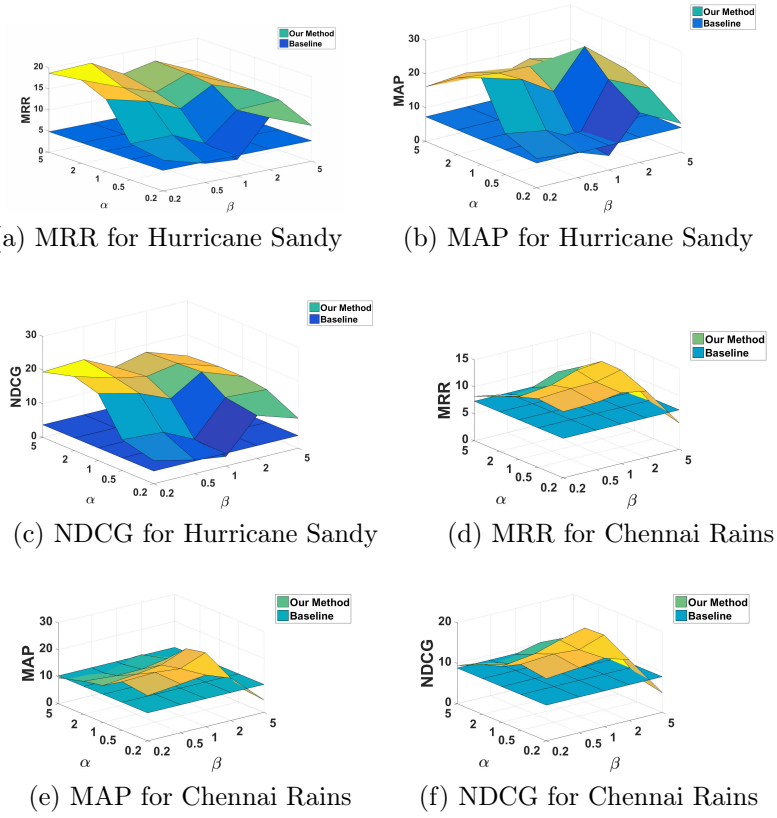
$$\text{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{\text{q}\epsilon\mathcal{Q}} \frac{1}{\text{rank}_\text{q}}$$

where $\mathcal{Q}$ is a set of questions, and rank_q is the rank of the first relevant responder for question q.

Mean Average of Precision (MAP): The average precision is the average of the Precision@K computed after each positive example appears in the ranked list (Bian *et al.*, 2008). MAP is computed as

$$\text{MAP} = \frac{1}{|\mathcal{Q}|} \sum_{\text{q}\epsilon\mathcal{Q}} \frac{\sum_{\text{n}=1}^{\text{N}_\text{q}} (\text{P(n)} \times \text{sui(n)})}{2 * |\text{R}_\text{q}|},$$

where N_q is the number of candidate responders for question q, $|\text{R}_\text{q}|$ is the number of responders for question q and P(n) is the Precision@K value computed when the n^{th}



(a) MRR for Hurricane Sandy (b) MAP for Hurricane Sandy

(c) NDCG for Hurricane Sandy (d) MRR for Chennai Rains

(e) MAP for Chennai Rains (f) NDCG for Chennai Rains

Figure 4: Performance of the framework with varying parameters for α and β as shown by (a) MRR (b) MAP (c) NDCG for Hurricane Sandy dataset, (d) MRR (e) MAP (f) NDCG for Chennai Rains dataset. The baseline is **Topics**, which is the closest in performance to our model among the baselines taken from prior work.

responder appears in the ranked list. The term sui(n) denotes the suitability of the r^{th} relevant responder and we compute this by

$$\text{sui(n)} = 2 - \frac{r_n}{\max(r)},$$

if n has replied to the question and 0 otherwise, where r_n is the time taken by user n to reply to the question. The suitability of the responder increases with the timeliness of his response.

Non-Discounted Cumulative Gain (NDCG): $NDCG_k$ takes the order of the

positive examples into consideration within the top K ranks (Wang *et al.*, 2013b). This measure is computed as

$$\begin{aligned} \mathrm{DCG_k} &= \sum_{\mathrm{r=1}}^{\mathrm{k}} \frac{2^{\mathrm{sui(i)}-1}}{\log_2(\mathrm{i}+1)} \\ \mathrm{nDCG_k} &= \frac{\mathrm{DCG_k}}{\mathrm{IDCG_k}} \end{aligned} \tag{3.20}$$

where $\mathrm{IDCG_k}$ computes the $\mathrm{DCG_k}$ for the optimal ordering of candidate responders.

In the rest of the section, we describe experiments designed to evaluate the framework using the evaluation metrics using the two datasets. In Section 3.4.2, we evaluate the performance of the framework in identifying users providing both timely and relevant responses. We keep the value of the parameters as $\alpha = 1$ and $\beta = 1$ in this section. Next, we evaluate the robustness of the framework due to parameter variation in Section 3.4.3 and change in training data size in Section 3.4.4. We conclude the experiments by examining if the framework can be used to estimate the response time for a given question in Section 3.4.5.

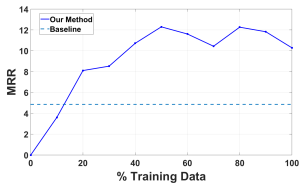
3.4.2 Timely and Relevant Responders

We now evaluate the performance of the framework in identifying responders who can provide both timely and relevant answers to questions in social media in both the datasets. We employ the procedure described in **Algorithm 1** for training and substitute the obtained latent matrices for scoring the questions and the candidate responders in the test set. For each question, we rank the candidate responders and evaluate the position of the relevant responders in the rank list. The results of the experiment are presented in Table 2, and we make the following observations.

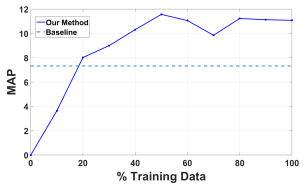
From the table, we can see that the performance of random ordering is low demonstrating the difficulty of the problem. The performance of (Nandi *et al.*, 2013) improves upon the random performance for both the datasets show the utility of modeling relevance for identifying responders for our task.The improved performance of [17] indicates the utility of response times from previous questions and supervised models for our task. The "Topics" baseline demonstrates an improved performance over (Mahmud *et al.*, 2014) and (Nandi *et al.*, 2013) in both the datasets, showing the effectiveness of the criterion in identifying timely and relevant responders.

The performance of "Future Availability" in both the datasets demonstrates the utility of estimating future user behavior in identifying timely responders to questions in social media. "Past Response," which models the relationship between the given question and the previous questions answered by the candidate, and "Relevance" which considers the relevance terms improves upon the baselines. This demonstrates the utility of the learning algorithm in determining suitable parameter values. Our combined framework considerably outperforms existing baselines by a significant margin in both the datasets, thus demonstrating its effectiveness in integrating information related to future availability, previous response patterns, and relevance in identifying responders who provide timely and relevant answers. The inclusion of dimension correlation matrix improves the performance of the algorithm, showcasing its utility. We performed a paired t-test to compare the results with the baselines that showed the improvement is significant with $p < 0.001$.

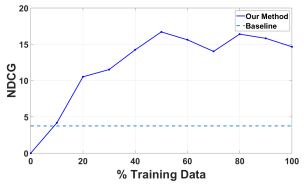
In summary, we can say from Table 2 that our framework is effective in identifying responders who provide timely and relevant answers in both the datasets. The results also demonstrate the ability of the framework to integrate effectively information crucial for identifying responders providing timely and relevant answers to a given



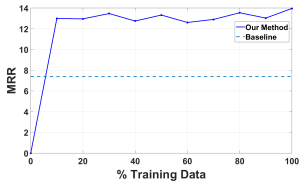
(a) MRR for Hurricane Sandy



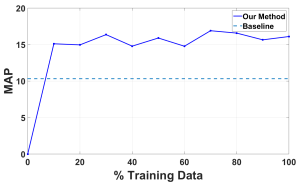
(b) MAP for Hurricane Sandy



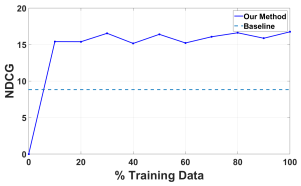
(c) NDCG for Hurricane Sandy



(d) MRR for Chennai Rains



(e) MAP for Chennai Rains



(f) NDCG for Chennai Rains

Figure 5: Performance of the framework with varying training data size as shown by (a) MRR (b) MAP (c) NDCG for Hurricane Sandy Dataset and (d) MRR (e) MAP (c) NDCG for Chennai Rains dataset. he baseline is **Topics**, which is the closest in performance to our model among the baselines taken from prior work.

social media question. Next, we will examine the effect on the performance of the framework due to variation in training data size and parameter values.

3.4.3 Effect of Parameter Variation

In the ranking criterion presented in Eq 3.4, α and β control the proportion of temporal patterns in the answering behavior and his relevance to the question respectively. In order to evaluate the performance of the framework to different proportions of temporal and relevance information, we vary the values of $\alpha = [0.2, 0.5, 1.0, 2.0, 5.0]$

and $\beta = [0.2, 0.5, 1.0, 2.0, 5.0]$ and plot the performance of the framework for these values in Fig 4 for both the datasets using the MRR, MAP and NDCG metrics. We make the following observations from the figure.

From Fig 4, we notice that the framework performs the best when there is equal proportions information related to timeliness and relevance (α=β). This shows the equal importance of information related to timeliness and relevance for this task and the effectiveness of the proposed framework for integrating them. The performance takes a dip in performance where information related to past response is relatively low, showing its importance in identifying prompt responders. The framework shows a good performance outperforms the nearest baselines **Topics** in both the datasets for a large range of parameter values, demonstrating its robustness to the variation of parameters.

In summary, the framework performs well over different proportions of information from temporal and relevance patterns and is robust to their variation in both the datasets. An appropriate combination of these kinds of information can optimize the effectiveness of the framework for identifying responders who can provide timely and relevant answers to social media questions.

3.4.4 Effect of Variation in Training Data Size

We now examine the relation between the performance of the framework with varying proportions of training data. This enables us to examine the performance of the framework when less amount of training information is available and also assess the robustness of the framework with varying training data size. We keep 50% of the data for training and the rest 50% for testing. We further divide the training data

into ten equal parts and vary the proportion from 10% to 90% in steps of 10% and measure the performance of the framework using MRR, MAP and NDCG metrics using the test dataset. We plot the results of the experiment in Fig 5 and make the following observations.

From the figure, we can say that the performance of the framework increases with increasing proportion of training data in both the datasets. The framework shows a good performance when 30% of the data is used for training, outperforming the nearest supervised baseline (Mahmud *et al.*, 2014), demonstrating that it performs well for fairly low training data sizes. We observe a small dip in performance for higher proportions of training data, and this may be due to insufficient testing data. The performance increases with increasing training data in the all three metrics in both the datasets, showing the ability of the framework to utilize the training data points effectively to identify timely and relevant responders to social media questions.

In summary, the figure demonstrates that the framework is effective in learning the latent parameters from a small amount of training data and that the framework is effective in learning when more training data is available. This also demonstrates the effectiveness of the proposed ranking criterion and the learning framework in exploiting information crucial for identifying such responders. Next, we investigate the effectiveness of the framework in estimating the response time for a social media question.

3.4.5 Estimating Response Time

We now describe the baselines and the metrics used to evaluate the method described in **Algorithm 2**. We compare the method with the following baselines

Method	Hurricane Sandy MAE	Hurricane Sandy RMSE	Chennai Rains MAE	Chennai Rains RMSE
Prev Asker	$1.01 * 10^6$	$3.09 * 10^6$	$4.66 * 10^4$	$1.69 * 10^4$
Past Replies	$3.69 * 10^4$	$4.50 * 10^4$	$2.54 * 10^3$	$1.12 * 10^3$
Our Model	$2.60 * 10^4$	$1.07 * 10^4$	$2.50 * 10^3$	$1.10 * 10^3$

Table 3: Estimating response time : Comparison of the framework with baselines

- **Prev Asker (Mahmud *et al.*, 2013)**: This work estimates the time taken to a given question to get a reply from the reply times to the asker's previous questions.
- **Past Response**: For each question, we compute the mean response time for previous questions answered by the candidate responders. The response time for the given question is estimated as the minimum of the mean response times of the candidate responders.

The response time estimated by the proposed method and the baseline is compared to the true response times using MAE (Mean Absolute Error) and RMSE (Root Mean Square Error), and the results are illustrated in Table 3.

From Table 3, we can see that modeling response time for the previous questions of the asker (Mahmud *et al.*, 2013) gives a poor RMS and MAE values. This performance is improved upon by "Past Responses" baselines showing that the response time for a question has a higher correlation with the temporal behavioral patterns of the candidate responders in both the datasets. The proposed method outperforms the baselines significantly, demonstrating that modeling the similarity of the given question and previous questions answered by the user will better estimate the response time for the asker. We performed a t-test between the results provided by the proposed method and the baselines that showed that the improvement is significant.

In summary, we show that the proposed framework can be applied to estimate the response time for a given question in both the datasets. The results showed that the proposed behavioral patterns of the candidate responders are effective in estimating the response time for a social media question.

3.5 Summary

In this chapter, we propose a novel framework to identify responders who can provide timely and relevant answers to questions in social media by integrating information related to their future availability, past response behavior, and interests. We evaluate the framework on using two datasets of questions posted in Twitter and demonstrate its effectiveness in identifying users to satisfy time-critical information needs.

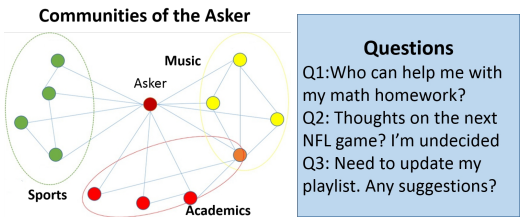
Chapter 4

FAST IDENTIFICATION OF PERSONAL RESPONDERS

When the information need of the user is subjective or personal, his social context might be useful to find appropriate people able to satisfy it (Hecht *et al.*, 2012). Users with higher tie strength with the asker were shown to better satisfy information needs in social media (Panovich *et al.*, 2012). For example, to assist a person looking to get a new hairstyle, finding people from his social connections who share related context with him can be more useful to him than finding web pages related to hair salons.

Inferring and utilizing the social context of the asker and his social connections in the question domain can be challenging. I make use of the social foci theory, which postulates that interactions between people are organized around relevant entities known as foci (Feld, 1981). A focus can be the activities, interests, and various affiliations of a user. Different groups of social connections of a user share different foci with him. For example, from Fig. 6a I see that the user shares an interest in sports with his connections in green, an interest in music with his connections in yellow and academic interests with his connections in red.

Inspired by the social foci theory I propose that, people in social media sharing social foci related to the question with the asker are suitable to answer them. Illustrative examples of questions are given in Fig. 6b. The asker of Q1 is seeking assistance in his math homework, and this might be best responded by users sharing academic foci with him. The answer for Q2 might be best provided by his connections sharing foci related to sports with the asker. Similarly, Q3 might be best answered by connections sharing music related foci.



(a) (b)

Figure 6: (a) Different foci a user shares with his social connections. (b) Questions of users. Users sharing different foci with the asker are more likely to answer related questions.

I propose a framework to investigate the utility of social context derived from network and content information in identifying answerers to social media questions. Specifically, I address the following questions: How to utilize the network and content information of the asker and his social connections to better identify answerers for social media questions? Are approaches based on the shared context in the question domain useful in identifying answerers to different kinds of social media questions? The main contributions of our work are as follows:

- Formally defining the problem of finding suitable users to answer questions in online social media platforms.
- Proposing a framework to exploit network and content information to identify answerers to social media questions, and
- Conducting experimental evaluations of the framework on a dataset of social media questions.

4.1 Problem Definition

We now define some terms related to the questions asked, the network and content of the asker and his social connections. We define attributes of a question q as the set of words used in the question i.e. $\mathbf{w}_{\mathrm{q}} = [\mathrm{w}_{\mathrm{q1}}, \mathrm{w}_{\mathrm{q2}}, ..., \mathrm{w}_{\mathrm{ql}}]$. Since we are dealing with subjective questions, the asker marking the answer to be useful or publicly acknowledging the answerer gives the evidence of its acceptance.

Let A denote the asker of the question q and $\mathbf{f}_{\mathrm{A}} = [\mathrm{f}_1, \mathrm{f}_2, ..., \mathrm{f}_{\mathrm{m}}]$ denote the social connections of A and m is the number of social connections of A. We define the egonetwork of each asker A as consisting of the asker, the social connections of the asker and the links among his social connections. The egonetwork of asker A, $\mathbf{N} \in \mathbb{R}^{(m+1)\times(m+1)}$ is given by

$$\mathbf{N}_{ij} = \begin{cases} 1 & \text{directed edge from } \mathrm{f_j} \text{ to } \mathrm{f_i}, \mathrm{i} \neq \mathrm{j},\ \mathrm{i,j} \in \{\mathrm{A}, \mathrm{f_A}\} \\ 0 & \text{otherwise} \end{cases}$$

We collect the status messages of the asker and his social connections. We apply basic preprocessing steps such as removal of stop words and stemming. We then define the user-word matrix $\mathbf{S} \in \mathbb{R}^{(m+1)\times w}$ of asker A as

$$\mathbf{S}_{ij} = \begin{cases} \text{num*tfidf}_{\mathrm{j}} & \text{if user } \mathrm{u_i} \text{ has used word } \mathrm{w_j} \text{ num times} \\ 0 & \text{if user } \mathrm{u_i} \text{ has not used the word } \mathrm{w_j}, \end{cases}$$

where num is the number of times the user $\mathrm{u_i}$ has used the word $\mathrm{w_j}$, w is the total number of words used by the asker and his social connections and $\mathrm{tfidf_j}$ is the tf-idf score of word $\mathrm{w_j}$. A single user will only use a small subset of the total number of words, resulting in $\mathbf{S}$ being sparse.

With the terminologies and the notations described above, we formally define the problem as follows *"Given a question q, an asker A, the network neighborhood of the*

asker $\mathbf{f}_A$*, find a suitable set of people among* $\mathbf{f}_A$ *whose responses for the question q that the asker accepts".*

4.2 Information Seeking via Social Foci

In this section, we describe our framework to identify answerers for social media questions in detail. First, we infer social foci memberships of the asker and his social connections from their network and content information. We then compute the overlap in foci memberships of the asker and his social connections in the question domain to identify answerers to these questions.

4.2.1 Modeling Content Information

We model the content information to infer major foci of the asker and his social connections. We draw from Non-negative Matrix Factorization (NMF) presented in (Seung and Lee, 2001) to infer foci from the user-word matrix $\mathbf{S} \in \mathbb{R}^{(m+1)\times w}$. We factorize the matrix $\mathbf{S}$ into two low dimensional sparse non-negative matrices, $\mathbf{U} \in \mathbb{R}^{(m+1)\times k}$ and $\mathbf{P} \in \mathbb{R}^{w\times k}$ such that k $\ll$ m by solving the following optimization problem.

$$\min_{\mathbf{U}\geq 0,\mathbf{P}\geq 0} ||\mathbf{S} - \mathbf{U}\mathbf{P}^{\mathrm{T}}||_F^2 \tag{4.1}$$

Here, k is the number of latent foci in the neighborhood of the asker and m is the number of his social connections. $\mathbf{U}$ denotes the latent foci membership of the asker and his social connections and $\mathbf{P}$ denotes the latent foci memberships of words. The correlation between foci memberships of the words can be obtained by the overlap in the corresponding rows of $\mathbf{P}$. The constraints $\mathbf{U} \geq 0$ and $\mathbf{P} \geq 0$ denote that the

matrices have all non-negative elements. The non-negativity ensures an intuitive decomposition of the matrix into its constituent parts.

4.2.2 Integrating Network Information

In a social setting, the interests or affiliations of a user are correlated with the interests of his social connections, thereby affecting his memberships to different foci (Feld, 1981). This notion is also supported by network homogeneity (Marsden, 1988), which says that people connected to each other display similar interests and affiliations. Therefore, it is essential to utilize network structure to determine foci memberships of the asker and his social connections.

To utilize the network structure, we first factorize the ego network of the asker $\mathbf{N}$ into two low rank non-negative matrices $\mathbf{U} \in \mathbb{R}^{(m+1)\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times k}$ s.t. k $\ll$ m by solving the following optimization problem.

$$\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0} ||\mathbf{N} - \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}||_F^2, \tag{4.2}$$

where $\mathbf{U}$ contains the membership of the asker and his social connections to different latent foci and $\mathbf{V}$ contains the correlations between the foci. The constraints $\mathbf{U} \geq 0$ and $\mathbf{V} \geq 0$ denote that the matrices have only non-negative elements.

We then integrate network and content information to infer the foci membership of the asker and his social connections by formulating the following optimization problem.

$$\begin{aligned}\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0,\mathbf{P}\geq 0} \quad & \alpha||\mathbf{S} - \mathbf{U}\mathbf{P}^{\mathrm{T}}||_F^2 + \beta||\mathbf{N} - \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}||_F^2 \\ & +\gamma(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2 + ||\mathbf{P}||_F^2)\end{aligned} \tag{4.3}$$

Here $\mathbf{U}$ contains the latent foci membership of the asker and his connections obtained by integrating network and content information, $\mathbf{P}$ shows the latent foci memberships

of the words and $\mathbf{V}$ represents the correlation between the latent foci. $||\mathbf{U}||_F^2$, $||\mathbf{V}||_F^2$, and $||\mathbf{P}||_F^2$ are regularization terms introduced to prevent overfitting and γ is the positive parameter for control the proportions of the regularization terms. The constraints $\mathbf{U} \geq 0$, $\mathbf{V} \geq 0$, and $\mathbf{P} \geq 0$ denote that the matrices do not contain negative elements. α and β are positive parameters to control the effects of content and network proportions respectively.

We draw from the concepts of the social foci theory illustrated in Fig. 6 to propose that users sharing a lot of foci memberships with the asker in the question domain can effectively answer social media questions. The shared foci memberships of the asker with his social connections are given by the overlap between their corresponding rows in $\mathbf{U}$. The question domain in the latent foci space is obtained by combining the rows of $\mathbf{P}$ corresponding to the words in the question. Before formalizing these notions, we optimally derive the latent matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{P}$ by solving Eq. (4.3).

4.2.3 Deriving the Optimal Latent Matrices

The problem presented in Eq. (4.3) belongs to a class of constrained convex minimization problems. Motivated by (Ding *et al.*, 2006), we describe an algorithm to find optimal solutions for $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{P}$. The key idea is to optimize the objective concerning one variable while fixing others. The three variables are iteratively updated until convergence.

From Eq.(4.3), we let

$$\mathcal{J} = \alpha||\mathbf{S} - \mathbf{U}\mathbf{P}^{\mathrm{T}}||_F^2 + \beta||\mathbf{N} - \mathbf{U}\mathbf{V}\mathbf{U}^{\mathrm{T}}||_F^2 + \gamma(||\mathbf{U}||_F^2 + ||\mathbf{V}||_F^2 + ||\mathbf{P}||_F^2) \tag{4.4}$$

We then take the Lagrangian of the objective function $\mathcal{J}$. Let the Lagrange multiplier

for the constraints $\mathbf{U} \geq 0$, $\mathbf{V} \geq 0$, and $\mathbf{P} \geq 0$ be Λ_{u}, Λ_{v}, and Λ_{p} respectively. Then

$$\mathcal{L} = \mathcal{J} + tr(\Lambda_{\mathrm{u}}\mathbf{U}^{\mathrm{T}}) + tr(\Lambda_{\mathrm{v}}\mathbf{V}^{\mathrm{T}}) + tr(\Lambda_{\mathrm{p}}\mathbf{P}^{\mathrm{T}}) \tag{4.5}$$

We compute the partial derivatives of the lagrangian $\mathcal{L}$ with respect to $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$ keeping the other variables fixed as shown below.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{U}} &= 2(\alpha(-\mathbf{SP} + \mathbf{UP}^{\mathrm{T}}\mathbf{P}) + \beta(-\mathbf{N}^{\mathrm{T}}\mathbf{UV} - \mathbf{NUV}^{\mathrm{T}} \\
&\qquad + \mathbf{UVU}^{\mathrm{T}}\mathbf{UV}^{\mathrm{T}} + \mathbf{UV}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UV}) + \gamma\mathbf{U}) + \Lambda_{\mathrm{u}} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{V}} &= 2(\beta(-\mathbf{U}^{\mathrm{T}}\mathbf{NU} + \mathbf{U}^{\mathrm{T}}\mathbf{UVU}^{\mathrm{T}}\mathbf{U}) + \gamma\mathbf{V}) + \Lambda_{\mathrm{v}} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{P}} &= 2(\alpha(-\mathbf{S}^{\mathrm{T}}\mathbf{U} + \mathbf{PU}^{\mathrm{T}}\mathbf{U}) + \gamma\mathbf{P}) + \Lambda_{\mathrm{p}}.
\end{aligned} \tag{4.6}$$

Substituting the KKT complementary conditions in Eq. (4.6) and rearranging we get the following update rules for latent matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{P}$.

$$\begin{aligned}
\mathbf{U}_{ij} &\leftarrow \mathbf{U}_{ij}\sqrt{\frac{\alpha\mathbf{SP} + \beta(\mathbf{N}^{\mathrm{T}}\mathbf{UV} + \mathbf{NUV}^{\mathrm{T}})}{\alpha\mathbf{UP}^{\mathrm{T}}\mathbf{P} + \beta(\mathbf{UVU}^{\mathrm{T}}\mathbf{UV}^{\mathrm{T}} + \mathbf{UV}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{UV}) + \gamma\mathbf{U}}} \\
\mathbf{V}_{ij} &\leftarrow \mathbf{V}_{ij}\sqrt{\frac{\beta\mathbf{U}^{\mathrm{T}}\mathbf{NU}}{\beta(\mathbf{U}^{\mathrm{T}}\mathbf{UVU}^{\mathrm{T}}\mathbf{U}) + \gamma\mathbf{V}}} \\
\mathbf{P}_{ij} &\leftarrow \mathbf{P}_{ij}\sqrt{\frac{\alpha\mathbf{S}^{\mathrm{T}}\mathbf{U}}{\alpha\mathbf{PU}^{\mathrm{T}}\mathbf{U} + \gamma\mathbf{P}.}}
\end{aligned} \tag{4.7}$$

The optimization algorithm is summarized in Steps 1-7 in **Algorithm 1**. The square root on the update rules is added to ensure convergence (Ding *et al.*, 2008). The correctness and convergence of the rules can be proved by the axillary function method (Lee and Seung, 2000).

4.2.4 Identifying Answerers from Foci Information

We now identify relevant answerers from the social connections of the asker using the latent matrices $\mathbf{U}$,$\mathbf{V}$ and $\mathbf{P}$. We first extract the words from the question attribute

vector $\mathbf{w}_{\mathrm{q}}$ and obtain the foci memberships of each word from the corresponding rows in matrix $\mathbf{P}$. We then compute the domain of the question in the latent foci space as a combination of individual word membership vectors as

$$\mathbf{d}_{\mathrm{q}} = \sum_{\mathrm{w_i} \epsilon \mathbf{w}_{\mathrm{q}}} \mathbf{P}_i, \tag{4.8}$$

where $\mathbf{d}_{\mathrm{q}}$ represents the domain of the question q in the latent foci space and $\mathrm{w_i}$ is the word corresponding to the i^{th} row of $\mathbf{P}$.

We next compute the foci memberships of the asker and his social connections in the question domain. The Hadamard product of two vectors is the pointwise product of their respective elements, and it exactly captures this notion. For each question, we compute the Hadamard product of the row of $\mathbf{U}$ corresponding to the asker, $\mathbf{U}_{\mathrm{A}}$ and the vector representing the question domain $\mathbf{d}_{\mathrm{q}}$.

$$\mathbf{g}_{\mathrm{A}} = \mathbf{U}_{\mathrm{A}} \circ \mathbf{d}_{\mathrm{q}}, \tag{4.9}$$

where $\mathbf{g}_{\mathrm{A}}$ contains the foci membership of the asker in the domain of the question. Similarly, we compute the foci memberships of each social connection of the asker in the domain of the question q by

$$\mathbf{g}_{\mathrm{f_m}} = \mathbf{U}_{\mathrm{f_m}} \circ \mathbf{d}_{\mathrm{q}}, \tag{4.10}$$

where $\mathrm{f_m}$ is the m^{th} social connection of the asker, $\mathbf{U}_{\mathrm{f_m}}$ is the row of matrix $\mathbf{U}$ corresponding to $\mathrm{f_m}$ and $\mathbf{g}_{\mathrm{f_m}}$ contains the foci membership of $\mathrm{f_m}$ w.r.t the domain of the question.

Finally, we find the overlap in foci memberships of the asker and his social connections in the question domain as

$$\mathbf{rs}(\mathrm{q}, \mathrm{A}, \mathrm{f_m}) = sim(\mathbf{g}_{\mathrm{A}}, \mathbf{g}_{\mathrm{f_m}}), \tag{4.11}$$

where $\mathbf{rs}(\mathrm{q}, \mathrm{A}, \mathrm{f_m})$ denotes the score of the answerer $\mathrm{f_m}$ to the question q by the asker A. We sort the answerers according to their score and return them to the asker as a ranked list, **ra**. Results with different similarity metrics is presented in Table 5. The method for identifying answerers from foci information is summarized in Steps 8-11 in **Algorithm 1**. The quantity $\mathbf{rs}(\mathrm{q}, \mathrm{A}, \mathrm{f_m})$ signifies the context in terms of network and content shared between asker A and his social connection $\mathrm{f_m}$ in the domain of question q.

ALGORITHM 3: Automatic Identification of Answerers

Data: Question q of asker (A), friends and followers of A, $\mathbf{f}_\mathrm{A} = [\mathrm{f}_1, \mathrm{f}_2, ..., \mathrm{f_m}]$, Egonetwork of the asker ($\mathbf{N}$), user-word matrix of the asker and his connections ($\mathbf{S}$), and $\{\alpha, \beta, \gamma, k\}$
Result: A ranked list of the potential answerers **ra**
Initialize $\mathbf{U}$, $\mathbf{V}$, $\mathbf{P}$ randomly;
for *i=1:Q* **do**
  Compute latent representations of q_i $\mathbf{a} = \mathbf{q}\mathbf{S}^\mathrm{T}\mathbf{T}$;
  do
    update $\mathbf{U}, \mathbf{V}, \mathbf{P}$ using Eqn 4.7;
  while *not convergent*;
  $\mathbf{et}_\mathrm{i} = \mathbf{et}_\mathrm{i} + c$;
  $\mathbf{w}_\mathrm{q} = [\mathrm{w_{q1}}, \mathrm{w_{q2}}, ..., \mathrm{w_{ql}}]$, $\mathbf{d}_\mathrm{q} = \sum_{\mathrm{w_i} \epsilon \mathbf{w}_\mathrm{q}} \mathbf{P}_i$;
  $\mathbf{g}_\mathrm{A} = \mathbf{U}_\mathrm{A} \circ \mathbf{d}_\mathrm{q}$, $\mathbf{g}_{\mathrm{f_m}} = \mathbf{U}_{\mathrm{f_m}} \circ \mathbf{d}_\mathrm{q}$;
  $\mathbf{rs}(\mathrm{q}, \mathrm{A}, \mathrm{f_m}) = sim(\mathbf{g}_\mathrm{A}, \mathbf{g}_{\mathrm{f_m}})$;
end

4.2.5 Time Complexity

The highest time cost results from updating the latent matrices in steps 4-6. In the updating terms, the complexity of the terms $\mathbf{SP}$ and $\mathbf{S}^\mathrm{T}\mathbf{U}$ is low due to sparsity of $\mathbf{S}$. The terms $\mathbf{N}^\mathrm{T}\mathbf{U}\mathbf{V}$, $\mathbf{N}\mathbf{U}\mathbf{V}^\mathrm{T}$ and $\mathbf{U}^\mathrm{T}\mathbf{N}\mathbf{U}$ have a complexity of $\mathrm{O}(\mathrm{mk}^2)$ where m is the number of friends and k is the number of latent dimensions due to the sparsity

of $\mathbf{N}$. The terms ($\mathbf{U}(\mathbf{V}(\mathbf{U}^{\mathrm{T}}\mathbf{U})\mathbf{V}^{\mathrm{T}})$, ($\mathbf{U}(\mathbf{V}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}}\mathbf{U})\mathbf{V})$ and ($(\mathbf{U}^{\mathrm{T}}\mathbf{U})\mathbf{V}(\mathbf{U}^{\mathrm{T}}\mathbf{U})$) has a complexity of $O(mk^2)$ when computed as shown in the brackets. The complexity of $\mathbf{P}\mathbf{U}^{\mathrm{T}}\mathbf{U}$ and $\mathbf{U}\mathbf{P}^{\mathrm{T}}\mathbf{P}$ is $O((w+m)k^2)$ where w is the number of words. Therefore, the overall complexity of a single iteration is $O((w+m)k^2)$, which is low owing to the few number of latent dimensions. In addition, notice that steps 1-7 can be computed offline and only steps 8-10 are computed when the question is asked, further reducing the time required to identify answerers for a given question.

4.3 Experiments

In this section, we first present a dataset of questions posted on Twitter and then conduct experiments to answer the following questions that help in understanding the framework better: How does the proposed framework perform in comparison to existing baselines? What is the effect of the amount of network and content information on the performance of the framework?

4.3.1 Dataset

The dataset consists of subjective questions from the social media platform Twitter. We follow the literature on questions in Twitter (Morris *et al.*, 2010) to construct a keyword set related to subjective questions. We append "?" to each keyword to collect questions from the Twitter Streaming API. Texts having "?" in online content are shown to be questioned with high precision (Cong *et al.*, 2008). We deem replies to have been accepted by the asker if he has marked it as "favorite" or acknowledged the answerer by using "thanks" or "thank you". We mark the users who provided these

Parameter	Statistics
# of Questions	1065
# of Askers	1026
# of Selected Answers	1450
# of Followers and Friends of the askers	966,117
Median # of Followers and Friends per asker	588
Median # Tweets per user	479

Table 4: Dataset containing questions posted in Twitter with statistics related to network and content information.

answers as the ground truth for each question following (Hecht *et al.*, 2012). Some important statistics of the dataset are given in Table 4. The first question was posted on Dec 27, 2013, and the last one on Jan 15, 2014. We use the methods in the public Twitter API to collect the friends, followers and public status messages of the asker to obtain the asker's social connections and their interests (Kumar *et al.*, 2013a). We use the data to construct the ego network $\mathbf{N}$ and user-word matrix $\mathbf{S}$ for each asker.

4.3.2 Experiment Settings

We introduce the following metrics to evaluate the performance of our framework: The Mean Reciprocal Rank (MRR) (Radev *et al.*, 2002) is a measure of the overall likelihood of the framework to identify an answerer for a question, the Mean Average of Precision (MAP) (Bian *et al.*, 2008) measures the potential satisfaction of the asker with the top K results and the Normalised Discounted Cumulative Gain (NDCG)@K considers the order within the top K rankings (Wang *et al.*, 2013b).We use the following baselines to evaluate the performance of our framework.

Random: We randomly order the friends and followers of the asker 100 times and return the mean ordering.

Aardvark (Horowitz and Kamvar, 2010): This paper describes a search engine which directed questions posted by the system to users with a formulation to compute affinity with the asker and interest in the question topics. It does not consider the network structure and also does not contain experimental evaluations of its formulation.

Content based Methods (Riahi *et al.*, 2012): The paper focuses on community Q&A like Yahoo! Answers and compares the similarity of the question topic with the interests of the answerers derived only from their content. Two topic models inferred the interests: LDA and the Segmented Topic Model (STM) (Du *et al.*, 2010).

Topic Sensitive Page Rank (Zhou *et al.*, 2012): This paper employs a PageRank based approach to find subject matter experts in the question topic by combining network and content information of the potential answerers. The paper identifies topical authorities not considering the shared context between the asker and the answerers.

Shared Foci: This baseline measures the effect of shared user context. It computes the shared foci memberships of the asker and his social connections derived from either network (α=0) or content (β=0) information. The question information is not taken into consideration. This also helps in evaluating methods using only network structure.

For initial experiments, we set the parameters in Eq. (4.3) as follows. The regularization parameter is set at γ=0.01. The number of topics in the baselines and the number of foci k is set as 50. For initial evaluation of the framework, we choose α=1 and β=1. The performance for different values of α and β will be presented in future subsections.

Method	MRR	MAP@5	NDCG@5
Random	1.20%	1.12%	0.25%
Content-LDA	1.56%	1.46%	0.30%
Content-STM	1.93%	2.27%	0.50%
TSPR	1.64%	1.63%	0.45%
Aardvark	2.11%	2.53%	0.50%
Shared Foci (Network)	3.43%	3.66%	0.97%
Shared Foci (Content)	3.60%	3.87%	1.17%
Our Model (Cosine)	3.91%	4.63%	1.25%
Our Model (PCC)	3.80%	4.73%	1.31%
Our Model (Euclidean)	4.36%	5.54%	1.41%

Table 5: Comparison of performance of the proposed framework with baselines.

4.3.3 Performance Evaluation

The results of the evaluations are presented in Table 5. From Table 5, we can see that the proposed framework has outperformed the baselines by a considerable margin. We conducted a paired t-test to compare the performance of our framework with that of the baselines, and the results indicated the difference between them is significant. We make the following observations from the table.

The proposed framework gives more than 300% improvement over random selection. We can see that simple formulation like the one in Aardvark that considers social network information performs on par with complex topical models using only content such as STM. The proposed framework also performs significantly better than methods identifying subject matter experts as answerers such as TSPR. This emphasizes the importance of social context to identify answerers to social media questions.

Considering shared foci between the asker and the answerer improves the performance over methods like Aardvark not utilizing community memberships. This shows the effectiveness of using social foci to exploit social context. Incorporating

question information to consider the overlap only in the foci related to the question gives further improvement in the performance.

In summary, by designing approaches based on shared social context and exploiting the structure of social ties, the proposed framework can effectively identify answerers for social media questions in the dataset. Next, we wish to understand the effect of content and network information on the performance of our framework.

4.3.4 Effect of Content and Network Information

In the model presented in Eq. (4.3), α and β control the proportion of the network and content information respectively. In order to evaluate the framework for different proportions of content and network, we set $\alpha = [0.1, 1, 10]$ and $\beta = [0, 0.1, 1, 10]$ and plot the values for MAP in Fig. 7 arbitrarily using cosine similarity as the similarity metric. We make the following observations from the figure.

A general trend in Fig. 7 has peaked at the main diagonal of the α and β axes and an off-diagonal dip. This shows that the framework works best for nearly equal proportions of network and content information. The MAP value is greater than 3% for all α and β except for low proportions of network information ($\alpha = 10$, $\beta = [0, 0.1]$). This emphasizes the importance of social connections of the asker for identifying answerers to social media questions. The lowest performance across all parameter values is more than twice than random ordering indicating the effectiveness of the framework for low relative proportions of content or network information. Overall, the MAP value is above 3% for different combinations of α and β indicating the effectiveness of the framework for a wide range of parameter values.

In summary, the framework performs well over different proportions of network

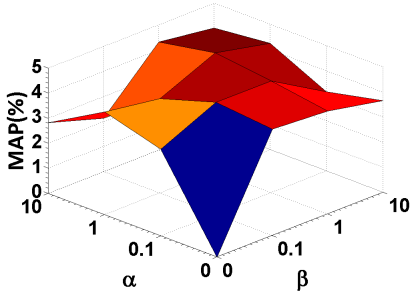


Figure 7: Effect of variation of content and network proportions on the framework performance for MAP.

and content and is robust to their variation. An appropriate combination of network and content information can optimize the effectiveness of the framework for identifying answerers to social media questions.

4.3.5 Performance across Question Categories

The literature on social media questions have identified kinds of questions people ask on Twitter. The recommendation, opinions, factual and rhetorical questions are popular questions asked on Twitter (Morris *et al.*, 2010; Paul *et al.*, 2011). We select four categories related to subjective questions, "Suggestions", "Opinion", "Favor", and "Rhetorical", and evaluate our framework in identifying answerers for different question categories.

We employed human labeling to assign category labels to questions. Three people independently labeled the questions, and the labels were assigned using majority selection. Employing this procedure, 93.5% of the questions were assigned to either of the four categories, and the framework was evaluated on them. The results of the evaluations are presented in Table 6. The distribution of different question categories

Categories	**Parts**	**MRR**	**MAP@5**
Suggestions	39.83%	4.27%(+2.23%)	4.68%(+1.78%)
Opinion	16.42%	2.67%(+1.43%)	2.38%(+1.61%)
Favor	30.51%	3.65%(+1.55%)	4.39%(+1.01%)
Rhetorical	6.74%	1.75%(+1.17%)	0%(+0%)

Table 6: Performance for different question categories.

is given in the first column. The performance for different categories is listed in the other columns. The improvement over (Horowitz and Kamvar, 2010), the nearest baseline not a part of our method, for different question categories are shown in the brackets.

From the table, we see that the framework gives considerable improvements over all the selected question categories. A paired t-test suggested that the improvements are significant, indicating that the framework is effective in finding answerers to a wide range of question categories in Twitter. The best performance can be seen in "Suggestions" and "Favor" categories and the performance in "Opinions" is relatively lower. These results suggest that identifying answerers for the "Opinion" category might depend on additional factors such as similarity of views in a given topic. The framework gives the lowest performance for questions in the "Rhetorical" category. Rhetorical questions are classified as conversational questions in the literature (Harper *et al.*, 2009). They might be used as an expression of opinion or to initiate a conversation and not to express an information need.

4.4 Summary

In this chapter, we draw from sociological theories to present a novel framework to identify possible answerers to personal questions. We evaluate the framework on questions on Twitter and demonstrate its effectiveness in identifying answerers. The framework is robust to a wide range of proportions of network and content information and categories of social media questions.

Chapter 5

UNDERSTANDING AND IDENTIFYING ADVOCACY

Social media is emerging to be a popular information channel for sociopolitical issues of broad importance e.g., elections and gun rights. It provides access to a wide range of perspectives on these issues, enabling users to form independent opinions. Owing to this, millions of people are using social media to seek information on these important issues. This has given rise to individuals who use it to try and push their agenda for political campaigns (Guo and Saxton, 2013). Media advocacy is defined in the literature as "the strategic use of mass media to advance a social or public initiative" (Jernigan and Wright, 1996). During the 2014 Indian elections, for example, a set of individuals formed an organization called NaMo Brigade with the motto of "Mission: Narendra Modi as PM" and used social media platforms to advocate for the election of Narendra Modi as Prime Minister (Lulla, 2014).

Although these campaigns have considerable social media presence (Today, 2014), it is difficult to identify individual accounts of advocates. Designing algorithms to identify accounts of individual advocates can better inform users as they navigate through social media spaces. People are accessing information about an issue, e.g., an election, through social media can be notified whether a given account is an advocate before reading their messages.

This task faces several challenges. First, advocates employ nuanced message construction and propagation strategies to shape user opinion and increase the spread of their messages, making it difficult to distinguish them from random users posting on issues related to the campaign. Second, these strategies are very diverse, manifesting

both in the activity patterns restricted to individual advocates like constructing persuasive messages, and multiple relational patterns like shared language and interactions, making it a challenge to study them collectively in a unified model.

Theoretical constructs of strategies for message construction, propagation, and community formation by advocates have been extensively studied in social sciences. Social movement theory records persuasive language and high degrees of emotion in the messages of advocates in their attempts to shape the opinion of people (McCarthy and Zald, 1977). The literature on campaign communications (Farrell and Webb, 2006) studies the widespread use of focused messaging for effective communication during political campaigns. Also, distinctive language patterns shared among people with similar affiliations foster easier communication of messages between them (Philipsen *et al.*, 1997). To increase the reach of messages during political campaigns, the utility of popular users for widespread propagation has been studied in (González-Bailón *et al.*, 2013). Formation of social and interaction networks between advocates for easier coordination and communication has been investigated in social movement literature (McCarthy and Zald, 1977).

In this chapter, we model the nuanced message strategies, propagation strategies, and community structure of advocates guided by sociological literature and integrate them to identify advocates for political campaigns on social media. We primarily focus on the following questions: How to model the nuanced strategies of advocates for political campaigns on social media? How to integrate them to design a unified framework for identifying advocates for political campaigns in social media?

The primary contributions of this work are:

- A definition of the problem of identifying advocates for political campaigns on social media,

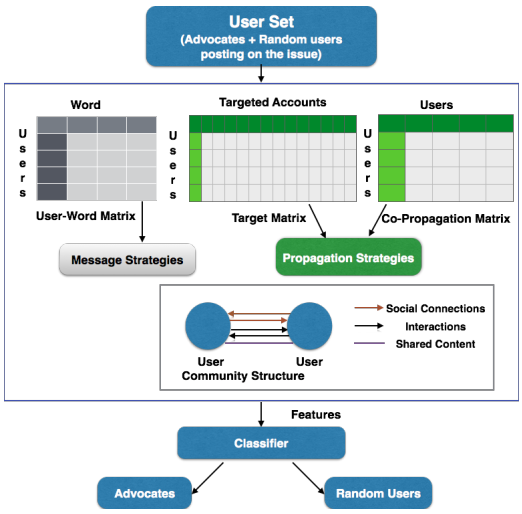


Figure 8: The proposed framework to identify advocates for political campaigns on social media

- A computational framework to gain insights into the strategies of the advocates and identify them by collectively modeling their strategies, and
- Evaluation of the framework in identifying advocates of political campaigns in the social media platform Twitter using two real-world datasets.

5.1 Problem Statement

In this section, we introduce notations and terms used and formally define the problem statement. We first define some notations. The n mode vector product of a tensor $\mathcal{M} \in \mathbb{R}^{D_1 \times D_2 \times .. \times D_N}$ with $\mathbf{x} \in \mathbb{R}^{D_n}$ is given by $\mathcal{M} \times_n \mathbf{x}$ and results in a tensor of size $D_1 \times D_2 \times .. \times D_{n-1} \times D_{n+1} \times ... \times D_N$ whose each element is given by $(\mathcal{M} \times_n \mathbf{x})_{d_1 d_2 .. d_{n-1} d_{n+1} ... d_N} = \sum_{d_n=1}^{D_n} \mathcal{M}_{d_1 d_2 .. d_N} \mathbf{x}_{d_n}$. The set of advocates is denoted as $\mathbf{a}$ and the set of random users using keywords related to the issue as $\mathbf{v}$. The set of users is denoted as $\mathbf{u} = [\mathbf{a}, \mathbf{v}]$ and the total number of users as N.

We now define terms related to different kinds of strategies of advocates that can be characterized. Message strategies deal with the construction of status messages of advocates with the aim of shaping the opinions of people. We develop characterizations on four types of message strategies possibly present in the status messages of advocates: persuasive language, a high degree of emotion, topical focus and shared language patterns. We construct the user word matrix $\mathbf{S} \in \mathbb{R}^{N \times l}$ from the status messages with tf-idf weighting, where l is the total number of words. Shared language patterns between two users in $\mathbf{u}$ are modeled taking hashtags as instances of language patterns. For each user $\mathbf{u}_i \in \mathbf{u}$, we construct a vector of hashtags $\mathbf{h}_{\mathbf{u}_i}$ from his status messages. We define matrices capturing shared hashtag information as $\mathcal{Z}^1$, where $\mathcal{Z}^1_{ij} = \text{jac\_sim}(\mathbf{h}_{\mathbf{u}_i}, \mathbf{h}_{\mathbf{u}_j})$. jac_sim($\mathbf{x}$,$\mathbf{y}$) indicates the Jaccard similarity between $\mathbf{x}$ and $\mathbf{y}$.

Propagation strategies comprise of the strategies employed by advocates to increase the spread of their messages. We model the propagation strategies of advocates from their targeting and co-propagation behavior. For each user in $\mathbf{u}_i \in \mathbf{u}$, let $\mathbf{ta}_u$ denote the set of people targeted by the user. We define the targeting matrix $\mathbf{T} \in \mathbb{R}^{N \times R}$ where $\mathbf{T}_{ij}$ is equal to the number of times $\mathbf{u}_i$ has targeted $\mathbf{r}_j$ where $\mathbf{r} = \bigcup_{i \in \mathbf{u}} \mathbf{ta}_i$ and R is the total number of users in $\mathbf{r}$. We next define the co-propagation network as $\mathbf{P} \in \mathbb{R}^{N \times N}$, where $\mathbf{P}_{ij}$ is the number of times $\mathbf{u}_j$ has propagated a message of $\mathbf{u}_i$.

Community structure deals with the patterns arising from the networks formed by advocates to facilitate easier communication and coordination. We model the community structures arising from social connections and interactions between advocates. The social connection matrix is defined as $\mathcal{Z}^2 \in \mathbb{R}^{N \times N}$. The value of $\mathcal{Z}^2_{ij}$ is 1 if $\mathbf{u}_i$ connects to $\mathbf{u}_j$ and 0 otherwise. We also capture the interactions between the users in $\mathcal{Z}^3 \in \mathbb{R}^{N \times N}$ where $\mathcal{Z}^3_{ij}$ is equal to the number of times $\mathbf{u}_i$ interacts with $\mathbf{u}_j$. In

addition, we define the tensor $\mathcal{Z}$ to hold the information contained in the relational matrices $\mathcal{Z}^1, \mathcal{Z}^2$ and $\mathcal{Z}^3$, where $\mathcal{Z}_{ijt} = \mathcal{Z}^t_{ij}$.

We model the message strategies, propagation strategies, and community structure and study them in a unified supervised learning framework to identify advocates for political campaigns on social media. The problem statement can then be stated as follows: *"Given a set of advocates for a given political campaign on social media* $\mathbf{a}$ *and a set of random users posting on the campaign* $\mathbf{v}$*, their status message matrix* $\boldsymbol{S}$*, the targeting matrix* $\boldsymbol{T}$*, the propagation network* $\boldsymbol{P}$*, and the relational tensor* $\mathcal{Z}$*, determine if a new user x is an advocate of the political campaign."*

5.2 Quantifying Strategies

In this section, we study the different nuanced strategies employed by advocates for political campaigns on social media drawing from theoretical constructs present in sociological literature and present ways to model them. We first study the employed strategies regarding their message strategies, propagation strategies, and community structure. We explore each of them in detail and then present ways to model them to derive characteristics possibly capable of distinguishing between advocates for political campaigns and random users posting on the campaign.

5.2.1 Quantifying Message Strategies

Message strategies deal with patterns from the construction of the status messages of advocates. Advocates can employ persuasive language in their status messages and attempt to sway opinions of other users. Parallels can be seen in the sociological liter-

ature which documents the high use of persuasive language during social movements (McCarthy and Zald, 1977). The status messages of advocates can also contain a high degree of emotions both positive; when they try to generate positive feelings about their campaign, or negative; when they try to generate feelings of anger, fear, and anxiety (McCarthy and Zald, 1977).

Message strategies of advocates can also manifest across a set of status messages. For instance, advocates can concentrate their status messages around a small number of topics in their attempts at effective communication, resulting in high topical focus. Parallels can be seen in studies of campaign communications (Farrell and Webb, 2006), which shows a widespread use of focused messaging. An advocate can also share distinctive language patterns with other advocates to support common issues and facilitate easier communication. Shared language patterns among people with similar affiliations have been shown in speech codes theory to foster easier communication (Philipsen *et al.*, 1997). Next, we present ways to model characteristics arising from persuasive language, emotions, focused messaging, and shared language patterns.

The use of persuasive language can be quantified by modeling theoretical principles of persuasion (Cialdini, 1993). We consider two principles of reason and affinity (Cialdini, 1993) and model their occurrence. People employ reason as a rational justification for their views while persuading others. The number of words related to reason (Gilbert and Henry, 2010) in the status messages of a user is used to quantify reason. Expressions of affinity can also be used as a tool for persuasion by using words conveying liking, compliments, and association. We use social words in the LIWC corpus (Pennebaker *et al.*, 2001) to model expressions of affinity and count their occurrence in the messages of each user. We postulate that advocates in **a** use

a higher number of words denoting persuasion than random users $\mathbf{v}$ posting on the campaign.

The emotional content in the posts of the users in $\mathbf{u}$ can be modeled using the positive and negative emotional words from the LIWC corpus (Pennebaker *et al.*, 2001). We postulate that advocates in social media use a higher number of emotional words, both positive and negative than random users posting on an issue. A higher use of emotion can be an indication that advocates have a strong belief in their cause, which separates them from paid workers posting promotional comments for a campaign, who display fewer emotions in their posts (Lee *et al.*, 2013).

To model the topical focus of a user in $\mathbf{u}$, we first compute the topic distribution using LDA (Blei *et al.*, 2003) on the user word matrix $\mathbf{S}$. The topic model results in the user topic matrix $\mathbf{DT} \in \mathbb{R}^{n \times t}$, where $\mathbf{DT}_{ij}$ is the number of times a word of user $\mathbf{u}_i$ has been assigned to topic $\mathbf{t}_j$ and t is the number of topics. The document-topic matrix can be normalized $\mathbf{DT}$ to obtain $\mathbf{DT'}$, where each row of $\mathbf{DT'}$ contains the probability distribution over topics of a user. The entropy of the topic distribution from the corresponding row of $\mathbf{DT}$ for each user i to construct a vector $\mathbf{e}_i$ where $\mathbf{e}_i = \sum_{j=1}^{j=t} -\mathbf{DT'}_{ij} \log(\mathbf{DT'}_{ij})$. It is evident that a lower value of entropy for a user implies greater concentration on fewer topics in his messages and a higher value implies distribution over a larger number of topics. Therefore, the higher the topical focus in the status messages of a user, the lower the value of his entropy. We postulate that advocates have a higher topical focus in their messages than random users posting on the campaign.

We next model the shared language patterns among users in $\mathbf{u}$ and use hashtags as instances of language patterns. For each advocate $a \in \mathbf{a}$, let $\mathcal{Z}^1_{ab}$ denote the amount of hashtags he shares with any other advocate $b \neq a \in \mathbf{a}$ measured by Jaccard similarity.

Similarly, let $\mathcal{Z}^1_{\text{av}}$, where v $\in$ **v** is a random user posting on issues related to the campaign. We postulate that an advocate for a political campaign shares a higher amount of hashtags with other advocates than with random users in **v** posting on the campaign. We evaluate these characteristics using the datasets in Section 5.3.3.

Until now, we characterized and modeled message strategies capable of distinguishing advocates for political campaigns from random users posting on the issue of persuasive language, emotion, topical focus and shared language patterns with other advocates. We next examine characteristics of propagation strategies employed by advocates to increase the reach of these messages.

5.2.2 Quantifying Propagation Strategies

We examine the propagation strategies of advocates for political campaigns on social media, focusing on their targeting and co-propagation behavior. Social media enables advocates to target specific users for propagating information. We propose that advocates target popular users more frequently than random users posting on the campaign as popular users help to get messages across to a wider audience (González-Bailón *et al.*, 2013). We then investigate how advocates assist each other in spreading their messages.

We first model the targeting behavior of the users in **u**. Let **r** is the set of people targeted by all users in **u** as defined in Section 5.1. Taking the number of users connecting to a user as a measure of his popularity, we construct the vector **c** where $\mathbf{c}_{\text{i}}$ is the number of people connecting to $\mathbf{r}_{\text{i}}$. We postulate that the attention of advocates is more skewed towards users with higher popularity than the attention of random users posting on the campaign. To model this postulate, we compute the vector

$\mathbf{sta} = \mathbf{Tc}$, where $\mathbf{sta}_k$ is the sum of number of times a user k targets an user $\mathbf{r}_i$ weighted by $\mathbf{c}_i$, the number of users connecting to $\mathbf{ta}_i$. The value of $\mathbf{sta}_k$ is higher if the user k targets popular users a higher number of times. Therefore, our postulate is satisfied when advocates for a political campaign have a higher value of **sta** than random users posting on the campaign.

We next model the co-propagation behavior of users in **u**. Advocates will be more interested in propagating messages of other advocates, and also, their messages will be more likely to be propagated by other advocates than random users posting on the campaign. Based on this, we characterize advocates by their hubs and authority scores (Kleinberg, 1999) in the information propagation network **P**. We compute hubs and authority scores of users in **u** using the information propagation network **P** and postulate that advocates have higher hub and authority scores than random users posting on the campaign. We evaluate these characteristics using the datasets in Section 5.3.4.

Until now, we proposed characteristics of advocates for political campaigns on social media from their message strategies and propagation strategies along with methods for modeling them. We next propose characteristics related to the community structure arising from their relationships with other advocates for the campaign.

5.2.3 Quantifying Community Structure

Community structure deals with the patterns arising from the networks formed by advocates to facilitate easier communication and coordination. Social media provides opportunities for advocates to connect to each other through many different types of relationships. Advocates can form social connections, interact with each other

for coordination, and carry out conversations. The formation of networks of social connections and interactions between advocates for communication and coordination have been studied in theoretical studies of social movements (McCarthy and Zald, 1977). Social connections and interactions between advocates can give rise to the similarity in community memberships. We now postulate a few underlying hypothesis to establish the similarity in community memberships between advocates in social media

We first present postulates are underlying community structure arising from social connections and interactions of advocates. For each advocate a $\in \mathbf{a}$, let $\mathcal{Z}^2_{\text{ab}}$ be 1 if a connects to b and 0 otherwise, where b is another advocate b $\neq$ a $\in \mathbf{a}$ and $\mathcal{Z}^2$ is defined in Section 5.1. Similarly, let $\mathcal{Z}^2_{\text{av}}$ be 1 if a connects to b and 0 otherwise, where v $\in \mathbf{v}$ is a random user posting on issues related to the campaign We then postulate that advocates are more likely to form social connections with other advocates than with random users posting on the campaign. We follow a similar procedure using $\mathcal{Z}^3$ to postulate that advocates are more likely to interact with each other than with random users posting on the campaign. These postulates underly that advocates have similar community memberships for different types of relationships. Are these community memberships of advocates are similar when measured across relationship types?

To model this, we first select the users with whom the advocates have at least one type of relationship with. For each advocate, we construct a vector $\mathbf{co}_{\text{ab}}$, b $\neq a \in \mathbf{a}$, where each element is the number of types a pair of advocates have relations in. Similarly, we construct the vector $\mathbf{co}_{\text{av}}$, v $\in \mathbf{v}$, where each element is the number of types of relations between a pair of advocate and a random user posting on the campaign. We postulate that given an advocate has established one type of relationship

with a user; he has a significantly higher propensity to form more types of relation if the user is another advocate than if he is a random user posting on the campaign. The postulate, if verified, underlies that community memberships of advocates are shared across different relationship types, and hence, they can be jointly inferred by efficiently combining different relationship types.

In this section, we draw from theoretical constructs in sociological literature to propose different characterizations of the nuanced message strategies, propagation strategies, and community structure of advocates for political campaigns on social media. Next, we are going to use two real-world datasets from Twitter to evaluate these characteristics in their ability to distinguish between advocates and random users posting on the issue.

5.3 Evaluating Strategies

In this section, we describe the datasets used to evaluate our characterizations of advocates of political campaigns in social media. We have two datasets from Twitter, each related to a political campaign carried out using Twitter. We then use these datasets to evaluate the ability of the proposed characterizations of strategies to distinguish between advocates for a given political campaign and random users posting on issues related to the campaign.

5.3.1 Datasets

We have two datasets related each related to a political campaign. The first dataset is focused on advocates for the Indian election campaign. The rise of around 200

Parameter	Elections	Gun Rights
Total # of Users	9390	7695
# of Tweets	20,362,442	19,275,481
# of Links	514,501	899,535
Users posting on the campaign	8500	7000
# of Advocates	890	695

Table 7: Statistics of the datasets of advocates.

million users of social media in India has made it an important platform for political discourse during elections (Khullar and Haridasan, 2014). Independent groups like NaMo Brigade (Lulla, 2014) were formed to advocate for the political campaign of Narendra Modi. The second dataset is related to the issue of gun rights in the United States. This campaign is focused on preserving gun rights, which is being questioned in the wake of increasing gun-related violence. Organizations advocating to preserve gun rights as the National Rifles Association (NRA) have considerable social media presence (Palmer, 2014).

Although these organizations have a considerable media presence, it is a challenge to obtain labels for individual users involved in advocacy. Previous literature proposes the use of publicly compiled lists as an effective alternative for inferring affiliations of social media users (Kim *et al.*, 2010). The authors in (Bhattacharya *et al.*, 2014) have utilized lists in Twitter to characterize topical-identity based groups. Informed by this literature, we identified two public Twitter lists, titled "NaMo Brigade" (Baruah, 2014), for advocates for the election campaign of Narendra Modi and "NRA" (Robinson, 2014), for the advocates for gun rights.

To validate the datasets, we apply the mark and recapture technique, drawing from population estimation methodologies (Brower *et al.*, 1998). For each of the two lists, we draw two random samples and estimate the total number of errors in the list as follows. Let the probability of finding errors in the random samples r_1 and r_2 be p_{r_1}

and p_{r_2}. The number of errors in both the samples will then be given by $e_{r_1} = p_{r_1} N_e$ and $e_{r_2} = p_{r_2} N_e$, where N_e is the total number of errors to be estimated. The number of errors in the intersection of the two samples is then $e_{r_1 r_2} = p_{r_1} p_{r_2} N_e$. The number of errors in the dataset N_e can then be estimated as $N_e = e_{r_1} e_{r_2} / e_{r_1 r_2}$ This is shown to overestimate the population, and hence we use a variation to estimate the number of errors as

$$N_e = \frac{(e_{r_1} + 1) \times (e_{r_2} + 1)}{e_{r_1 r_2} + 1} - 1 \tag{5.1}$$

A measure of uncertainty is given by the standard error, which estimates of the variability of N if the above experiment is conducted repeatedly. This is computed as follows

$$SE = \sqrt{\frac{(e_{r_1} + 1) \times (e_{r_2} + 1) \times (e_{r_2} - e_{r_1 r_2}) \times (e_{r_1} - e_{r_1 r_2})}{(e_{r_1 r_2} + 1)^2 \times (e_{r_1 r_2} + 2)}} \tag{5.2}$$

From the standard error, we then calculate the 95% confidence interval i.e. the range within which the number of errors lies with 95% certainty as $I_e = N_e \pm 1.96 \times SE$, where I_e is the 95% confidence interval of the error estimate.

We draw random samples of 10% of the size of the list and use external evaluators to verify them. The evaluators are asked to assess whether a user is an advocate of a given political campaign in social media. The definition of advocates is given to "individuals who use social media to advance their agenda for a given political campaign" strategically, according to the definition provided. The evaluators mark 1 if they think the member is an advocate and 0 otherwise. We then estimate the total number of errors from Eq 5.1 and the confidence interval. The percentage accuracy of the two lists 'NaMo Brigade' and 'NRA' are $92.10\% \pm 5.29\%$ and $90.07\% \pm 6.27\%$ respectively. Validating the accuracy of the lists, we use their members as ground truths for the set of advocates **a**.

Factors	Features	Elections	Gun Rights
Message Strategy			
Persuasion	Reason	3.52*	6.52**
	Liking	1.717***	1.04***
Focus	Entropy	-3.63***	-0.21**
Emotion	+ve	0.5731**	0.3669**
	-ve	0.7782***	0.2621**
Shared Language		3.8571***	10.57***
Propagation Strategy	Targeting	2.15***	1.91*
	Hubs	2.91***	-.3185
	Authorities	6.39***	0.51*
Community Structure	Following	1.5625***	0.9086***
	Followers	2.045*	0.9554***
	Interactions	1.441***	0.0112**
	Multiple	.468***	4.08***

Table 8: Evaluating strategies using logistic regression coefficients with p-value from t test (*-$p < 0.05$,**-$p < 0.01$,***-$p < 0.0001$)

To construct the set **v**, we collect a set of random users who posted with hashtags related to the given campaigns. We obtain the related hashtags from (RiteTag, 2014) by giving the initial hashtag as "#modi" for the dataset related to the elections and "#progun" for the dataset related to gun rights and assign the set of users posting using the hashtags as **v**. We randomly sampled **v**, and a very few number of users were ascertained as advocates, which we removed from the set. We collect the friends, followers, profile, and statuses of both advocates in **a** and random users in **v** for the two campaigns. Some statistics of the datasets is presented in Table 7.

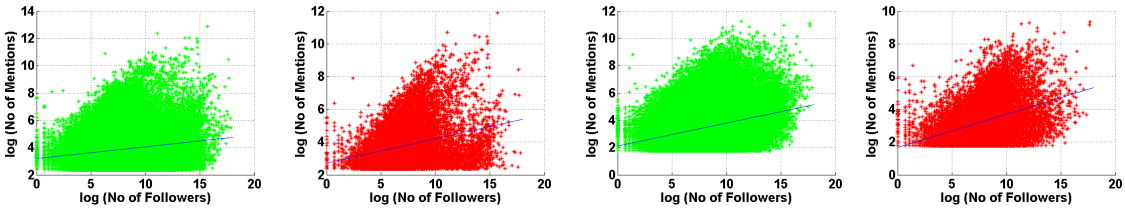
5.3.2 Evaluation

We now evaluate the characterizations of advocates to distinguish between advocates and random users posting on the issue using the datasets. Characterizations restricted to individual advocates such as persuasion, focus, emotion and propagation strategies are evaluated using logistic regression. The positive class is given by the advocates from the set **a** and the negative class consists of the random users posting on the campaign from **v**. The regression coefficients for all the characterizations along with the significance values derived from the t-test are shown in Table 8.

Characterizations of pairwise relational patterns such as shared language and community structure are evaluated using a paired t-test. We present the coefficients of the t-test along with its significance values in Table 8. We next evaluate the characterizations grouping them into message strategies, propagation strategies, and community structure.

5.3.3 Message Strategies

We first evaluate the ability of characterizations of message strategies to distinguish between advocates and random users posting on the campaign. From Table 8, we can see a strong evidence of the use of persuasive language by advocates in both the datasets, indicative of the attempts of advocates to sway the opinions of others. A significantly higher use of words denoting reason and affinity by advocates can be observed, demonstrating that the proposed model of persuasive language is effective. High positive and negative emotional content, which can be used to generate strong feelings about the campaign, is a strong characteristic in the tweets of advocates. This



(a) (b) (c) (d)

Figure 9: Interacting with Influencers (a) random users posting on the election campaign ($\rho = 0.08$) (b) advocates for election campaign ($\rho = 0.15$) (c) random users posting on gun rights ($\rho = 0.15$) (d) advocates for gun rights ($\rho = 0.20$)

indicates that unlike workers who are paid to comment or post on a particular issue leading to a lack of emotion in their posts (Lee *et al.*, 2013), advocates show a higher level of emotion in their messages.

A negative coefficient in the topical focus of advocates with high significance demonstrates that they concentrate their posts around fewer topics for effective messaging compared to random users posting on the campaign. In our experiments, we assign the number of topics t=20. The speech code theory states that users with common affiliations develop a common lingo to foster easier communication (Philipsen *et al.*, 1997). This is borne out by a significantly high coefficient of common hashtags, showing that advocates share higher amount of hashtags with other advocates than with random users posting on the campaign.

These observations indicate that the characterizations of message strategies drawn from theoretical constructs from sociological literature and the proposed approaches to modeling them are effective in distinguishing between advocates and random users posting on the campaign. We next evaluate the ability of the characterizations of the propagation strategies of advocates to distinguish between advocates and random users posting on the campaign.

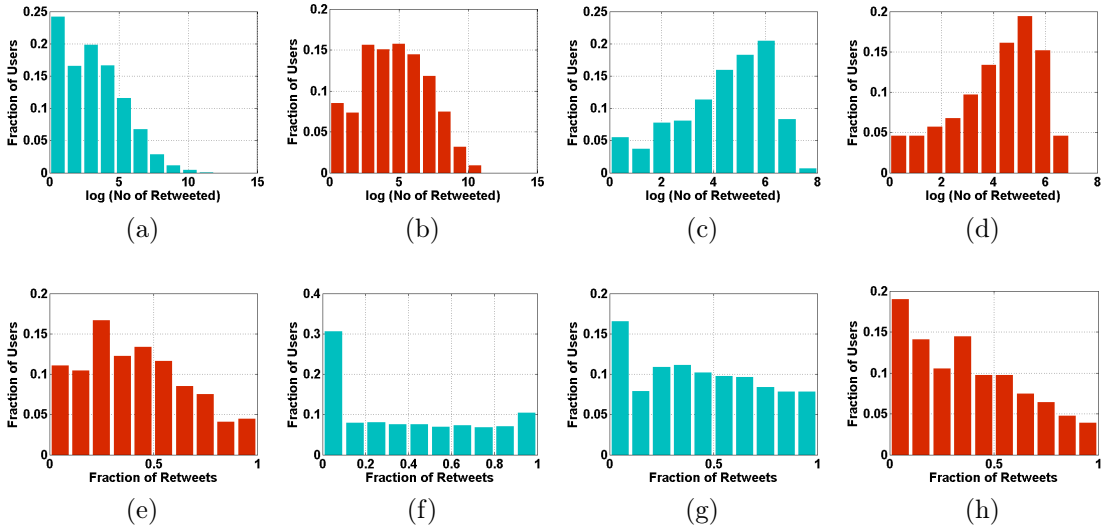


Figure 10: Retweet Patterns. Fraction of Users v/s Fractions of retweets in status for (a) random users posting on and (b) advocates for the election campaign, (c) random users posting on and (d) advocates related for gun rights. Fraction of users v/s no of times users are retweeted by (e) random users posting on and (f) advocates for election campaigns, (g) random users posting on, and (h) advocates for gun rights

5.3.4 Propagation Strategies

We first focus on the patterns arising from advocates targeting specific users, which can be performed in Twitter through the mention feature. The number of people who connect to the set of people targeted by the users in **u**, given by **c** in Section 5.2.2 can be measured by their follower count in Twitter. We first examine if the targeting patterns of advocates differ from those of normal users posting on the issue. We plot the targeting patterns of users in Fig 9 with the log of the number of mentions on the y-axis and the log of the number of followers on the x-axis. For each plot, we fit a line to the data that minimizes the least square error in the data and its slope, ρ, given in the caption. A higher value of ρ indicates that people with a higher number of followers are targeted more frequently.

We can see that mentioning patterns of advocates from Fig 9 that Fig 9b and 9d have a greater slope than that of random users posting on the campaign Fig 9a and 9c. Advocates might target popular users as they have the potential to increase the reach of information. We model this notion in Section 5.2.2 and present the results in Table 8. We can say from a significantly high coefficient value of the targeting characteristic that the targeting patterns of advocates are more significantly skewed towards popular users than those of random users posting on the campaign.

We next evaluate the characterizations of the co-propagation behavior of users in **u**. Fig 10 (a), (b), (c), and (d) illustrates the number of times each user has been retweeted. The logarithm of the times the user is retweeted in shown in the x-axis and the fraction of users who have been retweeted the corresponding number of times, choosing users with more than 100 tweets, on the y-axis. A larger fraction of advocates is retweeted than random users posting on the campaign.

For each user in **u**, we next compute the ratio of a number of their retweets to his total number of tweets. The retweet ratio is plotted on the x-axis and the fraction of users with the retweet ratio on the y-axis, choosing users with more than 100 tweets, in Fig 10 (e), (f), (g), and (h). A similar pattern can be observed, where the retweet ratio of random users posting on the campaign follows a power law distribution and those of advocates have a skewed distribution with a higher fraction of users actively involved in information propagation. These observations are reflected in Table 8, in the effectiveness of hubs and authority scores for distinguishing advocates and random users posting on the campaign.

These observations indicate that the proposed characterizations of propagation strategies drawn from sociological theories and approaches for modeling them can effectively distinguish between advocates and random users posting on the campaign.

We next evaluate the ability of the characterizations of the community structure in their ability to distinguish between advocates for political campaigns and random users posting on the campaign.

5.3.5 Community Structure

We first evaluate the characteristics of social relationships of advocates. From Table 8, we can see that advocates tend to connect with each other significantly more than with random users posting on the topic. This indicates advocates are utilizing social media to build a strong network of connections with each other. Advocates interact significantly more with other advocates for the campaign than with random users posting on it, indicating that they maintain a high level of interactions with each other. These provide evidence of a community structure between advocates.

We next examine if advocates tend to establish multiple types of relationships with each other. From Table 8, we see that given that an advocate has established one type of relationship with a user, he has a significantly higher propensity to form multiple types of relation if the user is another advocate than if he is a normal user posting on the campaign. These results indicate a strong network of social connections and interactions between advocates, verifying our postulates. This also provides a basis to infer community membership of users across different relationship types jointly.

Until now, we proposed a set of characteristics capable of effectively distinguishing between advocates and random users posting on the issue. The set of characterizations include both individual characteristics from their messages and propagation strategies and multiple relational patterns like social connections, shared content patterns, and interactions, forming a heterogeneous feature space. We next design a mathematical

formulation to combine individual and multiple pairwise characteristics in a unified framework to identify advocates for a political campaign.

5.4 A Unified Model

Integration of individual and relational characteristics in a unified, homogenous space for classification can be performed by deriving latent variables from the relational matrix (Tang and Liu, 2009). A difference here is that we have multiple types of pairwise relations between users, and we need to derive latent dimension memberships by jointly exploiting pairwise connections across multiple types of relations. Individual characteristics can be then be combined with the latent variables derived from the relational characteristics to construct features for classification.

Let the proposed individual characteristics are denoted as $\mathcal{I}$, and the relational characteristics are arranged in a tensor $\mathcal{Z}$ as defined in Section 5.1. Tensors have been used in literature to analyze jointly multi-modal relationships in applications such as community detection (Lin *et al.*, 2011) and link prediction (Dunlavy *et al.*, 2011). We first factorize the tensor $\mathcal{Z}$ to derive latent dimension memberships from the relational characteristics. Different types of connections between two users can be captured by their similarity in their memberships of latent dimensions. Two users who have pairwise relationships with each other across different types will have higher similarity in their latent dimension memberships than two users who do not have pairwise relationships with each other.

We factorize the tensor $\mathcal{Z}$ to obtain the user matrices $\mathbf{U} \in \mathbb{R}^{\mathrm{N}\times\mathrm{K}}$ and $\mathbf{V} \in \mathbb{R}^{\mathrm{N}\times\mathrm{K}}$, and the relationship type dimension matrix $\mathbf{T} \in \mathbb{R}^{\mathrm{T}\times\mathrm{K}}$, where K is the number of

latent dimensions, by solving the following optimization problem

$$\min_{\mathbf{U},\mathbf{V},\mathbf{T}} ||\mathcal{Z} - [\![\mathbf{U},\mathbf{V},\mathbf{T}]\!]||_{\mathrm{F}}^2, \tag{5.3}$$

where $[\![\mathbf{U},\mathbf{V},\mathbf{T}]\!] \in \mathbb{R}^{\mathrm{N}\times\mathrm{N}\times\mathrm{T}}$ is given by

$$[\![\mathbf{U},\mathbf{V},\mathbf{T}]\!] = \sum_{k=1}^{K} \mathbf{u}_k \circ \mathbf{v}_k \circ \mathbf{t}_k.$$

Here $\mathbf{u}_k$, $\mathbf{v}_k$, $\mathbf{t}_k$ are the k^{th} column vectors of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{T}$ respectively. The symbol $\circ$ represents the vector outer product such that if the tensor $\mathcal{Y} = \mathbf{u}_k \circ \mathbf{v}_k \circ \mathbf{t}_k$ then $\mathcal{Y}_{\mathrm{efg}} = (\mathbf{u}_k)_{\mathrm{e}}(\mathbf{v}_k)_{\mathrm{f}}(\mathbf{t}_k)_{\mathrm{g}}$. Substituting this in Eqn 5.3, we get

$$f = \min_{\mathbf{U},\mathbf{V},\mathbf{T}} ||\mathcal{Z} - \sum_{k=1}^{K} \mathbf{u}_k \circ \mathbf{v}_k \circ \mathbf{t}_k||_{\mathrm{F}}^2, \tag{5.4}$$

We optimize this function motivated by the conjugate linear optimization method (Acar *et al.*, 2011). We first arrange the vectors $\mathbf{u}$,$\mathbf{v}$ and $\mathbf{t}$ in a single vector $\mathbf{x} = [\mathbf{u},\mathbf{v},\mathbf{t}]$ and calculate the gradient of $\mathbf{f}(\mathbf{x})$ with respect to each $\mathbf{x}_k^n$ where $\mathbf{x}^1 = \mathbf{u}, \mathbf{x}^2 = \mathbf{v}, \mathbf{x}^3 = \mathbf{t}$. f can be rewritten as

$$f = \underbrace{\frac{1}{2}||\mathcal{Z}||^2}_{f_1} \underbrace{- \langle \mathcal{Z}, \sum_{k=1}^{K} \mathbf{x}_k^1 \circ \mathbf{x}_k^2 \circ \mathbf{x}_k^3 \rangle}_{f_2} + \underbrace{\frac{1}{2}||\sum_{k=1}^{K} \mathbf{x}_k^1 \circ \mathbf{x}_k^2 \circ \mathbf{x}_k^3||^2}_{f_3} \tag{5.5}$$

The gradient of f is obtained by computing its partial derivative with respect to each element in $\mathbf{x}$ denoted by $\mathbf{x}_k^m$. So

$$\frac{\partial f_1}{\partial \mathbf{x}_k^m} = 0 \tag{5.6}$$

as $\mathcal{Z}$ is a constant with respect to $\mathbf{x}_k^m$. The partial derivative of f_2 with respect to each element in $\mathbf{x}$ denoted by $\mathbf{x}_k^m$ can then be computed as follows.

$$\frac{\partial f_2}{\partial \mathbf{x}_k^1} = \mathcal{Z} \times_2 \mathbf{x}_k^2 \times_3 \mathbf{x}_k^3, \frac{\partial f_2}{\partial \mathbf{x}_k^2} = \mathcal{Z} \times_1 \mathbf{x}_k^1 \times_3 \mathbf{x}_k^3, \frac{\partial f_2}{\partial \mathbf{x}_k^3} = \mathcal{Z} \times_1 \mathbf{x}_k^1 \times_2 \mathbf{x}_k^2, \tag{5.7}$$

where $\times_n$ is the n-mode multiplication operator as defined as Section 5.1. The partial derivative of f_3 with respect to each term in $\mathbf{x}$ denoted by $\mathbf{x}_k^m$ can be computed as

$$\frac{\partial f_3}{\partial \mathbf{x}_k^m} = \sum_{j=1}^{K} (\prod_{\substack{r=1 \\ r \neq m}} \mathbf{x}_k^{r\mathrm{T}} \mathbf{x}_j^r) \mathbf{x}_j^r \tag{5.8}$$

The overall gradient can then be computed as

$$\frac{\partial f}{\partial \mathbf{x}_k^m} = \frac{\partial f_1}{\partial \mathbf{x}_k^m} - \frac{\partial f_2}{\partial \mathbf{x}_k^m} + \frac{\partial f_3}{\partial \mathbf{x}_k^m} \tag{5.9}$$

where $\frac{\partial f_1}{\partial \mathbf{x}_k^m}$, $\frac{\partial f_2}{\partial \mathbf{x}_k^m}$, $\frac{\partial f_3}{\partial \mathbf{x}_k^m}$ are as described in Eqn 5.6, Eqn 5.7 and Eqn 5.8 respectively. The gradient descent step repeated for all values of m and k is continued until convergence. As the objective function in Eqn 5.4 is convex, the optimization is guaranteed to converge. The computational complexity of the iterations is low due to the high sparsity of $\mathcal{Z}$.

To give an intuitive understanding of the optimization term in Eqn 5.4, we rewrite it as follows

$$\min_{\mathbf{U},\mathbf{V},\mathbf{T}} || \sum_{\mathrm{t}=0}^{\mathrm{T}} \mathcal{Z}_\mathrm{t} - \mathbf{U}\mathbf{D}_\mathrm{t}\mathbf{V}^\mathbf{T} ||_F^2, \tag{5.10}$$

where $\mathcal{Z}_\mathrm{t}$ represents the user relations of type t and $\mathbf{D}_\mathrm{t} \in \mathbb{R}^{\mathrm{K}\times\mathrm{K}}$ is a diagonal matrix whose diagonal elements are the t^{th} row of $\mathbf{T}$. The matrices $\mathbf{U}$ and $\mathbf{V}$ contains the latent dimension memberships of users jointly inferred across different relationship types. If $\mathcal{Z}_t \forall t$ is symmetric then $\mathbf{U} = \mathbf{V}$. The matrix $\mathbf{T}$ contains the contribution of each relation type to different dimensions. For example, a high value of $\mathbf{t_2}$ signifies that high-scoring users in $\mathbf{u_2}$ form connections with high-scoring users in $\mathbf{v_2}$ through relationship type 2. The latent features representing different kinds of pairwise relationships between users can be obtained from any linear combination of $\mathbf{U}$ and $\mathbf{V}$. For our experiments, we use $\mathbf{L} = \mathbf{U} + \mathbf{V}$ after column normalization as latent features. We combine the individual characteristics $\mathcal{I}$ and the latent features $\mathbf{L}$ to construct

Method	Elections		Gun Rights	
	AUC	F1	AUC	F1
Random	0.4983	0.1607	0.5053	0.1536
Retweet Ratio	0.5804	0.1797	0.5078	0.1118
Volume	0.6830	0.2406	0.6519	0.2332
Bag of Words	0.7379	0.3515	0.7305	0.2919
Combine	0.7599	0.3604	0.7460	0.3065
Our Method	**0.9301**	**0.6341**	**0.9431**	**0.6046**

Table 9: Comparison with different baselines.

a feature set $F = \{\mathcal{I}, \mathbf{L}\}$ for identifying advocates for political campaigns in social media.

5.5 Identifying Advocates

In this section, we evaluate the performance of our framework by answering the following questions. How effective is our framework for identifying advocates for political campaigns on social media? How good are the characteristics group in identifying advocates? How robust is the proposed framework for variation in training sizes?

5.5.1 Performance Evaluation

We classify the feature set derived in Section 5.4 using Linear Discriminant Analysis and perform 10-fold cross-validation to evaluate the performance of the framework in identifying advocates. We measure the performance using two metrics, AUC, and

F1-measure and present the results in Table 9. We have the following baselines to compare the performance of our framework.

- **Random :** We randomly assign labels to all the users.
- **Retweet Ratio (Lumezanu *et al.*, 2012):** The fraction of retweets per overall tweets is used as a feature.
- **Activity :** The total number of tweets of the users, is used for as a feature.
- **Bag of Words :** We use all the words in the status messages of users as features after tf-idf weighting,
- **Combine:** We combine all the proposed baselines by concatenating all the features.

We compare the performance of the baselines and the proposed framework and illustrate the results in Table 9. The random assignment gives an AUC value of around 0.5, and the F1 measure is low for both the datasets indicating the difficulty of the problem. The retweet ratio is used to characterize the propagating behavior of biased users in (Lumezanu *et al.*, 2012). We model a wider range of propagation strategies advocates employ in social media, and as we can see from Table 10, where our characterizations of propagation strategies outperform the retweet ratio. From Table 9, we see that advocates are more active than random users posting on the campaign. We model specific patterns in the messages based on the strategies of advocates and hence outperform this baseline that considers only the total number of messages.

The "Bag of Words" performs better than the other baselines, but the number of features here is high. We model specific strategies of advocates related to the message, propagation and community structure instead of using all the posted words, enabling us to outperform this baseline. Combining all the baselines gives a slight improvement

in the performance, indicating the potential benefits of integrating heterogeneous information. The proposed framework outperforms the baselines demonstrating that it effectively models and integrates characteristics useful for identifying advocates for political campaigns. This signifies the ability of the framework in understanding the strategies and model them to to identify advocates effectively. We perform the t-test between the results of our framework and the baselines and find that the difference is significant.

In summary, we can say that our framework outperforms the baselines demonstrating that it effectively models and integrates strategies useful for identifying advocates for political campaigns. We next analyze the contributions of different characteristic groups in identifying advocates.

5.5.2 Contributions of Characteristic Groups

We separately select characteristics related to message strategies, propagation strategies, and community structure and present them to the classifier. We compare the performance of different characteristic groups in identifying advocates using AUC and F1 measure with 10-fold cross validation and illustrate the results in Table 10.

We first examine the performance of message strategies by combining individual characteristics and the latent features from shared content. From Table 10, we can see that the characterizations of message strategies like persuasion, emotion, focus, and shared linguistic patterns outperform random characteristics by a significant margin. This demonstrates that proposed characterizations of messages strategies contribute significantly in identifying advocates.

The performance of characteristics related to propagation strategies is much

Method	Elections		Gun Rights	
	AUC	F1	AUC	F1
Random	0.4983	0.1607	0.5053	0.1536
Mess Strat	0.8240	0.4303	0.8517	0.4727
Prop Strat	0.8210	0.3689	0.5707	0.1904
Mess+ Prop Strat	0.8804	0.5152	0.8680	0.4934
Comm Struct	0.8918	0.5117	0.8859	0.4834
Overall	0.9301	0.6341	0.9431	0.6046

Table 10: Performance of different groups

higher in the dataset related to elections than in the dataset related to gun rights. This is an indication that advocates in election campaigns place more emphasis on information propagation. On combining characterizations from both message strategies with propagation strategies, we observe an improvement in performance. This demonstrates the contribution of characterizations of propagation strategies in identifying advocates.

The characteristics related to community structure perform well in both the datasets. This indicates that advocates in social media have strong relationships with other advocates of the issue and display strong interactions with each other. A combination of all the three characteristic groups performs significantly better than individual characteristic groups demonstrating the effectiveness of our framework for integrating these characterizations.

In summary, we can say that all the proposed characteristic groups contribute significantly to identifying advocates for political campaigns in social media. We next evaluate the robustness of the framework to variations of training data size and assess its effect on identifying advocates.

5.5.3 Performance with Varying Training Sizes

We now answer the following questions: How does the framework perform with varying proportions of training data? How effective is the framework when we use the labels of popular users, potentially more accessible, for training?

We vary the relative proportion of training and testing sizes from $\alpha = \{10\% - 90\%\}$ with increments of 10%. For each value of α, we take the mean performance of 100 random samples with $\alpha\%$ used for training and the rest for testing. We repeat the procedure for the different values of α and illustrate the results in Figure 11.

In many cases, the labels of popular users might be well known, and the labels might be therefore potentially easier to obtain. To evaluate the framework in this scenario, we take the number of followers of a user as an indication of his popularity and sort the users in decreasing number of their followers. For each value of α, we select the users in the top $\alpha\%$ of followers for training and the rest for testing and obtain the performance. We repeat the procedure for all the values of α and illustrate the results in Figure 11.We make the following observations from the figure.

The performance of the algorithm is significantly higher than the nearest baseline for all relative proportions of training data for both methods of sampling. The performance slightly increases for higher proportions of training data but is overall robust for varying size of training data. Finally, random selection of training data performs only slightly better than selecting users with a high number of followers. This demonstrates that labels of just the popular users, which are potentially easier to obtain, can be effective in identifying advocates for political campaigns on social media.

In summary, the experiment demonstrates that the framework is robust to variation

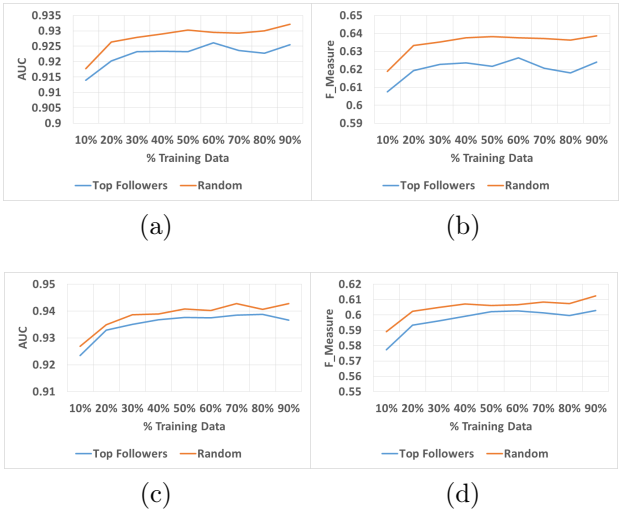


Figure 11: Effect of size of training data with training data randomly chosen and chosen ordered according to the number of followers for election dataset with (a) AUC (b) F1 measure and dataset on the gun rights with (c) AUC and (d) F1 Measure

in training data size and can also effectively identify advocates for political campaigns even when only the labels of popular users are known.

5.6 Summary

We present a framework to identify advocates for political campaigns on social media. We characterize advocates through their message strategies, propagation strategies, and community structure and propose different characterizations based on them. We evaluate the performance in identifying advocates of two political campaigns on the social media platform Twitter.

Chapter 6

IDENTIFYING RHETORICAL QUESTIONS

A popular way in which users express their views is through rhetorical questions(Paul *et al.*, 2011). Rhetorical questions are defined as "posts that have the form of a question but serve the function of a statement" (Anzilotti, 1982). For example, the rhetorical question "Would somebody willingly die for a claim he knew was a lie? ", where "he" refers to Christ, has the syntax of the question but is posted by the user to expresses his religious beliefs. Information seeking systems can be misled into finding responses to rhetorical questions, distracting them from addressing genuine information needs. Identifying them will assist in filtering them out.

Linguistic studies have provided theories for the function of the rhetorical questions by examining user motivations (Ilie, 1994). Rhetorical questions are stated as an indirect speech act (Schmidt-Radefeldt, 1977), meaning that the user posting rhetorical questions but implies the message from its context. It is hard to determine if the example question is rhetorical only from its text. However, when examining the post before the question, it is clear that the question is rhetorical. This indicates that rhetorical questions are likely to share context with the recent post of the user posting it.

The ability of rhetorical questions to imply a message can also be harnessed by the user to strengthen or mitigate a statement he previously made, as proposed in studies of conversation structure (Frank, 1990). In the previous example, the user posts the rhetorical question to mitigate the strong statement made in his most recent status message. Similarly, the user can employ rhetorical questions to strengthen a

statement in his last message. This indicates that rhetorical questions that it is likely to show a shift in the degree of sentiment from his last message.

I propose a framework to identify rhetorical questions in social media by modeling the motivations of the user to post them and evaluate it on two datasets of questions from the social media platform Twitter. Specifically; I address the following questions: How to model the motivations of the user for posting rhetorical questions to identify them? Are approaches based on motivations of the user useful in identifying rhetorical questions in social media? The primary contributions of the paper are the following:

- Formally defining the problem of identifying rhetorical questions in social media;
- Demonstrating the applicability of linguistic theories of user motivations to employ rhetorical questions in social media data;
- Proposing a framework to identify rhetorical questions in social media by modeling user motivations; and
- Evaluating the framework using two real-world datasets of questions posted on the social media platform Twitter.

6.1 Problem Statement

The outline of the proposed framework for identifying rhetorical questions is illustrated in Figure 12. We now define some terms in the framework related to the questions and the most recent status message from the user posting it. Let $\mathcal{R}$ denote the set of rhetorical questions, and the set of randomly sampled questions be denoted by $\mathcal{S}$. The combined set of questions is denoted by $\mathcal{F} = [\mathcal{R}, \mathcal{S}]$ and the total number of questions be denoted by Q. For each question $q \in \mathcal{F}$, we collect the most recent

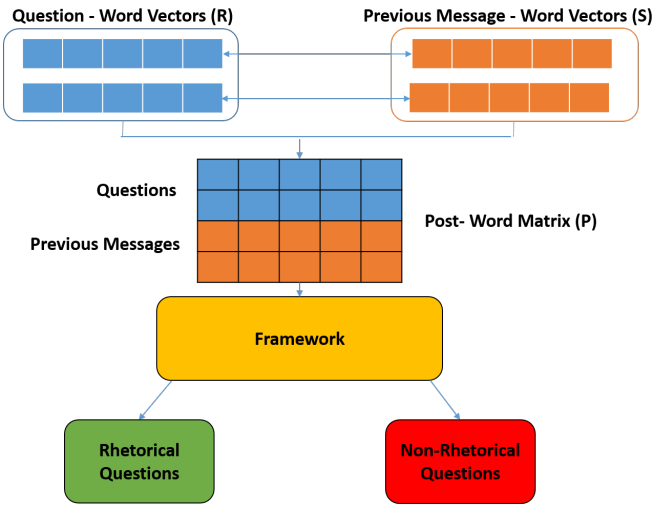


Figure 12: RhetId: The proposed framework to identify rhetorical questions in social media

status message the user posted previous to the question, and denote the set of most recent status messages as $\mathcal{M}$.

We construct a dictionary of words, $\mathcal{W}$, used in the questions and the last message of the users posting it. Let W be the number of words in the dictionary. We then construct word vectors from the content of the questions, the length of each being W. We then concatenate the question-word vectors to construct the question word matrix $\mathbf{F} \in \mathbb{R}^{Q \times W}$ from the question set $\mathcal{F}$, whose each element $\mathbf{F}_{ij}$ is given by

$$\mathbf{F}_{ij} = \begin{cases} \text{n} & q_i \text{ has used } w_j \text{ n times}, q_i \in \mathcal{F}, w_j \in \mathcal{W}, \\ 0 & \text{otherwise.} \end{cases} \tag{6.1}$$

Similarly, we construct message word vectors from the content of the previous message of the users posting the questions, each of length W. We concatenate the vectors to construct the message word matrix $\mathbf{M} \in \mathbb{R}^{Q \times W}$ from the set $\mathcal{M}$. The i^{th} row of the question word matrix $\mathbf{F}_i$ has a corresponding row $\mathbf{M}_i$ which contains word frequencies of the most recent status message of the user posting it. We concatenate the two

matrices $\mathbf{F}$ and $\mathbf{R}$ vertically to form the matrix $\mathbf{P} \in \mathbb{R}^{2Q \times W}$. The first Q rows of $\mathbf{P}$ contain the word frequencies of the questions from the matrix $\mathbf{F}$, and the last Q rows contain the word frequencies of the most recent status messages from the matrix $\mathbf{M}$. A single post will only use a small subset of the total number of words, resulting in $\mathbf{P}$ being sparse.

The problem can then be formally stated as follows: *"Given the question set $\mathcal{Q}$, consisting of a set of known rhetorical questions $\mathcal{R}$ and randomly sampled questions $\mathcal{S}$, and the post word matrix $\mathbf{P}$, determine if a new question q is a rhetorical question".*

6.2 Motivations behind Rhetorical Questions

Rhetorical questions share the form of a question, and it is difficult to identify them using syntactic characteristics (Li *et al.*, 2011). Understanding the motivations of the user in posting rhetorical questions, might give us more clues for identifying them. In this section, we will explore the possible motivations for users to post rhetorical questions. We draw concepts from linguistic literature to model these motivations, and design measures potentially useful for identifying rhetorical questions based on them. We collect two datasets of questions from the social media platform Twitter to evaluate the designed measures. We first present the two datasets with some relevant statistics. We then propose measures to quantify the concepts and use the datasets to verify if they are effective in identifying rhetorical questions in social media.

6.2.1 Datasets

The datasets consist of questions collected from the social media platform Twitter. To collect rhetorical questions, we use questions which the user has labeled as rhetorical with appropriate hashtags following (Ma *et al.*, 2014), where users have shown to employ hashtags to label their intention behind the tweet. We collect questions containing the hashtags related to rhetorical questions, along with "?" appended to each hashtag from the Twitter Streaming API, denoting them as positive examples for the two datasets respectively. Tweets containing "?" have been shown to be questions with high precision in (Cong *et al.*, 2008). We obtain the hashtags most related to rhetorical questions from (RiteTag, 2014) and construct two datasets with questions containing "#rhetoricalquestion" and "#dontanswerthat" as positive examples. To construct negative examples for the two datasets, we randomly sample some tweets equal to the number of positive examples containing "?". For each question in the two datasets, we collect the most recent status message from the user and construct the matrix $\mathbf{P}$ from the questions and most recent status messages using methods in the Twitter public API (Kumar *et al.*, 2013a). Some statistics of the datasets are given in Table 11.

The questions that constitute the negative examples might themselves have some rhetorical questions, so we use human assessments to validate the negative examples. The number of questions in the negative examples are large, and evaluating the entire set might be expensive. To address this, we apply the mark and recapture technique used in population estimation methodologies (Brower *et al.*, 1998). This technique involves drawing two random samples and using human assessments to estimate the

Parameter	Dataset 1	Dataset 2
# Questions	32,336	15,840
# Rhetorical Questions	16,168	7,920
# Randomly Sampled Questions	16,168	7,920
# Prev Status Messages	32,336	15,480
Avg Length (No of Words)	5.1300	5.3089
Word Frequency	1.7458	1.2424
Lexical Density	0.4678	0.4385

Table 11: Two Datasets containing questions posted in Twitter with relevant statistics. The positive examples of the first dataset contain the hashtag #rhetoricalquestion and the positive examples of the second dataset contain the hashtag #dontanswerthat

error. The error in the overall sample is then estimated from the number of errors in the two samples and their intersection.

We draw two random samples and assign the probability of finding errors in the random samples r_1 and r_2 be p_{r_1} and p_{r_2}. The total number of errors in the samples will be given by

$$e_{r_1} = p_{r_1} N_e, e_{r_2} = p_{r_2} N_e,$$

where N_e is the total number of errors to be estimated. The number of errors in the intersection of the two samples is then

$$e_{r_1 r_2} = p_{r_1} p_{r_2} N_e.$$

The number of errors in the dataset N_e , accounting for overestimation, is given by

$$N_e = \frac{(e_{r_1} + 1) \times (e_{r_2} + 1)}{e_{r_1 r_2} + 1} - 1, \tag{6.2}$$

with the standard error, SE, computed as

$$SE = \sqrt{\frac{(e_{r_1} + 1) \times (e_{r_2} + 1) \times (e_{r_2} - e_{r_1 r_2}) \times (e_{r_1} - e_{r_1 r_2})}{(e_{r_1 r_2} + 1)^2 \times (e_{r_1 r_2} + 2)}}. \tag{6.3}$$

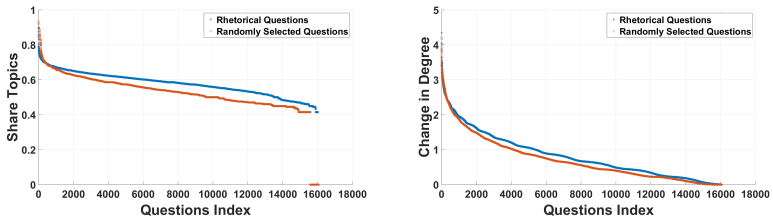
We draw two random samples of 1% of the size of the list from both the datasets and combined them. We then use human evaluators from Amazon Mechanical Turk

to verify them. The definition of rhetorical question is given to them as "posts that have the form of a question, but serve the function of a statement". The evaluators mark 1 if they think the question is a rhetorical question and 0 otherwise. The 95% confidence interval of the error is computed as $\text{Int}_\text{e} = \text{N}_\text{e} \pm 1.96\text{SE}$ and the accuracy of the negative examples as $1 - \text{Int}_\text{e}$. We compute the accuracy of the negative examples as $92.51\% \pm 8.3\%$. After validating the negative examples using the method described above, we use them in our experiments.

We next propose postulates to quantify the motivations of the user to post rhetorical questions drawing concepts from linguistic theories. We verify the postulates using the first dataset and later present the evaluation of the framework on both the datasets. We now state the postulates and using the dataset, examine if they can distinguish between rhetorical questions and randomly sampled questions.

6.2.2 Implying a Message

The first motivation for the user is to imply a message (Schmidt-Radefeldt, 1977), which indicates that the rhetorical question shares the context of his most recent status message. We characterize this by postulating "Rhetorical questions are more likely to share context with the most recent status message from the user than randomly sampled questions share with the most recent post of their user". We obtain the topical distribution of messages $\mathbf{P}$ using LDA (Blei *et al.*, 2003). We compute the cosine similarity in topic distributions of the question and the most recent message to measure the shared context between them and assign it to $\mathbf{s}_r$. We repeat this for a randomly selected question from the set of negative examples and assign it to $\mathbf{s}_n$. We repeat this procedure for all the rhetorical questions in $\mathcal{R}$.



(a) Shared topics (b) Degree change of sentiment

Figure 13: (a) Shared topics with the previous status message of rhetorical and randomly selected questions (b) Change of degree in sentiment of rhetorical and randomly selected questions from previous status messages. The x-axis in the two contains the questions arranged in descending order of the values of shared topics and degree change of sentiment respectively.

The results are plotted in Figure 13a. The quantity of topics rhetorical and randomly sampled questions share with their previous status messages is plotted in descending order on the y-axis. The figure indicates that rhetorical questions share a greater amount of topics with their previous status messages than randomly sampled questions. To test the significance of this observations, we perform a paired t-test. The null hypothesis is given by $\mathcal{H}_0 : \mathbf{s}_r = \mathbf{s}_n$ and the alternate hypothesis is given by $\mathcal{H}_1 : \mathbf{s}_r > \mathbf{s}_n$. A t-test verifies the postulate with $p < 0.0001$.

6.2.3 Modifying Expressed Sentiment

The second motivation for the user is to strengthen or mitigate the sentiment expressed in his last status message (Frank, 1990). This indicates that the rhetorical question will have a shift in the degree of the sentiment from his last status message. We state the postulate as "Rhetorical questions show a higher shift in the degree with the most recent status message of the user posting it than randomly sampled questions do with the last post of their user". We first construct the word level sentiment matrix

$\mathbf{O} \in \mathbb{R}^{W \times C}$ using (Hu *et al.*, 2013). Here C denotes the number of sentiment levels. To compute the sentiment in each word, we multiply it with the sentiment level vector $\mathbf{c} = \{-5, 5\}$. We then compute the sentiment of each tweet as $\mathbf{m} = \mathbf{POc}^T$. We select the elements of $\mathbf{m}$ containing the sentiment of the questions to construct $\mathbf{m}_q$ and the most recent status messages to construct $\mathbf{m}_p$. We compute the degree shift as $\mathbf{d} = |\mathbf{m}_q - \mathbf{m}_p|$. We select the elements corresponding to rhetorical questions and assign it to $\mathbf{d}_r$. We randomly select an equal number of elements of $\mathbf{d}$ corresponding to negative examples and assign it to $\mathbf{d}_n$. The paired t-test on $\mathbf{d}_r$ and $\mathbf{d}_n$ verifies the postulate with $p < 0.0001$.

The results are plotted in Figure 13b. The degree of sentiment shift in rhetorical and randomly sampled questions from their previous status messages is plotted in descending order on the y-axis. The figure indicates that rhetorical questions share a greater amount of topics with their previous status messages than randomly sampled questions. To test the significance of this observations, we perform a paired t-test. The null hypothesis is given by $\mathcal{H}_0 : \mathbf{d}_r = \mathbf{d}_n$ and the alternate hypothesis is given by $\mathcal{H}_1 : \mathbf{d}_r > \mathbf{d}_n$. A t-test verifies the postulate with $p < 0.0001$.

6.3 The Proposed Framework to Identify Rhetorical Questions

In this section, we present a framework to model the postulates and integrate them to identify rhetorical questions in social media. We first describe our approach to model the two motivations in detail and integrate them into an optimization function. We then present a method to solve the optimization function and derive a set of latent representations to be used for classification. We finally present the time complexity of the framework to analyze its scalability.

6.3.1 Modeling Shared Context

We now present a model for the first motivation. The first motivation that we explore is by a user to imply a message using the context of his recent message (Schmidt-Radefeldt, 1977). In the example introduced, the most recent status message before the question says "RT @PastorKentB: our pride keeps us from seeing who Jesus is... John 8...the Pharisees are too concerned with themselves to see the son of God!". In this example, it is clear that the asker is implying a message using the context derived from his previous post. This indicates that the question shares context with the most recent post of the user. The analysis in Section 6.2.2 showed that rhetorical questions share context with the previous message of their user as compared to randomly sampled questions. We next propose a method to compute the shared context between the question and the previous posts.

We first use latent dimensions to obtain concise representations of the questions and the previous messages. The latent dimension representation of $\mathbf{P}$ is given by the matrix $\mathbf{U} \in \mathbb{R}^{2Q \times K}$, with K latent dimensions. To capture the shared context between the question and the most recent message of the user, we make their latent dimensions in the corresponding rows of $\mathbf{U}$ close to each other. We do this by formulating a cost function that penalizes the distance between the corresponding rows of the matrix $\mathbf{U}$ and then minimizes the cost function. The cost function is defined as follows

$$\mathcal{F}_1 = \frac{1}{2} \sum_{i=1}^{2Q} \sum_{j=1}^{2Q} ||\mathbf{M}_{ij}\big(\mathbf{U}(i,*) - \mathbf{U}(j,*)\big)||_2^2, \tag{6.4}$$

where each element of matrix $\mathbf{M}$, $\mathbf{M}_{ij}$, is 1 if $|j - i| = Q$ and 0 otherwise. Note that in the matrix $\mathbf{P}$, the question and the previous statement of the user corresponding to it are Q rows apart. Therefore, the value of $\mathbf{M}_{ij}$ is 1 if row j of the matrix $\mathbf{P}$ contains the most recent message of the user posting the question whose word vector is in

a row i of $\mathbf{P}$. This loss function proposes a penalty if the latent dimensions of the question are far from the latent dimensions of the most recent status message of the user posting it, thus modeling the first motivation. The loss function can be rewritten as

$$\mathcal{F}_1 = \sum_{k=0}^{2Q} \mathbf{U}_k \mathcal{L} \mathbf{U}_k^T = tr(\mathbf{U}^T \mathcal{L} \mathbf{U}) = ||\mathbf{U}^T \mathcal{L}^{1/2}||_F^2, \tag{6.5}$$

where $\mathcal{L}$ is the Laplacian matrix of a graph whose adjacency matrix is $\mathbf{M}$. We next develop a loss function to model the second motivation concerning modifying the sentiment the user expresses in his previous message.

6.3.2 Modeling the Shift in the Expressed Sentiment

The second motivation that we explored is to strengthen or mitigate the sentiment expressed in the previous messages (Frank, 1990). In the example provided, the asker employs the rhetorical question "Would somebody willingly die for a claim he knew a lie?" is to mitigate the strength of his previous statement "RT @PastorKentB: our pride keeps us from seeing who Jesus is... John 8...the Pharisees are too concerned with themselves to see the son of God!". A possible reason can be that he considers the previous statement too offensive and he wants to make a milder statement to mitigate the effect of its earlier statements. This indicates that there might be a shift in the degree of the sentiment of the rhetorical question from the previous statement of the user. The analysis in Section 6.2.3 showed that rhetorical questions show a greater shift in the sentiment degree from the previous message of their user as compared to

randomly sampled questions. We next propose a method to compute the shift in the sentiment degree between the question and previous posts.

To obtain the sentiment distribution of each post, we compute $\mathbf{Q} \in \mathbb{R}^{2Q \times C}$ as $\mathbf{Q} = \mathbf{U}\mathbf{V}^T\mathbf{O}$. Here, we represent the latent dimensions of words present in the questions and the most recent status messages as $\mathbf{V} \in \mathbb{R}^{W \times K}$. We next need to model the notion that rhetorical questions show a greater shift in the sentiment degree from the previous message. To do this, we design a cost function that penalizes a small shift in the degree of sentiment between the two posts and then minimize the cost function. Let $\mathbf{q} \in \mathbb{R}^{1 \times C}$ and $\mathbf{p} \in \mathbb{R}^{1 \times C}$ be the sentiment distribution of the question and the most recent status message of the user. The loss function for this question can be defined as

$$f_2 = \mathbf{q}\mathbf{D}\mathbf{p}^T = \sum_{i}^{K}\sum_{j=1}^{K} \mathbf{q}_i \mathbf{D}_{ij} \mathbf{p}_j, \tag{6.6}$$

where

$$\mathbf{D}_{ij} = \begin{cases} \frac{c-1}{2} - |i-j| & \text{if } \left(\frac{c+1}{2} - i\right)\left(\frac{c+1}{2} - j\right) > 0. \\ \frac{c-1}{2} + 1 & \text{if } \left(\frac{c+1}{2} - i\right)\left(\frac{c+1}{2} - j\right) < 0. \\ 0 & \text{if } \left(\frac{c+1}{2} - i\right)\left(\frac{c+1}{2} - j\right) = 0. \end{cases} \tag{6.7}$$

The higher the degree of the sentiment shifts between $\mathbf{q}_i$ and $\mathbf{p}_j$, the lower the value of $\mathbf{D}_{ij}$. We also have a higher penalty if there is a change in sign. Aggregating this loss function over all the questions, we obtain the following loss function $\mathcal{F}_2 = tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q})$. We next integrate the two motivations into a unified framework to identify rhetorical questions.

6.3.3 Integrating the Models

Let us factorize the post-word matrix $\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0} ||\mathbf{P}-\mathbf{U}\mathbf{V}^T||_F^2$. We make the derived sentiment matrix $\mathbf{Q}$ closer to the post level sentiment matrix $\mathbf{T}$ obtained from (Hu *et al.*, 2013) using the regularization factor $||\mathbf{Q}-\mathbf{T}||_F^2$. We use the latent dimensions of the question from $\mathbf{U}$ to identify rhetorical questions with the least square loss function $||\mathbf{I}(\mathbf{U}\mathbf{W}-\mathbf{Y})||_F^2$. Here $\mathbf{I}\in\mathbb{R}^{Q\times Q}$ is a diagonal matrix where each diagonal element $\mathbf{I}_{ii}=1$, if the i^{th} question is labeled or 0 otherwise. $\mathbf{W}\in\mathbb{R}^{K\times 2}$ contains the weights given to each latent feature. $\mathbf{Y}\in\mathbb{R}^{2Q\times N}$ is the output of the classifier, and we label together the question and the most recent status message of the user. Each row of $\mathbf{Y}$ is given by {1,0} if the question labeled as rhetorical, {0,1} if the question is labeled as not rhetorical and {0,0}, if the question is not labeled. We then integrate this to form the final objective function, which is,

$$\begin{aligned}&\min_{\mathbf{U},\mathbf{V}\geq 0,\mathbf{W}} ||\mathbf{P}-\mathbf{U}\mathbf{V}^T||_F^2+\alpha||\mathbf{U}^T\mathcal{L}^{1/2}||_F^2+\beta tr((\mathbf{Q}\mathbf{D})^T\mathbf{C}\mathbf{Q})\\&+||\mathbf{I}(\mathbf{U}\mathbf{W}-\mathbf{Y})||_F^2+||\mathbf{Q}-\mathbf{T}||_F^2+\eta(||\mathbf{U}||_F^2+||\mathbf{V}||_F^2).\end{aligned}\tag{6.8}$$

Here, α and β control the contributions of the models based on the two motivations. We next minimize the optimization function to obtain the update rules for the latent dimension matrices $\mathbf{U}$, $\mathbf{V}$ and the weight matrix $\mathbf{W}$. We consider $\mathbf{U}$ and $\mathbf{V}$ as nonnegative to ensure an intuitive decomposition of the matrix $\mathbf{P}$ into its constituent parts. We randomly sample a fraction of the questions for training. We later use the matrices to determine whether the unlabeled question is rhetorical or not. The update rules for the three latent dimension matrices for $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ is provided below. The detailed derivation along with the complexity analysis to demonstrate the scalability

of the algorithm is illustrated in the appendix.

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij} \frac{(\mathbf{PV} + \mathbf{L}^- + \mathbf{M}^+ + \mathbf{TO}^T\mathbf{V})_{ij}}{(\mathbf{UV}^T\mathbf{V} + \alpha\mathcal{L}\mathbf{U} + \eta\mathbf{U} + \mathbf{L}^+ + \mathbf{M}^- + \beta\mathbf{AV}^T\mathbf{B} + \mathbf{QO}^T\mathbf{V})_{ij}}, \tag{6.9}$$

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij} \frac{(\mathbf{P}^T\mathbf{U} + \mathbf{OT}^T\mathbf{U})_{ij}}{(\mathbf{VU}^T\mathbf{U} + \beta\mathbf{BU}^T\mathbf{A} + \eta\mathbf{V} + \mathbf{OO}^T\mathbf{VU}^T\mathbf{U})_{ij}},$$

$$\mathbf{W} \leftarrow \mathbf{W} - 2\delta\mathbf{U}^T\mathbf{I}(\mathbf{UW} - \mathbf{Y}), \tag{6.10}$$

where $\mathbf{A} = \mathbf{CU}$, $\mathbf{B} = \mathbf{OD}^T\mathbf{O}^T\mathbf{V}$, $\mathbf{L}^s = (\mathbf{IYW}^T)^s$, $\mathbf{M}^s = (\mathbf{IUWW}^T)^s$, $s = \{+, -\}$.

6.3.4 Deriving the Question Labels

We now present a procedure to employ the derived matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ to obtain the labels denoting whether a question is rhetorical or not. We compute $\hat{\mathbf{Y}}$, the estimated value of $\mathbf{Y}$ as

$$\hat{\mathbf{Y}} = \mathbf{UW}. \tag{6.11}$$

We then select the rows of $\hat{\mathbf{Y}}$ corresponding to the questions in the test dataset to construct $\hat{\mathbf{Y}}_{test} \in \mathbb{R}^{Q \times 2}$. For each row of $\hat{\mathbf{Y}}_{test}$, we compare the values in the two columns. We assign the question in the row as rhetorical if the value in the first column is greater than the second column and not rhetorical otherwise.

The framework is summarized in **Algorithm 1**. The inputs of the framework is the post word matrix $\mathbf{P} \in \mathbb{R}^{2Q \times W}$ and the parameter values of α, β and η. The output of the framework is the labels of the questions as rhetorical or not. We first randomly initialize the latent dimension matrices $\mathbf{U}$, $\mathbf{V}$ and the weight matrix $\mathbf{W}$. We then compute the post sentiment matrix $\mathbf{Q}$. We then randomly pick questions for training and change the corresponding rows of the label matrix $\mathbf{Y}$ and the indicator matrix $\mathbf{I}$. We then compute the objective function from Eqn 6.10. We use alternate gradient descent to minimize the objective function in iterations, and the latent matrices $\mathbf{U}$,

ALGORITHM 4: RhetId: Identifying Rhetorical Questions

Data: $\mathbf{P}$, α, β, η
Result: Labels of Questions as rhetorical or not.
Randomly initialize $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$;
Compute post sentiment matrix $\mathbf{Q}$;
Randomly pick rows of $\mathbf{Y}$ to assign labels ;
Construct $\mathbf{I}$;
Compute objective function F from Eq.(6.8);
while *F does not converge* **do**
    Update $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ from Eq. (6.10);
    Update F;
end
Compute $\hat{\mathbf{Y}} = \mathbf{UW}$ and obtain $\mathbf{Y}_{test}$;
Obtain labels of the test questions from $\mathbf{Y}_{test}$;

$\mathbf{V}$ and the weight matrix $\mathbf{W}$ are updated in each iteration. The objective function is updated in each iteration, and the procedure is repeated until the function value converges. The label matrix is then estimated by Eqn 6.11 and the labels of the unlabeled questions are obtained from the corresponding rows of $\hat{\mathbf{Y}}_{test}$ by comparing the values in the two columns.

6.4 Derivation of Latent Dimension Matrices

We here provide detailed derivation of the update equations in latent dimension matrices $\mathbf{U}$, $\mathbf{V}$ and the weight matrix $\mathbf{W}$.

6.4.1 Computation of document-latent dimension matrix $\mathbf{U}$

We now present the closed form solution to the minimization problem in Eq.(6.8) to obtain the latent representations of questions and use it to identify rhetorical

questions. Motivated by (Ding *et al.*, 2006), we first introduce an algorithm to find optimal solutions for the three matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{W}$. The key idea is to optimize the objective on one variable while fixing others. The algorithm will keep updating the matrices until convergence with the following update equations. We first present a derivation of the latent dimension matrix $\mathbf{U}$ from the objective function and follow a similar procedure to derive latent dimension matrix $\mathbf{V}$ and the weight matrix $\mathbf{W}$. Solving the optimization function in Eq. (6.8) on $\mathbf{U}$, we get

$$\begin{aligned}\min_{\mathbf{U}\geq 0} \mathcal{J}_U =& ||\mathbf{P}-\mathbf{U}\mathbf{V}^T||_F^2 + \alpha||\mathbf{U}^T\mathcal{L}^{1/2}||_F^2 + \beta tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q}). \\ & ||\mathbf{I}(\mathbf{U}\mathbf{W}-\mathbf{Y})||_F^2 + \eta||\mathbf{U}||_F^2 + ||\mathbf{Q}-\mathbf{T}||_F^2.\end{aligned} \tag{6.12}$$

Let Λ_U, be the Langrangian multiplier for the condition $\mathbf{U} \geq 0$. The lagrangian function $\mathcal{L}(\mathbf{U})$ is then given as

$$\begin{aligned}\mathcal{L}(\mathbf{U}) =& \mathcal{J}_U - Tr(\Lambda_U\mathbf{U}). \\ =& ||\mathbf{P}-\mathbf{U}\mathbf{V}^T||_F^2 + \alpha||\mathbf{U}^T\mathcal{L}^{1/2}||_F^2 + \beta Tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q}) + ||\mathbf{I}(\mathbf{U}\mathbf{W}-\mathbf{Y})||_F^2 + \eta||\mathbf{U}||_F^2 \\ & + ||\mathbf{Q}-\mathbf{T}||_F^2 - Tr(\Lambda_U\mathbf{U}). \\ =& Tr(\mathbf{P}-\mathbf{U}\mathbf{V}^T)(\mathbf{P}-\mathbf{U}\mathbf{V}^T)^T + \alpha Tr(\mathbf{U}^T\mathcal{L}^{1/2})(\mathbf{U}^T\mathcal{L}^{1/2})^T + \beta Tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q}) \\ & + Tr(\mathbf{I}(\mathbf{U}\mathbf{W}-\mathbf{Y}))(\mathbf{I}(\mathbf{U}\mathbf{W}-\mathbf{Y}))^T + Tr(\mathbf{Q}-\mathbf{T})(\mathbf{Q}-\mathbf{T})^T + \eta Tr(\mathbf{U}\mathbf{U}^T). \\ =& Tr(\mathbf{P}\mathbf{P}^T - \mathbf{P}\mathbf{V}\mathbf{U}^T - \mathbf{U}\mathbf{V}^T\mathbf{P}^T + \mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T + \alpha\mathbf{U}^T\mathcal{L}\mathbf{U} + \beta\mathbf{D}^T\mathbf{Q}^T\mathbf{M}\mathbf{Q} + \\ & \mathbf{I}\mathbf{U}\mathbf{W}\mathbf{W}^T\mathbf{U}^T\mathbf{I}^T - \mathbf{I}\mathbf{U}\mathbf{W}\mathbf{Y}^T\mathbf{I}^T - \mathbf{I}\mathbf{U}\mathbf{W} - \mathbf{I}\mathbf{Y}\mathbf{W}^T\mathbf{U}^T\mathbf{I}^T + \mathbf{I}\mathbf{Y}\mathbf{Y}^T\mathbf{I}^T + \mathbf{Q}\mathbf{Q}^T + \mathbf{T}\mathbf{T}^T \\ & - 2\mathbf{Q}\mathbf{T}^T + \eta\mathbf{U}\mathbf{U}^T - \Lambda_U\mathbf{U}).\end{aligned} \tag{6.13}$$

By setting the derivative of $\mathcal{L}(\mathbf{U})$ with respect to $\mathbf{U}$ $\frac{\partial \mathcal{L}(\mathbf{U})}{\partial \mathbf{U}} = 0$, we get

$$\frac{1}{2}\Lambda_U = -\mathbf{PV} + \mathbf{UV}^T\mathbf{V} + \alpha\mathcal{L}\mathbf{U} + \mathbf{IUWW}^T - \mathbf{IYW}^T + \eta\mathbf{U} + \beta\mathbf{AV}^T\mathbf{B} + \mathbf{QO}^T\mathbf{V} - \mathbf{TO}^T\mathbf{V}, \tag{6.14}$$

where $\mathbf{A} = \mathbf{MU}$, $\mathbf{B} = \mathbf{OD}^T\mathbf{O}^T\mathbf{V}$.

The matrix $\mathbf{W}$ can have negative elements so we split the components containing $\mathbf{W}$, $\mathbf{IYW}^T$ and $\mathbf{IUWW}^T$ into positive and negative elements described as follows. Let a given matrix be denoted as $\mathbf{E}$ and $\mathbf{E}^+$ and $\mathbf{E}^-$ be the positive and negative components of $\mathbf{E}$ respectively. These matrices are defined as

$$\mathbf{E} = \mathbf{E}^+ - \mathbf{E}^-, \mathbf{E}^+ = \frac{|\mathbf{E}| + \mathbf{E}}{2}, \mathbf{E}^- = \frac{|\mathbf{E}| - \mathbf{E}}{2}. \tag{6.15}$$

Substituting this in the Eq. (6.14)

$$\begin{aligned}\Lambda_U = &-\mathbf{PV} + \mathbf{UV}^T\mathbf{V} + \alpha\mathcal{L}\mathbf{U} + (\mathbf{IYW}^T)^+ - (\mathbf{IYW}^T)^- - \eta\mathbf{U} - (\mathbf{IUWW}^T)^+ \\ &+ (\mathbf{IUWW}^T)^- + \beta\mathbf{AV}^T\mathbf{B} + \mathbf{QO}^T\mathbf{V} - \mathbf{TO}^T\mathbf{V}.\end{aligned} \tag{6.16}$$

The Karush-Kuhn-Tucker condition (Boyd and Vandenberghe, 2004) for the non-negative constraint $\mathbf{U} \geq 0$ gives

$$\Lambda_U(i,j)\mathbf{U}(i,j) = 0. \tag{6.17}$$

We substitute Eq. (6.14) in Eq. (6.17) and rearrange this equation to get the update rule for $\mathbf{U}$ as

$$\mathbf{U}_{ij} \leftarrow \mathbf{U}_{ij}\frac{(\mathbf{PV} + \mathbf{M}^- + \mathbf{L}^+ + \mathbf{TO}^T\mathbf{V})_{ij}}{(\mathbf{UV}^T\mathbf{V} + \alpha\mathcal{L}\mathbf{U} + \eta\mathbf{U} + \mathbf{M}^+ + \mathbf{L}^- + \beta\mathbf{AV}^T\mathbf{B} + \mathbf{QO}^T\mathbf{V})_{ij}}, \tag{6.18}$$

where $\mathbf{A} = \mathbf{MU}$, $\mathbf{B} = \mathbf{OD}^T\mathbf{O}^T\mathbf{V}$, $\mathbf{L}^s = (\mathbf{IYW}^T)^s$, $\mathbf{M}^s = (\mathbf{IUWW}^T)^s$.

6.4.2 Computation of word-latent dimension matrix $\mathbf{V}$

We follow a similar procedure to derive the latent dimension matrix $\mathbf{V}$ from the objective function. Solving the optimization function in Eq. (6.8) with respect to $\mathbf{V}$, we get

$$\min_{\mathbf{V}\geq 0} \mathcal{J}_V = ||\mathbf{P} - \mathbf{U}\mathbf{V}^T||_F^2 + \beta tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q}) + ||\mathbf{Q} - \mathbf{T}||_F^2 + \eta(||\mathbf{V}||_F^2). \tag{6.19}$$

Let Λ_V, be the langrangian multiplier for the condition $\mathbf{V} \geq 0$. The langragian function $\mathcal{L}(\mathbf{V})$ is then given as

$$\begin{aligned}
\mathcal{L}(\mathbf{V}) =& \mathcal{J}_V - Tr(\Lambda_V \mathbf{V}). \\
=& ||\mathbf{P} - \mathbf{U}\mathbf{V}^T||_F^2 + \beta Tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q}) + +\eta||\mathbf{V}||_F^2 + ||\mathbf{Q} - \mathbf{T}||_F^2 - Tr(\Lambda_U \mathbf{U}) \\
=& Tr(\mathbf{P} - \mathbf{U}\mathbf{V}^T)(\mathbf{P} - \mathbf{U}\mathbf{V}^T)^T + \beta Tr((\mathbf{Q}\mathbf{D})^T\mathbf{M}\mathbf{Q}) + Tr(\mathbf{Q} - \mathbf{T})(\mathbf{Q} - \mathbf{T})^T + \eta Tr(\mathbf{V}\mathbf{V}^T) \\
=& Tr(\mathbf{P}\mathbf{P}^T - \mathbf{P}\mathbf{V}\mathbf{U}^T - \mathbf{U}\mathbf{V}^T\mathbf{P}^T + \mathbf{U}\mathbf{V}^T\mathbf{V}\mathbf{U}^T + \beta\mathbf{D}^T\mathbf{Q}^T\mathbf{M}\mathbf{Q} + \mathbf{Q}\mathbf{Q}^T + \mathbf{T}\mathbf{T}^T - 2\mathbf{Q}\mathbf{T}^T \\
& \eta\mathbf{V}\mathbf{V}^T - \Lambda_U\mathbf{V})
\end{aligned} \tag{6.20}$$

By setting the derivative of $\mathcal{L}(\mathbf{V})$ with respect to $\mathbf{V}$, $\frac{\partial \mathcal{L}(\mathbf{V})}{\partial \mathbf{V}} = 0$, we get

$$\frac{1}{2}\Lambda_V = -\mathbf{P}^T\mathbf{U} + \mathbf{V}\mathbf{U}^T\mathbf{U} - \mathbf{O}\mathbf{T}^T\mathbf{U} + \beta\mathbf{B}\mathbf{U}^T\mathbf{A} + \eta\mathbf{V} + \mathbf{O}\mathbf{O}^T\mathbf{V}\mathbf{U}^T\mathbf{U}, \tag{6.21}$$

where $\mathbf{A} = \mathbf{M}\mathbf{U}$, $\mathbf{B} = \mathbf{O}\mathbf{D}^T\mathbf{O}^T\mathbf{V}$. The Karush-Kuhn-Tucker condition (Boyd and Vandenberghe, 2004) for the non-negative constraint $\mathbf{V} \geq 0$ gives

$$\Lambda_V(i,j)\mathbf{V}(i,j) = 0. \tag{6.22}$$

We substitute Eq. (6.21) in Eq.(6.22) and rearrange this equation to get the update rule for $\mathbf{V}$ as

$$\mathbf{V}_{ij} \leftarrow \mathbf{V}_{ij} \frac{(\mathbf{P}^T\mathbf{U} + \mathbf{O}\mathbf{T}^T\mathbf{U})_{ij}}{(\mathbf{V}\mathbf{U}^T\mathbf{U} + \beta\mathbf{B}\mathbf{U}^T\mathbf{A} + \eta\mathbf{V} + \mathbf{O}\mathbf{O}^T\mathbf{V}\mathbf{U}^T\mathbf{U})_{ij}}, \tag{6.23}$$

where $\mathbf{A} = \mathbf{M}\mathbf{U}$, $\mathbf{B} = \mathbf{O}\mathbf{D}^T\mathbf{O}^T\mathbf{V}$.

6.4.3 Computation of feature weight matrix $\mathbf{W}$

We next follow a similar procedure to derive the weight matrix $\mathbf{W}$ from the objective function. Unlike the latent dimension matrices $\mathbf{U}$ and $\mathbf{V}$, the weight matrix is not constrained to be non-negative. We, therefore, solved it by gradient descent with additive updates. Solving the optimization function in Eq. (6.8), we get

$$\min_{\mathbf{W}} \mathcal{J}_W = ||\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y})||_F^2. \tag{6.24}$$

$\mathcal{J}_W$ can be expanded as

$$\begin{aligned} \mathcal{J}_W =& Tr(\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y}))(\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y}))^T \\ =& Tr(\mathbf{I}\mathbf{U}\mathbf{W}\mathbf{W}^T\mathbf{U}^T\mathbf{I}^T - \mathbf{I}\mathbf{U}\mathbf{W}\mathbf{Y}^T\mathbf{I}^T - \mathbf{I}\mathbf{U}\mathbf{W} - \mathbf{I}\mathbf{Y}\mathbf{W}^T\mathbf{U}^T\mathbf{I}^T + \mathbf{I}\mathbf{Y}\mathbf{Y}^T\mathbf{I}^T). \end{aligned} \tag{6.25}$$

The gradient $\mathcal{J}_W$ with respect to $\mathbf{W}$, $\frac{\partial \mathcal{J}_W}{\partial \mathbf{W}} = 0$, is given by

$$\frac{\partial \mathcal{J}_W}{\partial \mathbf{W}} \leftarrow \mathbf{U}^T\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y}). \tag{6.26}$$

We obtain the updates by gradient descent with the following update equation

$$\mathbf{W} \leftarrow \mathbf{W} - 2\delta\mathbf{U}^T\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y}), \tag{6.27}$$

where δ is the gradient step.

6.5 Algorithm Complexity

The complexity of the algorithm comes mainly from two sources: the computation of the objective function in Eq. (6.8) and the update equations in Step 7 of **Algorithm 1**.

We first concentrate on the objective function. Computing the first term $||\mathbf{P} - \mathbf{U}\mathbf{V}^T||_F^2$ is low owing to the sparse nature of $\mathbf{P}$. The computational complexity of the second term $||\mathbf{U}^T\mathcal{L}^{1/2}||_F^2$ is low due to the sparsity of the Laplacian matrix $\mathcal{L}$. The computation of $\mathbf{Q}$ takes $O(WKC)$. The third and the fourth term each has a complexity of $O(QC^2)$. The complexity of $||\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y})||_F^2$ is $O(QK)$ multiplications. The complexity is considerably lessened due to the low value of K and the sparsity of $\mathbf{P}$ making the computation scalable.

We next focus on the complexity of the update equations. The complexities of $\mathbf{P}\mathbf{V}$ and $\mathbf{P}^T\mathbf{U}$ are low due to the sparsity of $\mathbf{P}$. The complexity of $\mathbf{I}\mathbf{Y}\mathbf{W}^T$ is $O(Q)$. The terms $\mathbf{U}\mathbf{V}^T\mathbf{V}$ and $\mathbf{V}\mathbf{U}^T\mathbf{U}$ have a time complexity of $O(WK^2)$. The time complexity of $\mathbf{I}\mathbf{U}\mathbf{W}\mathbf{W}^T$ is $O(QK^2)$. The computation of $\mathbf{A}$ has a low complexity due to the sparsity of $\mathbf{C}$. The complexity of $\mathbf{B}$ is $O(WCK)$ is reduced due to the low value of K and C. The complexity of $\mathbf{A}\mathbf{V}^T\mathbf{B}$ and $\mathbf{B}\mathbf{U}^T\mathbf{A}$ is $O(WK^2)$. The complexity of $\mathbf{Q}\mathbf{O}^T\mathbf{V}$ is $O(WK^2)$ and $\mathbf{O}\mathbf{O}^T\mathbf{V}\mathbf{U}^T\mathbf{U}$ is $O(WK^2)$. The complexity of $\mathbf{U}^T\mathbf{I}(\mathbf{U}\mathbf{W} - \mathbf{Y})$ is $O(QK)$.This complexity of the update equations is low owing to the low value of K and C.

From the above discussion, we can say that the framework is scalable and hence can be used to large datasets usually seen in social media. We next design experiments to evaluate the framework to identify rhetorical questions in social media data.

6.6 Experimental Evaluation

In this section, we use the two datasets of questions posted on Twitter described in Table 11 to conduct experiments to answer the following questions that help in understanding the concepts involved in the framework better: How does the proposed framework perform in identifying rhetorical questions in social media? What is the effect of the varying proportion of information from different motivations models on the performance of the framework? What is the effect of varying proportions of training data on the performance of the framework? We first describe the experiment settings and then address each of these questions.

6.6.1 Experimental Settings

We now present the metrics and baselines used for evaluating the algorithm. We use accuracy, AUC and F1 metrics to evaluate the algorithm and values are averaged over the positive and negative class. We choose AUC and F1 as they can handle the imbalance in the dataset. The following baselines are employed as performance benchmarks and compared with our framework.

- **Random Label Assignment**: Whether the question is rhetorical or not is assigned randomly. This is repeated for 100 trials, and the mean value of the metrics is presented. This baseline is employed to demonstrate the difficulty of the problem.
- **Topics**: We construct topic distributions of the questions using (Blei *et al.*, 2003). We present the predicted topics for classifying whether the questions are rhetorical or not rhetorical.

- **BOW**: We use the entire content contained in the candidate questions and present it for classifying whether the questions are rhetorical or not. The baselines **Topics** and **BOW** are introduced to evaluate the content information in the question for identifying whether the question is rhetorical.
- **InfoNeeds** (Zhao and Mei, 2013): The paper uses features such as tweet length and the use of capital letters in addition to **BOW** to classify whether a question posted in social media is seeking for information.
- **Qweet** (Li *et al.*, 2011): The algorithm uses linguistic features such as quotations, exclamations in addition to **BOW** to classify whether a question posted in social media is seeking information. We employ the baselines **Qweet** and **InfoNeeds** to determine if identifying rhetorical questions is similar to identifying questions not conveying information needs.
- **PrevMsg** (Bhattasali *et al.*, 2015): The algorithm identifies rhetorical questions by directly combining the textual features of questions and the neighboring statements. This baseline is included to evaluate whether information from the context of the question is useful in identifying whether the question is rhetorical or not.
- **RhetId (SC)**: This implements our algorithm using only the motivation of shared context (SC) i.e $\alpha = 1, \beta = 0$ in Eq. (6.8). This is introduced to assess the contribution of the motivation for implying a message in identifying rhetorical questions in social media
- **RhetId (SS)**: This baseline uses only the motivation of sentiment shift (SS), i.e., $\alpha = 0, \beta = 1$ in Eq. (6.8). This is introduced to assess the contribution of this motivation for modifying the previously expressed sentiment identifying rhetorical questions in social media.

We evaluate the proposed framework and the baselines with the following metrics suitable for the unbalanced datasets

Area Under the Curve (AUC): This metric computes the area under curve created by plotting the true positive rate against the false positive rate. The true positive rate defines how many of the rhetorical questions are identified correctly during the test. The false positive rate determines how many of the non-rhetorical questions are correctly identified as not rhetorical by the framework. The framework is repeatedly applied on the data to create the curve plotting the true positive rate and the false positive rate. The area under this curve is then computed and presented to evaluate the framework.

Accuracy (Accu): The accuracy computes the ratio of total number of questions correctly identified as either rhetorical or not rhetorical and the total number of questions. The accuracy is computed as

$$\text{Accu} = \frac{\sum_{\text{i=1}}^{\text{R}} \text{TP}_i + \sum_{\text{j=1}}^{\text{S}} \text{TN}_j}{\text{R} + \text{S}},$$

where TP_i denotes the true positives and is 1 if the i^{th} rhetorical question is identified as a rhetorical question and 0 otherwise and R denotes the number of rhetorical questions. Similarly, TN_j denotes the true negatives and is 1 if the j^{th} non-rhetorical question is identified as a non-rhetorical question and 0 otherwise, and S denotes the number of non-rhetorical questions.

F1 Measure (F1): The precision computes the ratio of the number of correctly identified rhetorical questions and the total of all correctly identified questions. The recall measures the ratio of the number of correctly identified rhetorical questions to the total number of rhetorical questions. The F1-measure computes the harmonic

Methods	Dataset 1 AUC	Dataset 1 Accuracy	Dataset 1 F1	Dataset 2 AUC	Dataset 2 Accuracy	Dataset 2 F1
Random	50.66	50.48	50.46	50.27	49.93	49.42
Topics	63.28	56.73	56.72	65.92	59.39	58.39
BOW	64.54	61.24	61.15	66.82	63.15	63.11
PrevMsg	65.35	63.86	63.93	66.76	65.47	65.48
InfoNeeds	67.55	61.43	61.04	67.72	64.06	64.03
Qweet	63.50	64.07	64.07	67.41	64.82	64.72
RhetId (SC)	**71.45**	**64.24**	**63.63**	**73.71**	**66.29**	**68.15**
RhetId (SS)	**69.49**	**63.17**	**62.83**	**72.10**	**64.88**	**64.69**
RhetId (SC&SS)	**72.61**	**65.31**	**65.07**	**74.08**	**67.04**	**69.73**

Table 12: Performance evaluation of the algorithm. It beats the proposed baselines by a significant margin in both the datasets.

mean of precision and recall and is denoted by

$$\text{F1} = \frac{2\sum_{\text{i=1}}^{\text{R}} \text{TP}_i}{2\sum_{\text{i=1}}^{\text{R}} \text{TP}_i + \sum_{\text{j=1}}^{\text{S}} \text{FP}_j + \sum_{\text{i=1}}^{\text{R}} \text{FN}_i},$$

where TP_i denotes the true positives and is 1 if the i^{th} rhetorical question is identified as a rhetorical question and 0 otherwise, R denotes the number of rhetorical questions and S denotes the number of non-rhetorical questions. FN_j denotes the false negatives and is 1 if the i^{th} rhetorical question is identified as a non-rhetorical question and 0 otherwise. Similarly, FP_j denotes the false positives and is 1 if the j^{th} non-rhetorical question is identified as a rhetorical question and 0 otherwise.

6.6.2 Performance Evaluation

We now evaluate the performance of the algorithm using Accuracy, AUC, and F1 measure, and compare it with the baselines. We set the number of latent dimensions $I = 50$. We randomly select 50% of the candidate questions for training and the rest for testing, experimenting with different values of α and β. Later in the section, we

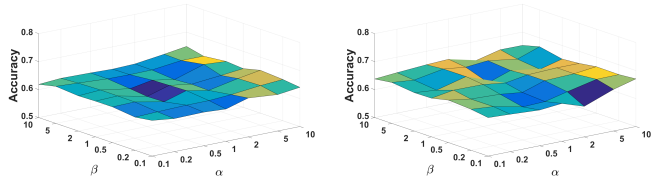
will study the variation of the performance with the various values of the parameters and training data size. We illustrate the results in Table 12.

From Table 12, we see that the baseline **Topics** shows an improvement over the performance of the random assignment in both the datasets showing linguistic characteristics are useful to identify rhetorical questions. The performance, however, is worse than **BOW**. This indicates that there is not a large difference in the topics of rhetorical questions and randomly sampled questions.

The algorithms **InfoNeeds** (Zhao and Mei, 2013) and **Qweets** (Li *et al.*, 2011) give a small improvement over **BOW** in both the datasets indicating that linguistic characteristics for categorizing information seeking questions are not similar to the characteristics of rhetorical questions. Hence, we need to use concepts unique to rhetorical questions to identify them. The algorithm **PrevMsg** (Bhattasali *et al.*, 2015) gives an improvement over BOW in both the datasets, showing the importance of contextual information for the identification of rhetorical questions.

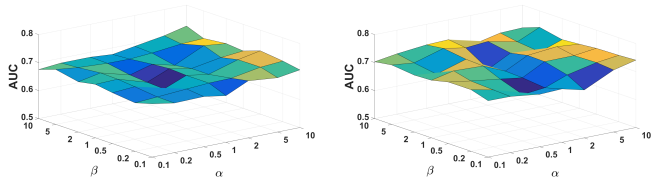
The improvement by **RhetId (SC)** and **RhetId (SS)** in both the datasets demonstrates that modeling motivations of users by utilizing specific relations of the questions and it's contextual guided by linguistic literature will better help in identifying them. This also demonstrates the ability of the framework to model these motivations. The significant improvement in **RhetId (SC&SS)** in both the datasets shows that integrating motivations of rhetorical questions are useful for identifying them in social media. It also shows that the two concepts supplement each other for identifying rhetorical questions and the effectiveness of the framework for integrating these concepts. A paired t-test showed that the method improves significantly over the baselines in both the datasets.

In summary, the proposed framework significantly outperforms the baselines and



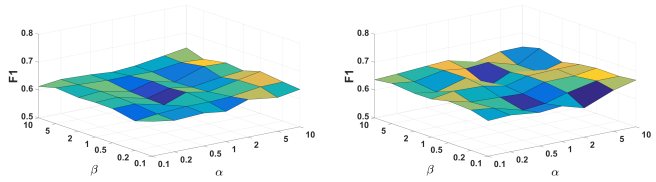
(a) AUC values for different values of α and β for Dataset 1

(b) AUC values for different values of α and β for Dataset 2



(c) Accuracy values for different values of α and β for Dataset 1

(d) Accuracy values for different values of α and β for Dataset 2



(e) F1 values for different values of α and β for Dataset 1

(f) F1 values for different values of α and β for Dataset 2

Figure 14: Performance of the framework for different values of $\alpha = \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$ and $\beta = \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$ for (a), (c), (e) Dataset 1 (b), (d), (f) Dataset 2. The framework is robust to the different values of the parameters.

is useful in identifying rhetorical questions in social media in both the datasets. The results show that modeling the motivations of the user to post rhetorical questions drawing concepts from linguistic theories is useful in identifying them. In the next section, we will examine the variation of the performance with different proportions of information modeling the motivations of implying the message and modifying the expressed sentiment.

6.6.3 Evaluation of Robustness across Parameter Values

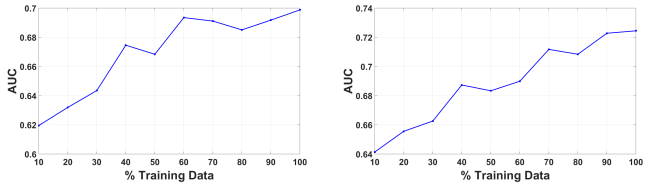
In the model presented in Eq. (6.8), α and β control the proportion of information modeling the motivations of implying the message and modifying the expressed sentiment respectively. To evaluate the framework for different proportions of two motivations, we set $\alpha = [0.1, 0.2, 0.5, 1, 2, 5, 10]$ and $\beta = [0.1, 0.2, 0.5, 1, 2, 5, 10]$ and plot the values for AUC, Accuracy and F1 measure in Figure 14 for both the datasets. We make the following observations from the figures.

From Figure 14, the general trend is that performance of the framework is maintained across different values of the parameters in both the datasets. This shows that the performance is robust to various proportions of information modeling the two motivations, indicating that effectiveness of the framework for integrating them. The performance is higher when both α β have non-zero values in both the datasets indicating that both motivations are important in identifying rhetorical questions. The framework outperforms baselines for all combination of the parameters in both the datasets demonstrating the effectiveness of the framework in modeling the motivations.

In summary, the framework performs well over different proportions of information modeling topic similarity and sentiment shift and is robust to their variation. An appropriate combination of information from the two motivations can optimize the effectiveness of the framework for identifying rhetorical questions in social media.

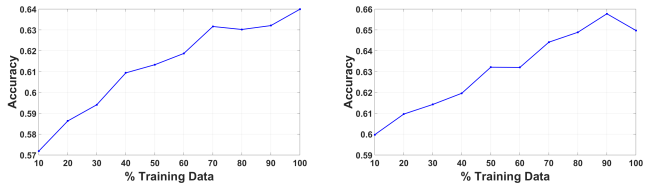
6.6.4 Identification with Less Training Data

We now evaluate the variation of performance of the framework with different proportions of training data for both the datasets. This experiment enables us to assess



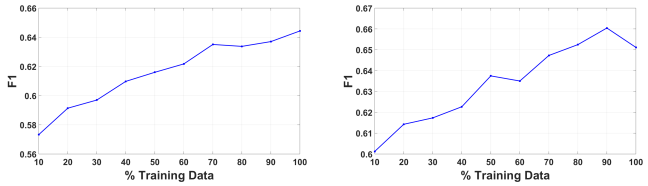
(a) AUC values for different training data size for Dataset 1

(b) AUC values for different training data size for Dataset 2



(c) Accuracy values for different training data size for Dataset 1

(d) Accuracy values for different training data size for Dataset 2



(e) F1 values for different training data size for Dataset 1

(f) F1 values for different training data size for Dataset 2

Figure 15: Performance of the framework for proportions of training data for (a), (c), (e) Dataset 1 (b), (d), (f) Dataset 2. The algorithm performs well for sufficiently low proportions of training data.

the amount of training information required for the framework and the adaptability of the framework when less amount of training information is available. This experiment will also help assess the robustness of the framework for varying proportions of training data. We use the training and testing data from the previous experiments. We train the model with different percentages of the training data set from 10% to 90% in steps of 10% and measure the performance of the framework keeping the testing data constant. We illustrate the results in Fig 15 and make the following observations.

From Fig 15 we can say that more training data is beneficial for increasing the performance of the framework in both the datasets. The framework outperforms the nearest baseline (Li *et al.*, 2011), for relatively low proportions for training data (50%-60%), demonstrating that it performs well for sufficiently small training data sizes. It is usually difficult to obtain the labels for rhetorical questions in social media, and this shows that the framework can apply when the labels are few. The performance shows consistent trends in all three metrics, showing that the framework efficiently utilizes the training data points across two datasets to identify rhetorical questions.

In summary, the results demonstrate that the framework can learn from a small amount of training data in both the datasets, and it efficiently utilizes training data to identify rhetorical questions in social media. The framework consistently performs well across all proportions of training data and hence is robust to its variations.

6.7 Further Applications of Identifying Rhetorical Questions

We now discuss the possible implications of identifying rhetorical questions in social media data. We first focus on applications related to improving information seeking systems and then discuss applications in studying social media campaigns.

One implication of our work is in developing better information seeking systems in social media. People post questions in their status messages to request personal information that is better obtained from their social circles (Morris *et al.*, 2010). Social media platforms provide timely information and hence is also used by people in the search for time-critical information during natural disasters (Ranganath *et al.*, 2015b). Information seeking systems can be misled into identifying rhetorical questions as genuine and try to find responses to them, leading to wastage of resources. Rhetorical questions are categorized as conversational questions, and identifying responders to them can initiate conversations and provide social support to users. One possible future direction can concentrate identifying appropriate responders to rhetorical questions in social media.

Social media is used in political campaigns owing to its broad reach and easy access. Examples can be advocacy groups in election campaigns or attempts of radicalization by groups like the ISIS (Guo and Saxton, 2013). Advocates of social media campaigns adopt nuanced strategies of message construction to shape user opinion (Ranganath *et al.*, 2016). Rhetorical questions which can be disguised as a question, but employed by users to state their views in a nuanced manner, can be useful for social media campaigners. Rhetorical questions are important means of persuasion (Petty *et al.*, 1981), and algorithms for identifying them can play a crucial part in monitoring the behavior of social media activists.

6.8 Summary

We develop a framework to identify rhetorical questions by modeling the possible motivations of the user for posting them. We focus on two motivations of the user, to imply a message and strengthen or mitigate the degree of a statement he previously made. We evaluate the framework on questions posted in Twitter and demonstrate its effectiveness in identifying rhetorical questions in social media.

Chapter 7

CONCLUSION AND FUTURE WORK

This chapter concludes the dissertation by summarizing the contributions of the work and exploring possible future directions

7.1 Conclusion

Online social media provides a new platform for people seeking information, as it enables them to reach out to a large number of people. However social media is not designed for information seeking, leading to users not getting adequate responses for their information needs. Designing algorithmic frameworks to facilitate information seeking in social media can help users fulfill their resource needs and social media platforms to increase their user satisfaction.

Facilitating information seeking in social media can give rise to several challenges. Information needs to be expressed in social media are personal to the asker and time-critical requiring prompt and satisfactory responses. The presence of people who try to shape seekers perspective on political issues can hinder them from getting balanced perspectives. Users can express viewpoints in the form of rhetorical questions, and this can mislead resource seeking systems. To address these challenges, we design an algorithmic framework to facilitate resource seeking in social media which makes the following contributions; identifying suitable responders for personal, time-critical requests; identifying users who push their agenda to shape user perspectives on a

given campaign; filtering out questions which do not seek resources to enable systems to better focus on genuine requests.

We draw from social foci theory to postulate that users who share context with the asker in the question domain are suitable to answer personal questions. We develop an algorithm to rank candidate responders according to their shared question-related context with the asker and evaluate it on personal questions posted on Twitter. We demonstrate its effectiveness in identifying responders to a wide range of question categories, with shared context measured by integrating network and content information.

Social media is a real-time platform and is hence used to seek information where promptness of reply is of the essence. We propose criteria to estimate the time taken for a user to respond to a question to identify responders who can provide timely and relevant answers. Our algorithm integrates information related to the responder's future availability, past response behavior, and interests. We evaluate the algorithm on the questions posted on Twitter during two natural disasters to demonstrate its effectiveness in identifying responders to time-critical questions.

An important component of information seeking behavior is present to identify advocates for political campaigns on social media. I characterize advocates through their message strategies, propagation strategies, and community structure and propose different characterizations based on them. I integrate heterogeneous information derived from these diverse characterizations and demonstrate that the characterizations can identify advocates effectively for political campaigns on social media. I also analyze contributions of individual characteristic groups like message strategies, propagation strategies, and community structure in identifying advocates.

Information seeking in social media goes hand in hand with users employing it

to express their viewpoints through rhetorical questions. It is difficult to identify rhetorical questions due to their similarity in syntactical form with other questions, and therefore we propose an algorithm that modeling the possible motivations of the user for posting them. We focus on two motivations of the user, to imply a message and strengthen or mitigate the degree of a statement he previously made. We demonstrate the effectiveness of the framework on questions posted on Twitter.

Through this systematic study on facilitating information seeking in social media, I focus mainly on two aspects: proposing concepts which can help in identifying responders who are most likely to satisfy personal, time-critical and financial requests in online social media; and introducing notions capable of differentiating between between genuine resource requests and rhetorical requests , which in turn, can increase the efficiency of systems facilitating resource seeking.

7.2 Future Work

This work can be extended in the following future directions

7.2.1 Resource Seeking in Social Media Campaigns

Resource seeking plays a vital role in conducting campaigns conducted for crisis response, elections and social issues. During emergencies, users have been shown to propagate time-critical requests to their followers to increase its reach. Identifying responders in such situations will require an understanding of the interplay between information propagation and information-seeking behavior of social media users. Identifying users who are providing a false response to questions will help to increase

the effectiveness of social media as a quality information source during campaigns. Analyzing different kinds of requests made by actors in a campaign can give insight into their type of informational needs and hidden ambitions. Rhetorical questions are a powerful tool and can be utilized by political and commercial entities to influence the opinion of the users. Studying the effects of rhetorical questions on different sets of users by social media campaigns will throw interesting insights into the ways in which various social media users respond to persuasive techniques.

7.2.2 Facilitating Financial Requests

Facilitating financial requests has several challenges and provides ample opportunities for future research. Estimating social influence of contribution amounts by taking network effects into consideration will provide insight into how contributors are affected by the donation behavior of their neighbors. Incorporating features like teams, contributor and request profile information to estimate amounts can further improve the performance of the framework in identifying contributors for a given request.The financial resources with each contributor are limited and lead to competition among the requesters. Modeling the effect of competition among requesters for identifying contributors will bring new insights on how people compete in a resource-constrained environment. Moving beyond the problem of responder identification, identifying borrowers who are prone to default and predicting them out at an early stage will help in increasing the satisfaction of the contributors and increase their faith in the platform.

7.2.3 Analyzing Conversations in Social Media

Another area we can extend the work is to facilitate conversations in social media. Conversations can be viewed as extensions of requests, where a large number of people discuss together on a task that they need to fulfill or a topic in which they have a different opinion. Identifying respondents who can help in different stages of conversations can help in taking them forward and increase the rate of satisfactory completion. Literature in dialog modeling is focused on formulating the next reply to a two-person conversation with chatbots. Extending this to formulate responses to multi-person conversations will help in bridging concepts of natural language conversations to resource seeking in social media. Analyzing response to these rhetorical questions and understanding the conversation dynamics driven by them is an interesting direction of future work. Designing algorithms to identify potential responders to rhetorical questions can help to increase user engagement in social media conversations.

REFERENCES

Acar, E., D. M. Dunlavy and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions", J.Chemo (2011).

Adamic, L. A., J. Zhang, E. Bakshy and M. S. Ackerman, "Knowledge sharing and yahoo answers: everyone knows something", in "WWW", (2008).

Anderson, A., D. Huttenlocher, J. Kleinberg and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow", in "KDD", pp. 850–858 (ACM, 2012).

Anzilotti, G. I., "The rhetorical question as an indirect speech device in english and italian.", CMLR (1982).

Baruah, A., "Namo Brigade", http://goo.gl/JKdmlP, [Online; accessed 7-May-2015] (2014).

Bhattacharya, P., S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly and K. P. Gummadi, "Deep twitter diving: exploring topical groups in microblogs at scale", in "CSCW", (ACM, 2014).

Bhattasali, S., J. Cytryn, E. Feldman and J. Park, "Automatic identification of rhetorical questions", in "ACL", (2015).

Bian, J., Y. Liu, E. Agichtein and H. Zha, "Finding the right facts in the crowd: Factoid question answering over social media", in "Proceedings of the 17th International Conference on World Wide Web", WWW '08, pp. 467–476 (ACM, New York, NY, USA, 2008), URL http://doi.acm.org/10.1145/1367497.1367561.

Blankenship, K. L. and T. Y. Craig, "Rhetorical question use and resistance to persuasion: An attitude strength analysis", Journal of language and social psychology **25**, 2, 111–128 (2006).

Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", JMLR (2003).

Boyd, S. and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).

Bozzon, A., M. Brambilla, S. Ceri, M. Silvestri and G. Vesci, "Choosing the right crowd: expert finding in social networks", in "EDBT", (ACM, 2013).

Brower, J. E., J. H. Zar and C. N. Von Ende, *Field and laboratory methods for general ecology* (WCB McGraw-Hill Boston, Massachusetts, 1998).

Burt, R. S., *Structural holes: The social structure of competition* (Harvard university press, 2009).

Case, D. O., *Looking for information: A survey of research on information seeking, needs and behavior* (Emerald Group Publishing, 2012).

Chan, W., W. Yang, J. Tang, J. Du, X. Zhou and W. Wang, "Community question topic categorization via hierarchical kernelized classification", in "CIKM", (ACM, 2013).

Chua, A. Y. and S. Banerjee, "So fast so good: An analysis of answer quality and answer speed in community question-answering sites", JASIST (2013).

Cialdini, R. B., *Influence: The psychology of persuasion* (Quill New York, NY, 1993).

Cong, G., L. Wang, C.-Y. Lin, Y.-I. Song and Y. Sun, "Finding question-answer pairs from online forums", in "SIGIR", (ACM, 2008).

Ding, C., T. Li and M. I. Jordan, "Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding", in "ICDM", (2008).

Ding, C., T. Li, W. Peng and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering", in "KDD", (ACM, 2006).

Du, L., W. Buntine and H. Jin, "A segmented topic model based on the two-parameter poisson-dirichlet process", JMLR (2010).

Dunlavy, D. M., T. G. Kolda and E. Acar, "Temporal link prediction using matrix and tensor factorizations", TKDD (2011).

Ellison, N. B., R. Gray, J. Vitak, C. Lampe and A. T. Fiore, "Calling all facebook friends: Exploring requests for help on facebook.", in "International AAAI Conference on Web and Social Media", (2013).

Farrell, D. M. and P. Webb, *Political parties as campaign organizations* (Sociology Institute of Zurich, 2006).

Feld, S. L., "The focused organization of social ties", AJS (1981).

Frank, J., "You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation", Journal of Pragmatics (1990).

Gao, H., G. Barbier and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief", IEEE Intelligent Systems (2011).

Garcia-Molina, H., G. Koutrika and A. Parameswaran, "Information seeking: convergence of search, recommendations, and advertising", CACM (2011).

Gass, R. H. and J. S. Seiter, *Persuasion: Social influence and compliance gaining* (Routledge, 2015).

Gilbert, I. and T. Henry, "Persuasion detection in conversation", Tech. rep., DTIC Document (2010).

Godin, F., V. Slavkovikj, W. De Neve, B. Schrauwen and R. Van de Walle, "Using topic models for twitter hashtag recommendation", in "WWW", (ACM, 2013).

González-Bailón, S., J. Borge-Holthoefer and Y. Moreno, "Broadcasters and hidden influentials in online protest diffusion", American Behavioral Scientist p. 0002764213479371 (2013).

Gray, R., N. B. Ellison, J. Vitak and C. Lampe, "Who wants to know?: question-asking and answering practices among facebook users", in "CSCW", (ACM, 2013).

Guo, C. and G. D. Saxton, "Tweeting social change: How social media are changing nonprofit advocacy", NVS (2013).

Harper, F. M., D. Moy and J. A. Konstan, "Facts or friends?: distinguishing informational and conversational questions in social q&a sites", in "CHI", (2009).

Harper, F. M., J. Weinberg, J. Logie and J. A. Konstan, "Question types in social q&a sites", First Monday (2010).

Hasanain, M., T. Elsayed and W. Magdy, "Identification of answer-seeking questions in arabic microblogs", in "CIKM", (ACM, 2014).

Hecht, B., J. Teevan, M. Morris and D. Liebling, "Searchbuddies: Bringing search engines into the conversation", in "International AAAI Conference on Web and Social Media", (2012).

Horowitz, D. and S. D. Kamvar, "The anatomy of a large-scale social search engine", in "WWW", (2010).

Hsiao, K.-J., A. Kulesza and A. Hero, "Social collaborative retrieval", in "WSDM", (ACM, 2014).

Hu, X., J. Tang, H. Gao and H. Liu, "Unsupervised sentiment analysis with emotional signals", in "WWW", (ACM, 2013).

Ilie, C., *What else can I tell you?: A pragmatic study of English rhetorical questions as discursive and argumentative acts* (Almqvist & Wiksell International, 1994).

Jeong, J.-W., M. R. Morris, J. Teevan and D. Liebling, "A crowd-powered socially embedded search engine", in "Seventh International AAAI Conference on Weblogs and Social Media", (2013).

Jernigan, D. H. and P. A. Wright, "Media advocacy: Lessons from community experiences", JPHP (1996).

Jurczyk, P. and E. Agichtein, "Discovering authorities in question answer communities by using link analysis", in "CIKM", (ACM, 2007).

Khullar, A. and A. Haridasan, "Politicians slug it out in India's first social media election", http://goo.gl/ZXZm3W, [Online; accessed 9-November-2014] (2014).

Kim, D., Y. Jo, I.-C. Moon and A. Oh, "Analysis of twitter lists as a potential source for discovering latent characteristics of users", in "CHI Workshop on Microblogging", (2010).

Kim, J. and J.-H. Kang, "Towards identifying unresolved discussions in student online forums", Applied intelligence (2014).

Kleinberg, J. M., "Authoritative sources in a hyperlinked environment", JACM (1999).

Korman, A. K., "Toward an hypothesis of work behavior.", JAP (1970).

Kumar, S., G. Barbier, M. A. Abbasi and H. Liu, "Tweettracker: An analysis tool for humanitarian and disaster relief", in "ICWSM", (2011).

Kumar, S., F. Morstatter and H. Liu, *Twitter Data Analytics* (Springer, 2013a).

Kumar, S., F. Morstatter and H. Liu, *Twitter data analytics* (Springer, 2014).

Kumar, S., F. Morstatter, R. Zafarani and H. Liu, "Whom should i follow?: identifying relevant users during crises", in "Hypertext", (ACM, 2013b).

Kywe, S. M., T.-A. Hoang, E.-P. Lim and F. Zhu, "On recommending hashtags in twitter networks", in "Social Informatics", (Springer, 2012).

Lampe, C., R. Gray, A. T. Fiore and N. Ellison, "Help is on the way: Patterns of responses to resource requests on facebook", in "Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing", pp. 3–15 (ACM, 2014).

Lee, D. D. and H. S. Seung, "Algorithms for non-negative matrix factorization", in "NIPS", (2000).

Lee, K., P. Tamilarasan and J. Caverlee, "Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media.", in "ICWSM", (2013).

Lee, U., H. Kang, E. Yi, M. Yi and J. Kantola, "Understanding mobile q&a usage: An exploratory study", in "Proceedings of the SIGCHI Conference on Human Factors in Computing Systems", pp. 3215–3224 (ACM, 2012).

Li, B., I. King and M. R. Lyu, "Question routing in community question answering: putting category in its place", in "CIKM", (ACM, 2011).

Liang, H., Y. Xu, D. Tjondronegoro and P. Christen, "Time-aware topic recommendation based on micro-blogs", in "CIKM", (ACM, 2012).

Lin, Y.-R., J. Sun, H. Sundaram, A. Kelliher, P. Castro and R. Konuru, "Community discovery via metagraph factorization", TKDD (2011).

Liu, Q. and E. Agichtein, "Modeling answerer behavior in collaborative question answering systems", in "AIR", (Springer, 2011).

Liu, Z. and B. J. Jansen, "Factors influencing the response rate in social question and answering behavior", in "CSCW", (ACM, 2013).

Lulla, A., "Achtung! It's the NaMo Brigade!", http://goo.gl/XeHEzf, [Online; accessed 14-December-2015] (2014).

Lumezanu, C., N. Feamster and H. Klein, "# bias: Measuring the tweeting behavior of propagandists", in "ICWSM", (2012).

Ma, Z., A. Sun, Q. Yuan and G. Cong, "Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter", in "CIKM", (ACM, 2014).

Mahmud, J., J. Chen and J. Nichols, "When will you answer this? estimating response time in twitter", in "ICWSM", (2013).

Mahmud, J., M. Zhou, N. Megiddo, J. Nichols and C. Drews, "Optimizing the selection of strangers to answer questions in social media", arXiv preprint arXiv:1404.2013 (2014).

Marsden, P. V., "Homogeneity in confiding relations", Social networks (1988).

McCarthy, J. D. and M. N. Zald, "Resource mobilization and social movements: A partial theory", AJS (1977).

Mishra, N., R. W. White, S. Ieong and E. Horvitz, "Time-critical search", in "Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval", pp. 747–756 (ACM, 2014a).

Mishra, N., R. W. White, S. Ieong and E. Horvitz, "Time-critical search", in "SIGIR", (ACM, 2014b).

Morris, M. R., J. Teevan and K. Panovich, "What do people ask their social networks, and why?: a survey study of status message q&a behavior", in "Proceedings of the SIGCHI conference on Human factors in computing systems", pp. 1739–1748 (ACM, 2010).

Nandi, A., S. Paparizos, J. C. Shafer and R. Agrawal, "With a little help from my friends", in "ICDE", (2013).

Pal, A., S. Chang and J. A. Konstan, "Evolution of experts in question answering communities.", in "ICWSM", (2012).

Pal, A. and S. Counts, "Identifying topical authorities in microblogs", in "WSDM", (ACM, 2011).

Pal, A., A. Herdagdelen, S. Chatterji, S. Taank and D. Chakrabarti, "Discovery of topical authorities in instagram", in "WWW", (ACM, 2016).

Palen, L., S. Vieweg and K. M. Anderson, "Supporting "everyday analysts" in safety- and time-critical situations", The Information Society (2010).

Palmer, A., "NRA campaign in high gear", http://goo.gl/PYBXDq, [Online; accessed 9-November-2014] (2014).

Panovich, K., R. Miller and D. Karger, "Tie strength in question & answer on social network sites", in "CSCW", (2012).

Paul, S. A., L. Hong and E. H. Chi, "Is twitter a good place for asking questions? a characterization study", in "International AAAI Conference on Web and Social Media", (2011).

Pennebaker, J. W., M. E. Francis and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001", Mahway: Lawrence Erlbaum Associates (2001).

Petty, R. E., J. T. Cacioppo and M. Heesacker, "Effects of rhetorical questions on persuasion: A cognitive response analysis.", JPSP (1981).

Philipsen, G., G. Philipsen and T. Albrecht, "A theory of speech codes", Dev.CommTheo **6** (1997).

Podgorny, I. A., M. Cannon, C. Gielow and T. Goodyear, "Real time detection and intervention of poorly phrased questions", in "CHI", (ACM, 2015).

Purohit, H., C. Castillo, F. Diaz, A. Sheth and P. Meier, "Emergency-relief coordination on social media: Automatically matching resource requests and offers", First Monday **19**, 1 (2013).

Qu, B., G. Cong, C. Li, A. Sun and H. Chen, "An evaluation of classification models for question topic categorization", JASIST (2012).

Qu, Y., C. Huang, P. Zhang and J. Zhang, "Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake", in "CSCW", (ACM, 2011).

Radev, D. R., H. Qi, H. Wu and W. Fan, "Evaluating web-based question answering systems", LREC (2002).

Ranganath, S., X. Hu, J. Tang and H. Liu, "Understanding and identifying advocates for political campaigns on social media", in "WSDM", (ACM, 2016).

Ranganath, S., J. Tang, X. Hu, H. Sundaram and H. Liu, "Leveraging social foci for information seeking in social media", in "AAAI", (2015a).

Ranganath, S., S. Wang, X. Hu, J. Tang and H. Liu, "Finding time-critical responses for information seeking in social media", in "ICDM", (2015b).

Riahi, F., Z. Zolaktaf, M. Shafiei and E. Milios, "Finding expert users in community question answering", in "WWW", (ACM, 2012).

RiteTag, "Twitter hashtags", http://goo.gl/X88aY7, [Online; accessed 7-August-2014] (2014).

Robinson, M., "NRA, PRO-GUN, GUN OWNER", http://goo.gl/OyDDGR, [Online; accessed 7-August-2014] (2014).

Schmidt-Radefeldt, J., "On so-called rhetorical questions", Journal of Pragmatics (1977).

Schuth, A., K. Hofmann, S. Whiteson and M. de Rijke, "Lerot: An online learning to rank framework", in "Proceedings of the 2013 workshop on Living labs for information retrieval evaluation", pp. 23–26 (ACM, 2013).

Seung, H. and D. Lee, "Algorithms for non-negative matrix factorization", NIPS (2001).

Starbird, K. and L. Palen, "Voluntweeters: Self-organizing by digital volunteers in times of crisis", in "CHI", (2011).

Starbird, K. and L. Palen, "Working and sustaining the virtual disaster desk", in "CSCW", (ACM, 2013).

Tang, L. and H. Liu, "Relational learning via latent social dimensions", in "KDD", (2009).

Teevan, J., M. R. Morris and K. Panovich, "Factors affecting response quantity, quality, and speed for questions asked via social network status messages.", in "ICWSM", (2011).

Today, I., "Rise of the Cyber Hindu", http://goo.gl/NvxAw6, [Online; accessed 7-May-2015] (2014).

Usunier, N., D. Buffoni and P. Gallinari, "Ranking with ordered weighted pairwise classification", in "ICML", (ACM, 2009).

Wang, G., K. Gill, M. Mohanlal, H. Zheng and B. Y. Zhao, "Wisdom in the social crowd: an analysis of quora", in "WWW", (2013a).

Wang, Y., L. Wang, Y. Li, D. He, T.-Y. Liu and W. Chen, "A theoretical analysis of ndcg type ranking measures", arXiv preprint arXiv:1304.6480 (2013b).

Wen, X. and Y.-R. Lin, "Information seeking and responding networks in physical gatherings: A case study of academic conferences in twitter", in "COSN", (ACM, 2015).

Weng, J., E.-P. Lim, J. Jiang and Q. He, "Twitterrank: finding topic-sensitive influential twitterers", in "WSDM", (ACM, 2010).

Weston, J., S. Bengio and N. Usunier, "Large scale image annotation: learning to rank with joint word-image embeddings", JMLR (2010).

Weston, J., C. Wang, R. Weiss and A. Berenzweig, "Latent collaborative retrieval", in "ICML", (2012).

Yang, J., J. McAuley and J. Leskovec, "Community detection in networks with node attributes", ICDM (2013a).

Yang, J., M. R. Morris, J. Teevan, L. A. Adamic and M. S. Ackerman, "Culture matters: A survey study of social q&a behavior.", International AAAI Conference on Web and Social Media (2011).

Yang, L., M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun and Z. Chen, "Cqarank: jointly model topics and expertise in community question answering", in "CIKM", (ACM, 2013b).

Yang, L., T. Sun, M. Zhang and Q. Mei, "We know what@ you# tag: does the dual role affect hashtag adoption?", in "WWW", (ACM, 2012).

Zangerle, E., W. Gassler and G. Specht, "Recommending#-tags in twitter", in "SASWeb", vol. 730 (2011).

Zhang, G. P. and M. Qi, "Neural network forecasting for seasonal and trend time series", EJOR (2005).

Zhao, Z. and Q. Mei, "Questions about questions: An empirical analysis of information needs on twitter", in "WWW", (ACM, 2013).

Zhou, G., S. Lai, K. Liu and J. Zhao, "Topic-sensitive probabilistic model for expert finding in question answer communities", in "CIKM", (2012).

Zhu, H., E. Chen, H. Xiong, H. Cao and J. Tian, "Ranking user authority with relevant knowledge categories for expert finding", WWW (2013).