Microblog Conversation Recommendation via Joint Modeling of Topics and Discourse

Xingshan Zeng[1,3], **Jing Li**[2], **Lu Wang**[4],
Nicholas Beauchamp[5,6], **Sarah Shugars**[6], **Kam-Fai Wong**[1,3]
[1]The Chinese University of Hong Kong, [2]Tencent AI Lab
[3]MoE Key Laboratory of High Confidence Software Technologies, China
[4]College of Computer and Information Science, Northeastern University, US
[5]Department of Political Science, Northeastern University, US
[6]Network Science Institute, Northeastern University, US
[1,3]{xszeng,kfwong}@se.cuhk.edu.hk
[2]ameliajli@tencent.com, [4]luwang@ccs.neu.edu
[5]n.beauchamp@northeastern.edu, [6]shugars.s@husky.neu.edu

Abstract

Millions of conversations are generated every day on social media platforms. With limited attention, it is challenging for users to select which discussions they would like to participate in. Here we propose a new method for microblog conversation recommendation. While much prior work has focused on post-level recommendation, we exploit both the conversational context, and user content and behavior preferences. We propose a statistical model that jointly captures: (1) *topics* for representing user interests and conversation content, and (2) *discourse modes* for describing user replying behavior and conversation dynamics. Experimental results on two Twitter datasets demonstrate that our system outperforms methods that only model content without considering discourse.

1 Introduction

Online platforms have revolutionized the way individuals collect and share information (O'Connor et al., 2010; Lee and Ma, 2012; Bakshy et al., 2015), but the vast bulk of online content is irrelevant or unpalatable to any given individual. A user interested in political discussion, for instance, might prefer content concerning a specific candidate or issue, and only then if discussed in a positive light without controversy (Adamic and Glance, 2005; Bakshy et al., 2015).

How do individuals facing such large quantities of superfluous material select which conversations to engage in, and how might we better algorithmically recommend conversations suited to individual users? We approach this problem from a *microblog conversation recommendation* framework. Where prior work has focused on the content of individual posts for recommendation (Chen et al., 2012; Yan et al., 2012; Vosecky et al., 2014; He and Tan, 2015), we examine the entire history and context of a conversation, including both topical content and *discourse modes* such as agreement, question-asking, argument and other dialogue acts (Ritter et al., 2010).[1] And where Backstrom et al. (2013) leveraged conversation reply structure (such as previous user engagement), their model is unable to predict first entry into new conversations, while ours is able to predict both new and repeated entry into conversations based on a

Conversation 1
...
[$\mathbf{U}_1$]: The sheer cognitive dissonance required for a "liberal" to say Clinton is as bad as Trump is just staggering.
[U_2]: Hillarists, Troll; they insult Liberals trying to distract from Hillary's Conseratism.
[U_3]: I still prefer Hillarist b/c it describes their Cultish and ideological aspects.
———————
...

Conversation 2
...
[U_4]: I do not like trump at all, but Comey left her in place knowing Bernie is much stronger.
[$\mathbf{U}_1$]: If you're going to actively start rooting against the Democrats, get off my mentions. I have enough GOP doing that.
[U_5]: Your tweets are an example of why open primaries are stupid. You're not a Dem, you're just for one guy.
———————
[$\mathbf{U}_1$]: No offense, but you've been wrong about pretty much everything so far. Why would I trust your prognostication now?
...

Figure 1: Two snippets of conversations on Twitter. [U_i]: The message is posted by user U_i. "—" is the dividing line between training history and test part. U_1 did not reengage in Conversation 1 but reengaged in Conversation 2.

[1]In this paper, *discourse mode* refers to a certain type of dialogue act, e.g., agreement or argument. The *discourse structure* of a conversation means some combination (or a probability distribution) of discourse modes.

combination of topical and discourse features.

To illustrate the interplay between topics and discourse, Figure 1 displays two snippets of conversations on Twitter collected during the 2016 United States presidential election. User U_1 participates in both conversations. The first conversation is centered around Clinton, and U_1, who is more typically involved with conversations about candidate Sanders, does not return. In the second conversation, however, U_1 is involved in a heated back-and-forth debate, and thus is drawn back to a conversation that they may otherwise have abandoned but for their enjoyment of adversarial discourse.

Effective conversation prediction and recommendation requires an understanding of both user interests and discourse behaviors, such as agreement, disagreement, inquiry, backchanneling, and emotional reactions. However, acquiring manual labels for both is a time-consuming process and hard to scale for new datasets. We instead propose a unified statistical learning framework for conversation recommendation, which jointly learns (1) hidden factors that reflect *user interests* based on conversation history, and (2) *topics* and *discourse modes* in ongoing conversations, as discovered by a novel probabilistic latent variable model. Our model is built on the success of collaborative filtering (CF) in recommendation systems, where latent dimensions of product ratings or movie reviews are extracted to better capture user preferences (Linden et al., 2003; Salakhutdinov and Mnih, 2008; Wang and Blei, 2011; McAuley and Leskovec, 2013). To the best of our knowledge, we are the first to model *both* topics and discourse modes as part of a CF framework and apply it to microblog conversation recommendation.[2]

Experimental results on two Twitter conversation datasets show that our proposed model yields significantly better performance than state-of-the-art post-level recommendation systems. For example, by leveraging both topical content and discourse structure, our model achieves a mean average precision (MAP) of 0.76 on conversations about the U.S. presidential election, compared with 0.70 by McAuley and Leskovec (2013), which only considers topics. We further conducted detailed analysis on the latent topics and discourse modes and find that our model can discover reasonable topic and discourse representations, which play an important role in characterizing reply behaviors. Finally, we also provide a pilot study on recommendation for first time replies, which shows that our model outperforms comparable recommendation systems.

The rest of this paper is structured as follows. The related work is discussed in Section 2. We then present our microblog conversation recommendation model in Section 3. The experimental setup and results are described in Sections 4 and 5. Finally, we conclude in Section 6.

2 Related Work

Social media has attracted increasing attention in digital communication research (Agichtein et al., 2008; Kwak et al., 2010; Wu et al., 2011). The problem studied here is closely related to work on recommendation and response prediction in microblogs (Artzi et al., 2012; Hong et al., 2013), where the goal is to predict whether a user will share or reply to a given post. Existing methods focus on measuring features that reflect personalized user interests, including topics (Hong et al., 2013) and network structures (Pan et al., 2013; He and Tan, 2015). These features have been investigated under a learning to rank framework (Duan et al., 2010; Artzi et al., 2012), graph ranking models (Yan et al., 2012; Feng and Wang, 2013; Alawad et al., 2016), and neural network-based representation learning methods (Yu et al., 2016).

Distinguishing from prior work that focuses on post-level recommendation, we tackle the challenges of predicting user reply behaviors at the conversation-level. In addition, our model not only captures latent factors such as the topical interests of users, but also leverages the automatically learned discourse structure. Much of the previous work on discourse structure and dialogue acts has relied on labeled data (Jurafsky et al., 1997; Stolcke et al., 2000), while unsupervised approaches have not been applied to the problem of conversation recommendation (Woszczyna and Waibel, 1994; Crook et al., 2009; Ritter et al., 2010; Joty et al., 2011).

Our work is also in line with conversation modeling for social media discussions (Ritter et al., 2010; Budak and Agrawal, 2013; Louis and Cohen, 2015; Cheng et al., 2017). Topic modeling has been employed to identify conversation con-

[2] To ensure the general applicability of our approach to domains lacking such information, we do not utilize external features such as network structure, but it may certainly be added in future, more narrowly targeted applications.

tent on Twitter (Ritter et al., 2010). In this work, we propose a probabilistic model to capture both topics and discourse modes as latent variables. A further line of work studies the reposting and reply structure of conversations (Gómez et al., 2011; Laniado et al., 2011; Backstrom et al., 2013; Budak and Agrawal, 2013). But none of this work distinguishes the rich discourse functions of replies, which is modeled and exploited in our work.

3 The Joint Model of Topic and Discourse for Recommendation

Our proposed microblog conversation recommendation framework is based on collaborative filtering and a novel probabilistic graphical model. Concretely, our objective function takes the form:

$$\min \mathcal{L} + \mu \cdot NLL(\mathcal{C} \mid \mathbf{\Theta}) \tag{1}$$

This function encodes two types of information. First, $\mathcal{L}$ models user reply preference in a similar fashion to collaborative filtering (CF) (Hu et al., 2008; Pan et al., 2008). It captures topics of interests and discourse structures users are commonly involved (e.g., argumentation), and takes the form of mean square error (MSE) based on user reply history. This part is detailed in Section 3.1.

The second term, $NLL(\mathcal{C} \mid \mathbf{\Theta})$, denotes the negative log-likelihood of a set of conversations $\mathcal{C}$, with $\mathbf{\Theta}$ containing all parameters. A probabilistic model is described in Section 3.2 that shows how the topical content and discourse structures of conversations are captured by these latent variables.

The hyperparameter μ controls the trade-off between the two effects. ℓ_2 regularization is also added for parameters to avoid model overfitting.

For the rest of this section, we first present the construction of $\mathcal{L}$ and $NLL(\mathcal{C} \mid \mathbf{\Theta})$ in Sections 3.1 and 3.2. We then discuss how these two components can be mutually informed by each other in Section 3.3. Finally, the generative process and parameter learning are described in Section 3.4.

3.1 Reply Preference ($\mathcal{L}$)

Our user reply preference modeling is built on the success of collaborative filtering (CF) for product ratings. However, classic CF problems, such as product recommendation, generally rely on explicit user feedback. Unlike user ratings on products, our input lacks explicit feedback from users about negative preferences and non-response. Therefore, we follow one-class Collaborative Filtering (Hu et al., 2008; Pan et al., 2008), which weights positive instances higher during training and is thus suited to our data. Formally, for user u and conversation c, we measure reply preference based on the MSE between predicted preference score $p_{u,c}$ and reply history $r_{u,c}$. $r_{u,c}$ equals 1 if u is in the conversation history; otherwise, it is 0. The first term of objective (Eq. 1) takes the following form:

$$\mathcal{L} = \sum_{u=1}^{|\mathcal{U}|} \sum_{c=1}^{|\mathcal{C}|} f_{u,c} \cdot (p_{u,c} - r_{u,c})^2 \tag{2}$$

where $\mathcal{U}$ consists of users $\{u\}$ and $\mathcal{C}$ is a set of conversations $\{c\}$ in a dataset. $f_{u,c}$ is the corresponding weight for a conversation c and a target user u. Intuitively, it has a large value if positive feedback (user replied) is observed. Therefore, we adapt the formulation from Pan et al. (2008):

$$f_{u,c} = \begin{cases} s & \text{if } r_{u,c} = 1 \text{ (i.e., user replied)} \\ 1 & \text{if } r_{u,c} = 0 \end{cases} \tag{3}$$

where $s > 1$, an integer hyperparameter to be tuned.

Inspired by prior models (Koren et al., 2009; McAuley and Leskovec, 2013), we propose the following latent factor model to describe $p_{u,c}$:

$$p_{u,c} = \lambda \cdot \gamma_u^U \cdot \gamma_c^C + (1-\lambda) \cdot \delta_u^U \cdot \delta_c^C + b_u + b_c + a \tag{4}$$

γ_u^U and γ_c^C are K-dimensional latent vectors that encode topic-specific information (where K is the number of latent topics) for users and conversations. Specifically, γ_u^U reflects the topical interests of u, with higher value $\gamma_{u,k}^U$ indicating greater interest by u in topic k. γ_c^C captures the extents that topics are discussed in conversation c.

Similarly, D-dimensional vectors δ_u^U and δ_c^C capture discourse structures in shaping reply behaviors (where D is the number of discourse clusters). δ_u^U reflects the discourse behaviors u prefers, such as u_1 often enjoys arguments as in the second conversation of Figure 1, while δ_c^C captures the discourse modes used throughout conversation c. By multiplying user and conversation factors, we can measure the corresponding similarity. The predicted score $p_{u,c}$ thereby reflects the tendency for a user u to be involved in conversation c.

As pointed out by McAuley and Leskovec (2013), these latent vectors often encode hidden factors that are hard to interpret under a CF framework. Therefore, in Section 3.2, we present a novel probabilistic model which can extract interpretable topics and discourse modes as word distributions. We then describe how they can be aligned with the latent vectors of γ^C and δ^U.

Parameter a is an offset parameter, b_u and b_c are user and conversation biases, and $\lambda \in [0, 1]$ serves as the weight for trading offs of topic and discourse factors in reply preference modeling.

3.2 Corpus Likelihood $NLL(\mathcal{C} \mid \mathbf{\Theta})$

Here we present a novel probabilistic model that learns coherent word distributions for latent topics and discourse modes of conversations. Formally, we assume that each conversation $c \in \mathcal{C}$ contains M_c messages, and each message m has $N_{c,m}$ words. We distinguish three latent components – *discourse*, *topic*, and *background* – underlying conversations, each with their own type of word distribution. At the corpus level, there are K topics represented by word distribution ϕ_k^T ($k = 1, 2, ..., K$), while ϕ_d^D ($d = 1, 2, ..., D$) represents the D *discourse modes* embedded in corpus. In addition, we add a *background* word distribution ϕ^B to capture general information (e.g., common words), which do not indicate either discourse or topic information. ϕ_d^D, ϕ_k^T, and ϕ^B are all multinomial word distributions over vocabulary size V. Below describes more details.

Message-level Modeling. Our model assigns two types of message-level multinomial variables to each message: $z_{c,m}$ reflects its latent *topic* and $d_{c,m}$ represents its *discourse mode*.

Topic assignments. Due to the short nature of microblog posts, we assume each message m in conversation c contains only one topic, indexed as $z_{c,m}$. This strategy has been proven useful to alleviate data sparsity for topic inference (Quan et al., 2015). We further assume messages in the same conversation would focus on similar topics. We thus draw topic $z_{c,m} \sim \theta_c$, where θ_c denotes the fractions of topics discussed in conversation c.

Discourse assignments. To capture discourse behaviors of u, distribution π_u is used to represent the discourse modes in messages posted by u. The discourse mode $d_{c,m}$ for message m is then generated from $\pi_{u_{c,m}}$, where $u_{c,m}$ is the author of m in c.

Word-level Modeling. We aim to separate *discourse*, *topic*, and *background* information for conversations. Therefore, for each word $w_{c,m,n}$ of message m, a ternary switcher $x_{c,m,n} \in \{\text{DISC}, \text{TOPIC}, \text{BACK}\}$ controls word $w_{c,m,n}$ to fall into one of the three types: *discourse*, *topic*, and *background*.

Discourse words (DISC) are indicative of the discourse modes of messages. When $x_{c,m,n} = \text{DISC}$ (i.e., $w_{c,m,n}$ is assigned as a discourse word), word $w_{c,m,n}$ is generated from the discourse word distribution $\phi_{d_{c,m}}^D$ where $d_{c,m}$ is discourse assignment to message m.

Topic words (TOPIC) describe the topical focus of a conversation. When $x_{c,m,n} = \text{TOPIC}$, $w_{c,m,n}$ is assigned as a topic word and generated from $\phi_{z_{c,m}}^T$ – word distribution given topic of m.

Background words (BACK) capture the general information that is not related to discourse or topic. When word $w_{c,m,n}$ is assigned as a background word ($x_{c,m,n} = \text{BACK}$), it is drawn from background distribution ϕ^B.

Switching among Topic, Discourse, and Background. We further assume the word type switcher $x_{c,m,n}$ is sampled from a multinomial distribution which depends on the current discourse mode $d_{c,m}$. The intuition is that messages of different discourse modes may show different distributions of the three word types. For instance, a statement message may contain more content words than a rhetorical question. Specifically, $x_{c,m,n} \sim Multi(\tau_{d_{c,m}})$, where τ_d is a 3-dimension stochastic vector that expresses the appearing probabilities of three kinds of words (DISC, TOPIC, BACK), when the discourse assignment is d. Stop words and punctuations are forced to be labeled as discourse or background. By explicitly distinguishing different types of words with switcher $x_{c,m,n}$, we can thus separate word distributions that reflect discourse, topic, and background information.

Likelihood. Based on the message-level and the word-level generation process, the probability of observing words in the given corpus is:

$$\begin{aligned} & Pr(\mathcal{C} \mid \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\tau}, \mathbf{z}, \mathbf{d}, \mathbf{x}) \\ = & \prod_{c=1}^{C} \prod_{m=1}^{M_c} \theta_{c,z_{c,m}} \pi_{u_{c,m},d_{c,m}} \\ & \times \prod_{x_{c,m,n}=\text{BACK}} \tau_{d_{c,m},\text{BACK}} \phi_{w_{c,m,n}}^B \\ & \times \prod_{x_{c,m,n}=\text{DISC}} \tau_{d_{c,m},\text{DISC}} \phi_{d_{c,m},w_{c,m,n}}^D \\ & \times \prod_{x_{c,m,n}=\text{TOPIC}} \tau_{d_{c,m},\text{TOPIC}} \phi_{z_{c,m},w_{c,m,n}}^T \end{aligned} \tag{5}$$

And we use negative log likelihood to model corpus likelihood effect in Eq. 1, i.e., $NLL(\mathcal{C} \mid \mathbf{\Theta}) = -\log(Pr(\mathcal{C} \mid \mathbf{\Theta})$, where parameters set $\mathbf{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\tau}, \mathbf{z}, \mathbf{d}, \mathbf{x}\}$.

3.3 Mutually Informed User Preference and Latent Variables

As mentioned above, the hidden factors discovered in Section 3.1 lack interpretability, which can be boosted by the learned latent topics and discourse modes in Section 3.2. However, it is non-trivial to link the topic-related parameters of γ_c^C to the conversation topic distributions of θ_c, since the former takes real values from $-\infty$ to $+\infty$ while the latter is a stochastic vector. Therefore, we follow the strategy from McAuley and Leskovec (2013) to apply a softmax function over γ_c^C:

$$\theta_{c,k} = \frac{exp(\kappa^T \gamma_{c,k}^C)}{\sum_{k'=1}^{K} exp(\kappa^T \gamma_{c,k'}^C)} \tag{6}$$

We further assume that the discourse mode preference by users, δ_u^U, can also be informed by the discourse mode distribution captured by π_u, i.e., a user who enjoys arguments may be willing to participate another. So similarly, we define:

$$\pi_{u,d} = \frac{exp(\kappa^D \delta_{u,d}^U)}{\sum_{d'=1}^{D} exp(\kappa^D \delta_{u,d'}^U)} \tag{7}$$

where κ^T and κ^D are learnable parameters that control the "peakiness" of the transformation. For example, a larger κ^T indicates a more focused conversation, while a smaller κ^T means users discuss diverse topics.

Finally, softmax transformation is also applied to ϕ_k^T, ϕ_d^D, ϕ^B, and τ_d, as done in McAuley and Leskovec (2013), with additional parameters ψ_k^T, ψ_d^D, ψ^B, and χ_d (as shown in Figure 2). This is to ensure that the distributions ϕ_*^* and τ_d are stochastic vectors. In doing so, these distributions can be learned via optimizing ψ_*^* and χ_d, which take any value and thus ensure that the cost function in Eq. 1 is optimized without considering any parameter constraints.

3.4 Generative Process and Model Learning

Our word generation process is displayed in Figure 2 and described as follows:

- Compute topic distribution θ_c by Eq. 6
- For message $m = 1$ to M_c:
  - Compute discourse distribution $\pi_{u_{c,m}}$ by Eq. 7
  - Draw topic assignment $z_{c,m} \sim Multi(\theta_c)$
  - Draw discourse mode $d_{c,m} \sim Multi(\pi_{u_{c,m}})$
  - For word index $n = 1$ to $N_{c,m}$:
    - Draw word type $x_{c,m,n} \sim Multi(\tau_d)$
    - if $x_{c,m,n} == BACK$:
      Draw word $w_{c,m,n} \sim Multi(\phi^B)$
    - if $x_{c,m,n} == DISC$:
      Draw word $w_{c,m,n} \sim Multi(\phi^D_{d_{c,m}})$
    - if $x_{c,m,n} == TOPIC$:
      Draw word $w_{c,m,n} \sim Multi(\phi^T_{z_{c,m}})$

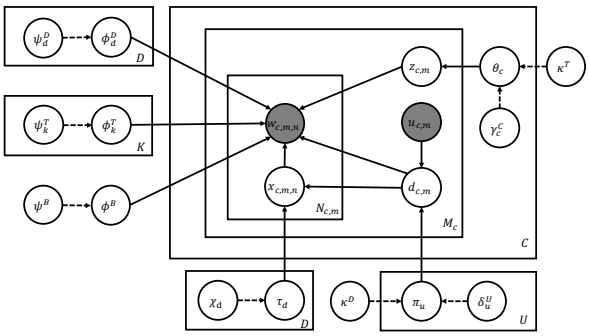


Figure 2: Generative process of our joint model of topic and discourse. u represents users. c represents conversations. Dotted arrows represent the softmax linkings, while solid arrows mean conditional priors.

Parameter Learning. For learning, we randomly initialize all learnable parameters and then alternate between the following two steps:

Step 1. Fix topic and discourse assignments z and d, and word type switcher x, then optimize the remaining parameters in Eq. 1 by L-BFGS (Nocedal, 1980):

$$\begin{aligned} &\text{Update } a, b, \gamma^*, \delta^*, \kappa^*, \psi^*, \chi = \\ &\quad argmin\, \mathcal{L} + \mu \cdot NLL(\mathcal{C} \mid \mathbf{\Theta}) \end{aligned} \tag{8}$$

Step 2. Sample topic and discourse assignments z and d at the message level and word type switcher x at the word level, using the distributions, computed according to parameters optimized in step 1:

$$\begin{aligned} &\text{Sample } z_{c,m},\ d_{c,m},\ x_{c,m,n} \text{ with probabilities} \\ &\quad p(z_{c,m} = k) = \theta_{c,k} \\ &\quad p(d_{c,m} = d) = \pi_{u_{c,m},d} \\ &\quad p(x_{c,m,n} = BACK) = \phi^B_{w_{c,m,n}} \tau_{d_{c,m},BACK} \\ &\quad p(x_{c,m,n} = DISC) = \phi^D_{d_{c,m},w_{c,m,n}} \tau_{d_{c,m},DISC} \\ &\quad p(x_{c,m,n} = TOPIC) = \phi^T_{z_{c,m},w_{c,m,n}} \tau_{d_{c,m},TOPIC} \end{aligned} \tag{9}$$

Step 2 is analogous to Gibbs Sampling (Griffiths, 2002) in probabilistic graphical models, such as LDA (Blei et al., 2003). However, distinguishing from previous models, the multinomial distributions in our models are not drawn from a Dirichlet prior. Instead, they are computed based on the parameters learned in Step 1.

Our learning process stops when the change of parameters is small (i.e., below a pre-specified threshold). Multiple restarts are tried, and similar results are achieved.

Dataset	# of user	# of conv	# of msg	Avg msg per user	Avg conv per user
US Election	4,300	2,013	22,092	5.14	1.23
TREC	10,122	7,500	38,999	3.85	1.71

Table 1: Statistics of two datasets.

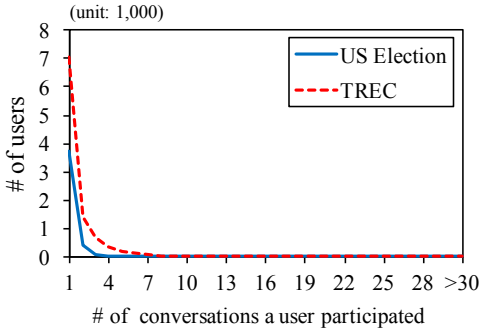


Figure 3: Horizontal axis: number of conversations that a user is involved. Vertical axis: number of users fall in the category (unit: 1,000). Notice that most of users (about 98%) participate in less than 10 conversations.

4 Experimental Setup

Datasets. We collected two microblog conversation datasets from Twitter for experiments[3]: one contains discussions about the U.S. presidential election (henceforth **US Election**), the other gathers conversations of diverse topics based on the tweets released by TREC 2011 microblog track (henceforth **TREC**)[4]. US Election was collected from January to June of 2016 using Twitter's Streaming API[5] with a small set of political keywords.[6] To recover conversations, Tweet Search API[7] was used to retrieve messages with the "in-reply-to" relations to collect tweets in a recursive way until full conversations were recovered.

Statistics of the datasets are shown in Table 1. Figure 3 displays the number of conversations individual users participated in. As can be seen, most users are involved in only a few conversations. Simply leveraging personal chat history will not produce good performance for conversation recommendation.

In our experiments, we predict whether a user will engage in a conversation given the previous messages in that conversation and past conversations the user is involved. For model training and testing, we divide conversations into three ordered segments, corresponding to training, development, and test sets at 75%, 12.5%, and 12.5%.[8]

Preprocessing and Hyperparameter Tuning. For preprocessing, links, mentions (i.e., @username), and hashtags in tweets were replaced with generic tags of "URL", "MENTION", and "HASHTAG". We then utilized the Twitter NLP tool[9] (Gimpel et al., 2011; Owoputi et al., 2013) for tokenization and non-alphabetic token removal. We removed stop words and punctuations for all comparisons to ensure comparable performance. We maintain a vocabulary with the 5,000 most frequent words.

Our model parameters are tuned on the development set based on grid search, i.e. the parameters that give the lowest value for our objective are selected. Specifically, the number of discourse modes (D) and topics (K) are tuned to be 10. The trade-off parameter μ between user preference and corpus negative log-likelihood takes value of 0.1, and λ, the parameter for balancing topic and discourse, is set to 0.5. Finally, the confidence parameter s takes a value of 200 to give higher weight for positive instances, i.e., a user replied to a conversation.

Evaluation Metrics. Following prior work on social media post recommendation (Chen et al., 2012; Yan et al., 2012), we treat our task on conversation recommendation as a ranking problem. Therefore, popular information retrieval evaluation metrics, including precision at K (P@K), mean average precision (MAP) (Manning et al., 2008), and normalized Discounted Cumulative Gain at K (nDCG@K) (Järvelin and Kekäläinen, 2002) are reported. The metrics are computed per user in the dataset and then averaged over all users. The values range from 0.0 to 1.0, with higher values indicating better performance.

Baselines and Comparisons. For comparison, we first consider three baselines: 1) ranking conversations randomly (RANDOM); 2) longer conversations (i.e., more words) ranked higher (LENGTH); 3) conversations with more distinct users ranked higher (POPULARITY).

[3]The datasets are available at http://www.ccs.neu.edu/home/luwang/

[4] http://trec.nist.gov/data/tweets/

[5]https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html

[6]Keyword list: "trump", "hillary", "clinton", "president", "politics", and "election."

[7]https://developer.twitter.com/en/docs/tweets/search/api-reference/get-saved_searches-show-id

[8]At least one turn per conversation is retained for training. It is possible that one user only replies in either development set or test set, but it is rather infrequent.

[9]http://www.cs.cmu.edu/~ark/TweetNLP/

Models	US Election			TREC		
	MAP	P@1	nDCG@5	MAP	P@1	nDCG@5
Baselines						
RANDOM	0.018	0.004	0.009	0.006	0.001	0.002
LENGTH	0.025	0.002	0.003	0.013	0.002	0.004
POPULARITY	0.050	0.010	0.025	0.023	0.005	0.010
Comparisons						
OCCF	0.637	0.589	0.649	0.410	0.385	0.425
RSVM	0.687	0.680	0.690	0.554	0.575	0.559
CTR	0.673	0.649	0.678	0.475	0.431	0.495
ADAPTED HFT	0.698	0.652	0.706	0.487	0.447	0.504
Our model	**0.762**	**0.750**	**0.757**	**0.591**	**0.591**	**0.600**

Table 2: Conversation recommendation results on US Election and TREC. The best result for each column is highlighted in **bold**. Our model performs significantly better than all the comparisons ($p < 0.01$, paired t-test).

We further compare results with three established recommendation models:

• OCCF: one-class Collaborative Filtering (Pan et al., 2008), which only considers users' reply history without modeling content in conversations.

• RSVM: ranking SVM (Joachims, 2002), which ranks conversations for each user with the content and Twitter features as in Duan et al. (2010).

• CTR: messages in one conversation are aggregated into one post and a state-of-the art Collaborative Filtering-based post recommendation model is applied (Chen et al., 2012).

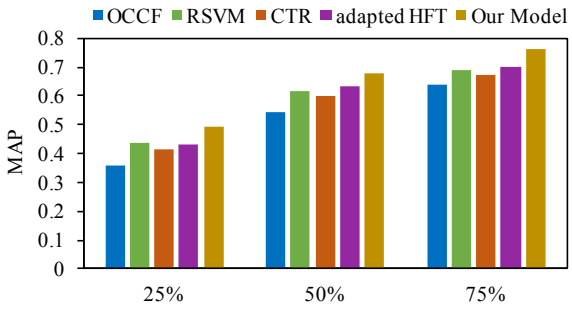
Finally, we also adapt the "hidden factors as topics" (HFT) model proposed in McAuley and Leskovec (2013) (henceforth ADAPTED HFT). Because the original model leverages the ratings for all product reviews and does not handle implicit user feedback well, we replace their user preference objective function with ours (Eq. 2).

5 Experimental Results

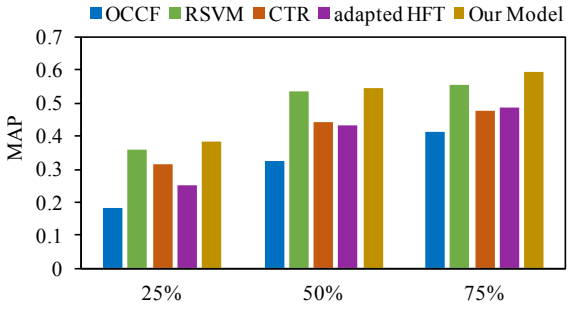
In this section, we first discuss our main evaluation in Section 5.1. A case study and corresponding discussion are provided in Section 5.2 to provide further insights, which is followed by an analysis of the topics and discourse modes discovered by our model (Section 5.3). We also examine our performance on first time replies (Section 5.4).

5.1 Conversation Recommendation Results

Experimental results are displayed in Table 2, where our model yields statistically significantly better results than baselines and comparisons (paired t-tests, $p < 0.01$). For P@K, we only report P@1, because a significant amount of users participate only in 1 or 2 conversations. For nDCG@K, different K values are experimented, which results in similar trend, so only nDCG@5 is reported.



(a) US Election dataset



(b) TREC dataset

Figure 4: MAP scores for models trained on 25%, 50%, and 75% of conversation history. For each quantile, from left to right shows the result of OCCF, RSVM, CTR, ADAPTED HFT, and our model. In general, longer conversation history leads to better performance, and our model outperforms compared systems in all settings.

We find that the baselines that rank conversations with simple features (e.g., length or popularity) perform poorly. This implies that generic algorithms that do not consider conversation content or user preference cannot produce reasonable recommendations.

Although some non-baseline systems capture content in one way or another, only ADAPTED HFT and our model exploit latent topic models to better represent content in tweets, and outperform other methods.

Compared to ADAPTED HFT, which only considers latent topics under a collaborative filtering framework, our model extracts both topics and discourse modes as latent variables, and shows superior performance on both datasets. Our discourse variables go beyond topical content to capture social behaviors that affect user engagement, such as arguments, question-asking, agreement, and other discourse modes.

Training with Varying Conversation History. To test the model performance based different levels of user engagement history, we further experiment with varying the length of conversations for training. Specifically, in addition to using 75% of

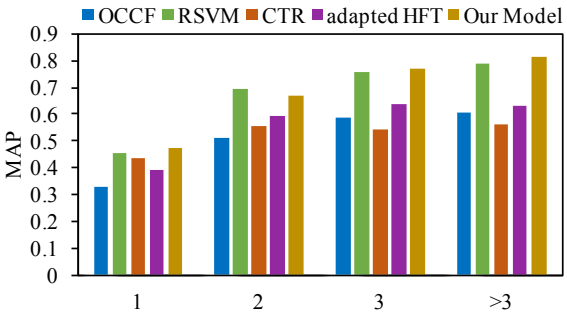


Figure 5: MAP scores of models for users involved in varying number of conversations on TREC dataset. Horizontal axis: degree of data sparsity indicated by the number of conversations a user involved in training data. Vertical axis: MAP scores. For each degree level, from left to right shows the results of OCCF, RSVM, CTR, ADPATED HFT, and our model.

conversation history, we also extract the first 25% and 50% of history as training. The rest of a conversation is separated equally for development and test. Figure 4 shows the MAP scores for US Election and TREC datasets. The increasing MAP for all methods as the training history increases indicates that generally, conversation history is essential for recommendation. Our model performs consistently better over different lengths of conversation histories.

Results for Varying Degree of Data Sparsity. From Table 1 and Figure 3, we observe that most users in our datasets are involved in only a few conversations. In order to study the effects of data sparsity on recommendation models, we examine in Figure 5 the MAP scores for users engaged in a varying number of conversations, as measured on the TREC dataset. The results on the US Election dataset have similar distributions. As we see, the prediction results become worse for users involved in fewer conversations. This indicates that data sparsity serves as a challenge for all recommendation models. We also observe that our model performs consistently better than other models over different degrees of sparsity. This implies that effectively capturing discourse structure in conversation context is useful to mitigating the effects of data sparsity on conversation recommendation.

5.2 Case Study and Discussion

Here we present a case study based on the sample conversations in Figure 1. Recall that user U_1 is interested in conversations about Sanders, and also prefers more argumentative discourse, and thus returns in conversation c_2 but not c_1.

Table 3 shows the predicted scores for the two conversations from OCCF, ADAPTED HFT, and our model (as in Eq. 2). Both ADAPTED HFT and our model more accurately recommend c_2 over c_1, with our model producing a slightly higher recommendation score for c_2.

Models	*Conv 1 (c_1)*	*Conv 2 (c_2)*
OCCF	0.941	0.922
ADAPTED HFT	0.923	0.954
Our model	0.924	0.961

Table 3: Predicted recommendation scores by different models of U_1 for conversations c_1 and c_2 in Figure 1. U_1 later replies to c_2 but not c_1, where our model predicts scores of 0.961 for c_2 (higher than 0.924 for c_1).

Latent Dim.	*User U_1*	*Conv 1 (c_1)*	*Conv 2 (c_2)*
Topic 1 (Sanders)	0.92 ($\gamma^U_{u_1,1}$)	0.10 ($\gamma^C_{c_1,1}$)	0.63 ($\gamma^C_{c_2,1}$)
Topic 2 (Clinton)	0.14 ($\gamma^U_{u_1,2}$)	0.84 ($\gamma^C_{c_1,2}$)	0.12 ($\gamma^C_{c_2,2}$)
Disc 1 (argument)	0.46 ($\delta^U_{u_1,1}$)	0.28 ($\delta^C_{c_1,1}$)	0.38 ($\delta^C_{c_2,1}$)
Disc 2 (statement)	-0.24 ($\delta^U_{u_1,2}$)	0.98 ($\delta^C_{c_1,2}$)	-0.09 ($\delta^C_{c_2,2}$)

Table 4: Sample latent dimensions of topics ($\gamma^U_{u_1}$ for user, and $\gamma^C_{c_*}$ for conversations) and discourse modes ($\delta^U_{u_1}$ for user, and $\delta^C_{c_*}$ for conversations). User U_1 shows interest in topic 1 (about Sanders), which is also a dominating topic in conversation c_2, but is not interested in topic 2 (about Clinton). U_1 shows a preference for discourse mode 1 (argument) over mode 2 (statement).

Table 4 shows the latent dimension values for the learned topics and discourse modes for this user and these two conversations. Based on human inspection, topic 1 appears to contain words about Sanders, which is the main topic in conversation c_2. Topic 2 is about Clinton, which is a dominating topic in conversation c_1. Our model also picks up user interest in topic 1 (Sanders), and thus assigns $\gamma^U_{u_1,1}$ a high value. For discourse modes, our model also generates a high score for "argument" discourse (labeled via human inspection) for both the user and c_2.

5.3 Further Analysis of Topic and Discourse

Ablation Study. We have shown that joint modeling of topical content and discourse modes produces the superior performance for our model. Here we provide an ablation study to examine the relative contributions of those two aspects by setting the trade-off parameter λ to 1.0 (topic only) or 0.0 (discourse only). Table 5 shows that topics or discourse individually improve slightly upon the comparison ADAPTED HFT, but only jointly do they improve significantly upon it.

Topic Coherence. To examine the quality of topics found by our model, we use the C_V topic coherence score measured via the open-source

Models	US Election	TREC
ADAPTED HFT	0.698	0.487
Our model (topic only)	0.711	0.491
Our model (discourse only)	0.705	0.483
Our model (full)	**0.762**	**0.591**

Table 5: MAP of different variants of our model. Best results in each column is in **bold**.

toolkit Palmetto[10], which has been shown to produce evaluation performance comparable to human judgment (Röder et al., 2015). Our model achieves topic coherence scores of 0.343 and 0.376 on TREC and US Election datasets, compared to 0.338 and 0.371 for the topics from ADAPTED HFT.

Sample Discourse Modes. While our topic word distributions are relatively unsurprising, of greater interest are the discourse mode word distributions. Table 6 shows a sample of discourse modes as labeled by human. Although this is merely a qualitative human judgment at this point, there does appear to be a notable overlap in discourse modes between the two datasets even though they were learned separately.

Discourse	Top 10 Terms	
	US Election	TREC
Question	? it if ... so all how because when any	? : HASHTAG or too with MENTION and ... what
Reaction	you like any good ! please no ~ lol what	all EMOTICON & !!! right ok u ;) thank haha
Statement	's do think the . should they from and have	i , a what all you be how then ...
Argument	but that all fuck without against out though ! anything	do would up that too even always never anything much
Reference	be i about that MENTION it " you -lrb- ?	MENTION ... ! : what rt it you URL :)

Table 6: Top 10 representative terms for sample discourse modes discovered by our model in two datasets. Names of discourse modes are our interpretations according to the word distributions generated by our model.

5.4 First Time Reply Results

From a recommendation perspective, users may be interested in joining new conversations. We thus compare each recommendation system for first time replies. For each user, we only evaluate for conversations where they are newcomers. Table 7 shows that, unsurprisingly, all systems perform poorly on this task, though our model performs slightly better. This suggests that other features, e.g., network structures or other discussion thread features, could usefully be included in future studies that target new conversations.

Models	US Election	TREC
OCCF	0.035	0.033
RSVM	0.023	0.002
CTR	0.029	0.016
ADAPTED HFT	0.054	0.058
Our model	**0.083**	**0.090**

Table 7: MAP of models considering only first time replies. Best results in each column is in **bold**.

[10] https://github.com/AKSW/Palmetto/

6 Conclusion

This paper has presented a framework for microblog conversation recommendation via jointly modeling topics and discourse modes. Experimental results show that our method can outperform competitive approaches that omit user discourse behaviors. Qualitative analysis shows that our joint model yields meaningful topics and discourse representations.

Acknowledgements

This work is partly supported by Innovation and Technology Fund (ITF) Project No. 6904333, General Research Fund (GRF) Project No. 14232816 (12183516), and National Science Foundation Grant IIS-1566382. We thank Shuming Shi, Yan Song, and the three anonymous reviewers for the insightful suggestions on various aspects of this work.

References

Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*. ACM, LinkKDD '05, pages 36–43.

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, pages 183–194.

Noor Aldeen Alawad, Aris Anagnostopoulos, Stefano Leonardi, Ida Mele, and Fabrizio Silvestri. 2016. Network-aware recommendations of novel tweets. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, pages 913–916.

Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 602–606.

Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. 2013. Characterizing and curating conversation threads: expansion, focus, volume, re-entry. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, pages 13–22.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Ceren Budak and Rakesh Agrawal. 2013. On participation in group chats on twitter. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, pages 165–176.

Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 661–670.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, CSCW '17, pages 1217–1230.

Nigel Crook, Ramón Granell, and Stephen G. Pulman. 2009. Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Proceedings of The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL 2009*. pages 341–348.

Yajuan Duan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. 2010. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pages 295–303.

Wei Feng and Jianyong Wang. 2013. Retweet or not?: personalized tweet re-ranking. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, pages 577–586.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. pages 42–47.

Vicenç Gómez, Hilbert J Kappen, and Andreas Kaltenbrunner. 2011. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*. ACM, pages 181–190.

Tom Griffiths. 2002. Gibbs sampling in the generative model of latent dirichlet allocation .

Yue He and Jinxiu Tan. 2015. Study on sina microblog personalized recommendation based on semantic network. *Expert Systems with Applications* 42(10):4797–4804.

Liangjie Hong, Aziz S Doumith, and Brian D Davison. 2013. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, pages 557–566.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, pages 263–272.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20(4):422–446.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pages 133–142.

Shafiq R. Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011*. pages 1807–1813.

D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.

Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *IEEE Computer* 42(8):30–37.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide Web*. ACM, pages 591–600.

David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *ICWSM*.

Chei Sian Lee and Long Ma. 2012. News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior* 28(2):331–339.

Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7(1):76–80.

Annie Louis and Shay B. Cohen. 2015. Conversation trees: A grammar model for topic structure in forums. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1543–1553.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, pages 165–172.

Jorge Nocedal. 1980. Updating quasi-newton matrices with limited storage. *Mathematics of computation* 35(151):773–782.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11(122-129):1–2.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. pages 380–390.

Rong Pan, Yunhong Zhou, Bin Cao, Nathan Nan Liu, Rajan M. Lukose, Martin Scholz, and Qiang Yang. 2008. One-class collaborative filtering. In *Proceedings of the 8th IEEE International Conference on Data Mining*. pages 502–511.

Ye Pan, Feng Cong, Kailong Chen, and Yong Yu. 2013. Diffusion-aware personalized social update recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, pages 69–76.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*. pages 2270–2276.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*. pages 172–180.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. pages 399–408.

Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*. ACM, pages 880–887.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* 26(3):339–373.

Jan Vosecky, Kenneth Wai-Ting Leung, and Wilfred Ng. 2014. Collaborative personalized twitter search with topic-language models. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, pages 53–62.

Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pages 448–456.

M Woszczyna and A Waibel. 1994. Inferring linguistic structure in spoken language. In *Proceedings of ICSLP*. IC-SLP.

Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World Wide Web*. ACM, pages 705–714.

Rui Yan, Mirella Lapata, and Xiaoming Li. 2012. Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, pages 516–525.

Yang Yu, Xiaojun Wan, and Xinjie Zhou. 2016. User embedding for scholarly microblog recommendation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. volume 2, pages 449–453.