Latent Variable Modeling for Natural Language Understanding

ZENG, Jichuan

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer Science and Engineering

The Chinese University of Hong Kong
September 2019

Thesis Assessment Committee

Professor LO Chi Lik Eric (Chair)

Professor KING Kuo Chin Irwin (Thesis Supervisor)

Professor LYU Rung Tsong Michael (Thesis Co-supervisor)

Professor LEE Pak Ching (Committee Member)

Abstract of thesis entitled:

Latent Variable Modeling for Natural Language Understanding

Submitted by ZENG, Jichuan

for the degree of Doctor of Philosophy

at The Chinese University of Hong Kong in September 2019

Natural Language Understanding (NLU) is focusing on enabling machine to understand and process unstructured human language. NLU is one of the core parts of natural language processing with a wide range of applications, such as text classification, sentiment analysis, question answering, etc. However, most learning based NLU methods either rely on pre-defined, hand-crafted features, or high-quality human annotated data for training.

With the risen of social media and arrival of information explosion era, the large volume and short nature of social media text data bring a lot of challenges to conversational NLU methods, making the feature engineering and data annotation very labor-intensive and domain-specific. In this thesis, we propose to model latent variables (i.e., topics and discourse) on social media text (e.g., microblog, online forum) in an unsupervised way, based on which, we design our NLU models and demonstrate their superior performance on two social media NLU tasks: short text classification and argumentation mining.

First, we propose an unsupervised framework for jointly

modeling topic content and discourse behavior in microblog conversations for understanding the semantics and interaction of social media messages. Concretely, we propose a neural model to discover word clusters indicating what a conversation concerns (i.e., topics) and those reflecting how participants voice their opinions (i.e., discourse). Extensive experiments show that our model can yield both coherent topics and meaningful discourse behavior. Our model can be easily extended with other neural networks. Further study shows that our topic and discourse representations can benefit the classification of microblog messages, especially when they are jointly trained with the classifier.

Second, we focus on the short text classification, which is one of the most fundamental techniques in social media text understanding. To address data sparsity issue in social media short text, we propose topic memory networks for short text classification with a novel topic memory mechanism to encode latent topic representations indicative of class labels. Different from most prior work that focuses on extending features with external knowledge or pre-trained topics, our model jointly explores topic inference and text classification with memory networks in an end-to-end manner. Experimental results on four benchmark datasets show that our model outperforms state-of-the-art models on short text classification, meanwhile generates coherent topics.

Third, we focus on the online argumentation, which is a growing and challenging field in social media text understanding. We present a novel study that automatically analyzes the key factors of persuasiveness in argumentation process, beyond simply predicting who will win the debate. Specifically, we

propose a novel neural model which is able to dynamically track the changes of latent topics and discourse in argumentative conversations, allowing the investigation of their roles in influencing the outcomes of persuasion. Extensive experiments have been conducted on argumentative conversations on both online forum and supreme court. The results show that our model outperforms state-of-the-art models in identifying persuasive arguments. We further analyze the effects of topics and discourse on persuasiveness, and draw some findings from our empirical results, which will help people better engage in future persuasive conversations.

論文題目：基于隱變量建模的自然語言理解
作者　　：曾紀川
學校　　：香港中文大學
學系　　：計算機科學與工程學系
修讀學位：哲學博士
摘要　　：
自然語言理解(NLU)關注于使機器能理解和處理非結構化的人類語言。NLU是自然語言處理核心部分之一，有廣泛的應用，比如文本分類，情感分析，問題解答等。然而，大多數基於學習的NLU方法或者依賴於預定義的人為製作的特徵，或者訓練所需的高質量的人工標註數據。隨著社交媒體和信息爆炸時代的到來，大量簡短的社交媒體文本給對話式NLU方法帶來了大量的挑戰，使得特徵工程和數據標註極為需要人力和領域特定。本論文中，我們提出了社交媒體文本（比如微博和在論壇）上的無監督模型隱變量（即主題和論述），基於此，我們設計了新的NLU模型並且展示了它們在兩個社交媒體自然語言理解任務上優越的性能：短文本分類和辯論挖掘。
首先，我們提出了微博對話中聯合建模主題內容和論述行為的無監督模型以理解社交媒體信息的意義和交互。具體地，我們提出了一個神經模型來發現表徵一組對話設計的內容（即主題）和反應參與者如何表達他們觀點（即論述）的詞聚類。大量實驗表明我們的模型可以生成連貫的主題和有意義的論述行為。我們的模型可以比較容易地跟其它的神經網絡結合擴展。進一步的研究表明我們的主題和論述表徵能夠有助於微博信息的分類，特別是當它們與分類器聯合訓練的時候。

其次，我們集中在短文本分類這個最基本的社交媒體文本理解的技術。為了解決社交媒體短文本的數據稀疏問題，我們提出了為短文本分類的主題記憶網絡，即用一個新的主題記憶機制來編碼表示類標註的隱含主題表徵。不同於大多數以往關注在用外部知識擴展特徵或者預訓練的主題，我們的模型用端到端的方式基於記憶網絡聯合探索了主題推斷和文本分類。在四個基準數據集上的實驗結果表明我們的模型在短文本分類上優於最新的模型，且生成連貫的主題。

再其次，我們關注在辯論這個在社交媒體文本理解中不斷增長的且有挑戰性的領域。我們呈現了一個新的自動分析辯論階段中說服的關鍵因素的研究，而不僅僅是簡單地預測誰講在辯論中獲勝。特別地，我們提出了一個新的神經模型，來自動地動態追蹤隱含主題和論述在辯論對話中的變化，並允許探索它們在影響說服結果中的角色。大量在辯論對話場景中的實驗被執行，包括在論壇和最高法院。結果表明我們的模型在識別說服性的辯論上優於最新的模型。我們進一步分析了主題和論述在說服中的影響，並發現它們都很有用。此外，我們總結了基於實驗結果的一些發現，可以幫助人們更好地參與未來的說服性對話。

Acknowledgement

This thesis would never be possible without the help and support from my family, friends, and colleagues, etc, and I would like to begin with the words in Bible, which is also the voice of my heart.

"Except the Lord build the house, they labour in vain that build it: except the Lord keep the city, the watchman waketh but in vain."

— *Psalms 127:1*

First and foremost, I would like to thank my supervisors, Prof. Michael R. Lyu and Prof. Irwin King for their kind supervision of my PhD study at CUHK. I am greatly lucky to have they as my supervisors. They help sharpen my research taste and build my research expertise, from maturing a research idea to designing implementation details, from paper writing to conference presentation. They also provide strong support and immense care for my personal life. Especially, greatly thanks Prof. Michael R. Lyu for shepherding me in my life of faith, and also Prof. Irwin King for delivering me valuable and helpful instructions on how to balance work and family. I will always be grateful to their advice, encouragement and support at all levels.

I would like to extend my gratitude to thank my thesis assessment committee members: Prof. Eric Lo, and Prof. Patrick Lee,

for their constructive comments and valuable suggestions to this thesis and all my term reports. Great thanks to Prof. Maggie Li from the Hong Kong Polytechnic University who kindly served as the external examiner for this thesis.
I would like to thank Prof. Jing Li, my mentor during the internship at Tencent AI Lab. She unreservedly provides me valuable support for my research in the thesis. I also thank friends met there, Haisong Zhang, Jialong Han, Guanlin li, Lingzhi Wang, Ming Fan and Lu Ji for their kindness and help.
I would like to thank Dr. Taifeng Wang, my mentor during the internship at Microsoft Research Asia. I also really thank friends met in MSRA, Yingce Xia, Fei Gao and Weiqing Liu, for their inspiration and encouragement.
I would like to thank my life-long friends, Linhao Ling, Li Huang, Zhifeng Liu, Runxiang Lin, Ke Wang, Jiahui Li, Jiazhi Tian, Benli Yang, Xiangfeng Zou, Guoming Chen, Yuan Li, Yongbin Wen, Huimiao Shi, Xuemin Cao, Jingmei Zhang and Yonghui Guo for their trust and patience.
I would like to thanks my brothers and sisters in the church, Joshua Lee, Yuqi Chen, Tommy Tseng, Jinghua Jiang, Ruiqing Fu, Chauncey Wang, Jiaojiao Fu, Yuanyuan Man, Carrie Zhou, Vitoria Liu, Jessica Feng, Minzhi Ou and Fan Zhang. Especially, I am highly appreciated Dr. Felicia Lyu and Dr. Janettey Chan for their firm trust and shepherd on me always. Their words and selfless care lighten the darkest period in my life.
I am also thankful to my other groupmates, Yangfan Zhou, Zibin Zheng, Haiqin Yang, Jian Li, Hou Pong Chan, Jiani Zhang, Shilin He, Chen Cheng, Yu Kang, Tong Zhao, Hongyi Zhang, Shenglin Zhao, Xixian Chen, Yuxin Su, Xiaotian Yu, Pengpeng Liu, Yue Wang, Wang Chen, Han Shao, Haoli Bai, Wenxiang

Jiao, Jingjing Li, Yifan Gao, Weibin Wu, and Zhuangbin Chen, who gave me encouragement and kind help.
Last but most important, I dedicate my greatest of gratitude to my wife Cuiyun Gao, my daughter Priscilla, and my parents. It is their deep and everlasting love that drives me through all the difficulties in my Ph.D study.

To my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis presents our research on latent variable modeling for natural language understanding, which is an important field of natural language processing with a wide range of applications. We provide a brief overview of the research problems under study in Section 1.1, and highlight the main contributions of this thesis in Section 1.2. Section 1.3 outlines the thesis structure.

1.1 Overview

In the last decades, we've witnessed the risen and popularity of the Internet. The way of communication between people has been revolutionized by breaking the limitation of region, space and time. Online platforms, such as Twitter[1], Sina Weibo[2], and Reddit[3] become important outlets for people to share information and voice opinion, evidenced by many recent events like "Notre-Dame de Paris fire", and Donald J. Trump on Twitter. The flourish of social media has led to the large amount of text data produced by various social media

[1] twitter.com
[2] weibo.com
[3] reddit.com

services every day. Those social media text data fertilize a wide range of real-life applications, for example, breaking event detection [74, 100, 138], real-time sentiment analysis [117, 132], user profiling [22, 139], advertising [67] and social chatbot [114]. However, the explosive growth of the social media text data far outpaces human beings' ability of reading and understanding. Table 1.1 tells two stories that commonly happen in our daily life. Although the advance of big data analysis technology enables large scale text analysis, people are still being exposed to superfluous information, facing the challenge of information explosion. Such problem can seriously affect lots of web applications, such as market decision [42], and stock prediction [15], if they have a superficial understanding of the social media messages. As a consequence, there is a pressing need for automatic language understanding techniques for processing and analyzing social media texts [163]

Social media text provides us rich information and insight for detecting social event and understanding social interaction. To help users to distill the useful information among the huge quantity of social media textual data, in this thesis, we focus on the research problem of natural language understanding of social media text. Unfortunately, different from the conventional formal and well-edit text, such as news article [162] and scientific paper [111], textual data in social media presents many new challenges due to its distinct characteristics. We conclude three challenges for understanding the social media text:

Short and noisy. Certain social media platforms have character limit on the message user created. In Twitter, users are restricted to type the message less than 280 characters. Similar, Pinterest board description is limited to 500 characters length.

Story of Justin Ng Justin Ng is a typical office worker in Hong Kong, another identify of him is a fan of Jay Chou, an pop singer. Once he knew that Jay Chou was going to give concert in Hong Kong at March 19, but unfortunately, he missed the sale time of ticket, and found that only expansive seats are available. So he started to search on the microblog hoping that there is someone that want to transfer the ticket. But he quickly gave up, since he was overwhelmed by enormous replicated irrelevant messages.
Story Jie Zhang Jie Zhang is a computer science Ph.D student in CUHK from Mainland China. Recently, she was invited to join to a Wechat group, which is for sharing research experience and communicating ideas, and the group has over 400 members. Jie Zhang edited her profile and turned off the notification of this Wechat group, since lots of conversations on it are irrelevant to her research. One day, Jie Zhang received a message from this Wechat group, which "@" her, and says "It is a nice work, should be helpful for you@jiezhang". She was very confused: which work it refers to? To find out the answer, she spent half an hour to trace the chat history.

Table 1.1: Two stories in the Internet era.

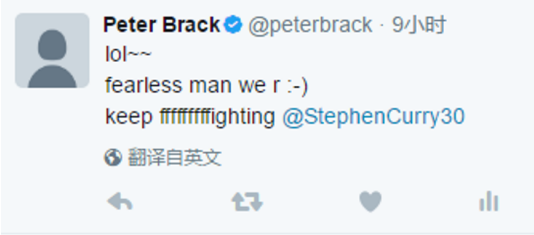


Figure 1.1: A message from Twitter.

Although the character limitations have been extended (e.g., Twitter [2]) or even canceled (e.g., Weibo [3]) over time, people still get used to writing shorter messages. For example, the average length of a tweet is 28 characters, and most Facebook posts are within 40-80 characters [2]. People like when a message makes its point quickly and concisely. Shorter tweets or posts usually receive more likes, comments, and shares. As a consequence, short length social media text is ubiquitous on the web. In addition to short in length, social media messages are typically informal and full of noise. Because most of social media usage coming via mobile devices, it is common that people write messages with misspelling words, emojis, abbreviations, and slang, as shown in a tweet example of Figure 1.1. Short and noisy social media text data have very limited textual features and lack the contextual information, current existing method for well-edited and formal text are inevitable compromised when facing such data. Therefore, successful processing of such short and noisy text is essential for social media text understanding.

Huge volume. With the flourish of social media, a massive amount of social media messages generated every day. Take Sina Weibo, the most widely-used microblog site in China, as an example. In 2018, the number of monthly active users has reached 462 million, over 130 million posts are created every

day [1]. Such huge volume of social media text far outpaces human beings' ability of reading and understanding. There is a pressing need for automatic language understanding techniques for processing and analyzing In addition to huge volume of data, informal and noisy social media messages make it extremely time-consuming for human editors to make annotations, which is the basis of many supervised learning models, such as sentiment classification, question answering. So how can we process and distill useful information from such huge volume of unlabeled data, is an important problem this thesis focuses on.

Open domain. Social media platforms are typically open-domain, which exhibit a rich variety of information sources. People with various interests keep posting, commenting, and sharing messages with a wide variety of topics, including entertainment, politics, social events, science, business, etc. In domain-specific corpus, where pre-defined schemes and hand-crafted features are designed to boost the performance of tasks, such as task-oriented chatbot, medical QA. In comparison, social media text is open-domain, it is hard to incorporate the domain-specific knowledge, such as knowledge base and hand-crafted features. Therefore, It requires us to develop fully data-driven methods that can handle such open domain social media text.

Although social media text data have the above challenges, there still existing lots of automatic methods with the help the advance of natural language understating techniques. Topic modeling, is one of the most famous and widely used techniques for understanding the social media text among them. Topic modeling can automatically cluster the massive and diverse social media text without any annotation according to the latent topics they belong to. Topic modeling is typically the

first step for many language understanding tasks, such as text classification and text summarization. However, topic modeling for social media text is another challenging task. Conventional topic modeling methods, such as LDA [8] and pLSA [41], can work very well in formal and well-edited text, but their performance drop sharply when facing short and noisy social media text.

Social media also provide abundant non-content information, which is helpful for understanding the semantic meaning of social media messages. For example, Twitter has a clear conversation structure information, indicating who replies to who. Such structure information is commonly used to study the interaction of social behaviors, which can help us to enrich the contextual information of each social media message. Therefore, if we can organize social media messages into the tree structure conversations, the data sparsity issue can be alleviated. Towards key focus understanding of a conversation, previous work has shown the benefits of discourse analysis [70, 72, 102], which shapes how messages interact with each other forming the discussion flow. Discourse, such as contrast, elaboration, statement, is originally defined to capture the semantic or pragmatic relation between sentences in a document. Recent study about discourse has involved identifying performative function (e.g., "question", "response") of each utterance in dialogue, i.e., dialogue acts, as the shallow conversation structure [25, 124]. Due to the short nature of social media text, each message in conversation may only contain one type of discourse, such as "statement", "comment", which capture the illocutionary meanings of an utterance. Table 1.2 gives an snippet of Twitter conversation. As we can see, message M_2 questions the

M_1: Louisiana flood victims praise Donald Trump for visiting damaged areas!
M_2: Total time: 49 seconds helping for photo op. Also Playdoh, damn helpful in a flood.
M_3: food water hygiene supply's Play Dough was for kids obvious the kids all in shelter from flood have no toys duh
M_4: Maybe he can bring you some punctuation
M_5: 😂funny all Libs do that when they have no answer they correct grammar ~fail~ this is Twitter not school

Table 1.2: A snippet of Twitter conversation about Trump visiting Louisiana flood victims.

statement of "Louisiana flood victims praise Donald Trump" posted by M_1 through emphasizing "Total time: 49 seconds". Message M_3 gives a comment of M_2's viewpoint, by saying "Play Dough was for kids". M_4 and M_5 are arguments about the punctuation. We can find that there is a clear discourse flow that carries the conversation forward in the above example, via making a statement, posting a question, giving a comment, and so on so forth. Therefore, discourse structure embedded in the social media conversation can usefully reflect salient topics raised in the discussion process.

The research of this thesis comprises three parts. In the first part, we focus on the study of unsupervised modeling social media conversation. Specifically, we explore the joint effect of modeling latent topics and discourse by utilizing the microblog conversation structure information. In the second part, we focus on the fundamental social media text understanding technique, short text classification. In particular, we propose to incorporate the latent topic representations indicative of class labels into neural classifier. In the third part, we focus on a more challenging task, exploring the reasons behind the persuasiveness of online argumentation.

1.2 Thesis Contributions

In this thesis, we make contributions to the understanding of social media text in the following ways:

1. **Joint modeling of topics and discourse in microblog conversations**

   Microblog conversation is ubiquitous in social media platform (e.g., Twitter, Weibo). We present an unsupervised neural network framework for jointly modeling topic content and discourse behaviors in microblog conversations. In particular, we propose a neural model to discover word clusters indicating what a conversation concerns (i.e., topics) and those reflecting how participants voice their opinions (i.e., discourse). Extensive experiments on Twiiter conversation dataset show that our model can generate coherent topics and meaningful discourse roles. Furthermore, Our model can be easily extended with other neural network architectures (such as CNN) to present better performance with end-to-end joint training.

2. **Short text classification for social media messages**

   Short text classification is one the most fundamental techniques for automatic social media text understanding. Due to data sparsity nature of social media messages, conventional classifier designed for formal and well-edited text work poorly facing such short text messages. To address this issue, we propose *topic memory networks* for short text classification with a novel topic memory mechanism to encode latent topical representations. Compared with existing work where most previous efforts focus on extending

features with external knowledge or pre-trained topics, our model jointly explores topic inference and text classification with memory networks in an end-to-end manner. Extensive experiments on four benchmark outperforms state-of-the-art models on short text classification, meanwhile generates coherent topics.

3. **Automatically identifying key factors of persuasiveness in online argumentation**

   Online argumentation mining is a growing field in social media text understanding. Most of the previous work focuses on crafting hand-made features to predict which side is more convincing. However, such a task has proven to be hard, and the prediction results are just slightly better than random guess [126]. Here we take a step further and presents a study that automatically analyzes the key factors in argument persuasiveness, beyond simply predicting who will win the debate. We propose a novel neural model which is able to dynamically track the changes of latent topics and discourse in argumentative conversations, allowing the investigation of their roles in influencing the outcomes of persuasion. We carry out extensive experiments on argumentative conversations from both social media and supreme court. The results show that our model can effectively identify persuasive arguments, significantly outperforming state-of-the-art methods on both datasets. We also draw some findings from our empirical results, which will help people better engage in future persuasive conversations.

1.3 Thesis Organization

The remainder of this thesis is organized as follows:

- **Chapter 2**

  In this chapter, we give a systematic review of the background knowledge and related work on latent variable modeling and natural language understanding for social media. First, we briefly introduce social media text data with a focus on its characteristics and applications. Then we review topic modeling, one of most Representative latent variable modeling methods for text data, including conventional topic models, topic modeling on social media text, and neural topic modeling. After that, we introduce discourse analysis, with a focus on the role of discourse in social media conversation. Finally, we review the task of natural language understanding for social media, providing the related work of short text classification and argumentation mining.

- **Chapter 3**

  This chapter presents an unsupervised neural network framework for jointly modeling topic content and discourse behavior in microblog conversations, with the aim to automatically analysis what a conversation is talking about and how the opinion is voiced in each message. To be specific, section 3.1 introduces social media and the motivation of understanding of microblog conversation. Section 3.2 presents our unsupervised neural framework that can joint exploration of word clusters to represent topic and discourse in microblog conversations. Evaluation

results are presented is Section 3.4. Finally, we conclude this chapter in Section 3.5.

- **Chapter 4**

  In this chapter, we focus on the fundamental language understanding technique for social media text, short text classification, and propose a novel topic memory mechanism to encode latent topic representations indicative of class labels, which outperforms the state-of-the-art short text classifier. Section 4.1 present the motivation and intuition of our short text classifier. Section 4.2 describe our topic memory networks, consisted of neural topic model, topic memory mechanism, and neural network classifier. We evaluate the performance of topic memory networks in Section 4.4, and conclude this chapter in Section 4.5.

- **Chapter 5**

  In this chapter, we focus on the online argumentation and present a novel study that automatically analyzes the key factors in argument persuasiveness, beyond simply predicting who will win the debate. More specifically, Section 5.1 introduce the motivation of analysis the key factors of persuasiveness in argumentation process. We briefly describe our data and problem formulation in Section 5.2. In Section 5.3, We propose a novel neural model which is able to dynamically track the changes of latent topics and discourse in argumentative conversations. We evaluate the proposed model in Section 5.5 and draw some insights from our empirical results in Section 5.6.

- **Chapter 6**

The last chapter summarizes this thesis and provides some potential future directions for social media text understanding that deserve for further exploration.

□ **End of chapter.**

Chapter 2

Background Review

This chapter briefly reviews some background knowledge and related work of our research. First, we provide the background knowledge about social media text data. Then we explain the two basic techniques this thesis focuses on, topic modeling in Section 2.2 and discourse analysis in Section 2.3. Next, we describe the natural language understanding for social media text in Section 2.4, including two tasks: text classification in Section 2.4.1, and argumentation mining in Section 2.4.2.

2.1 Social Media Text Data

Social media such as microblog, online forum, multimedia sharing sites, are extensively used in our daily, playing an important role for people to communicate breaking news, voice opinions, participate in events, and connect to each other from anywhere, at anytime. Social media is an important traffic channel for current web applications, accounts for 60% of top 10 sites according to the statistics from RankRanger [1] as shown in Table 2.1. These social media provide rich information of

[1] https://www.rankranger.com/top-websites

Rank	Website	Rank	Website
1	**Wikipedia**	6	**Imdb**
2	**Twitter**	7	Apple
3	Google	8	Amazon
4	**Youtube**	9	Merriam-webster
5	**Facebook**	10	**Instagram**

Table 2.1: Internet Traffic Report by RankRanger on August 3rd, 2019. Social media websites are in **boldface**.

people interaction and collective behavior, thus have attracted widespread attention in sociology, psychology, business, political science, computer science, economics, and other social and other disciplines.

The rapid growth of social media leads to large quantity of user-created text. For example, there are 500 million tweets sent each day on Twitter, that means 5,787 tweets generated per second [2]. Figure 2.1 gives an example of Twitter messages about Donald Trump visiting Louisiana flood victims. Compared with traditional text data from news article or well-edited books, social media text is typically short in length. The average length of a tweet is 28 characters, and most Facebook posts are within 40-80 characters [2]. Moreover, social media messages are typically informal and full of noise. Because most of social media messages are composed in mobile devices, it is common that people write messages with syntax errors, abbreviations, and slang, as shown in Figure 2.1. As a result, there are very limited textual features for understanding the meaning of social media messages without giving the context information. Another feature of social media text is the abundant non-text information, which is helpful for understanding the semantic

[2] https://blog.hootsuite.com/twitter-statistics/

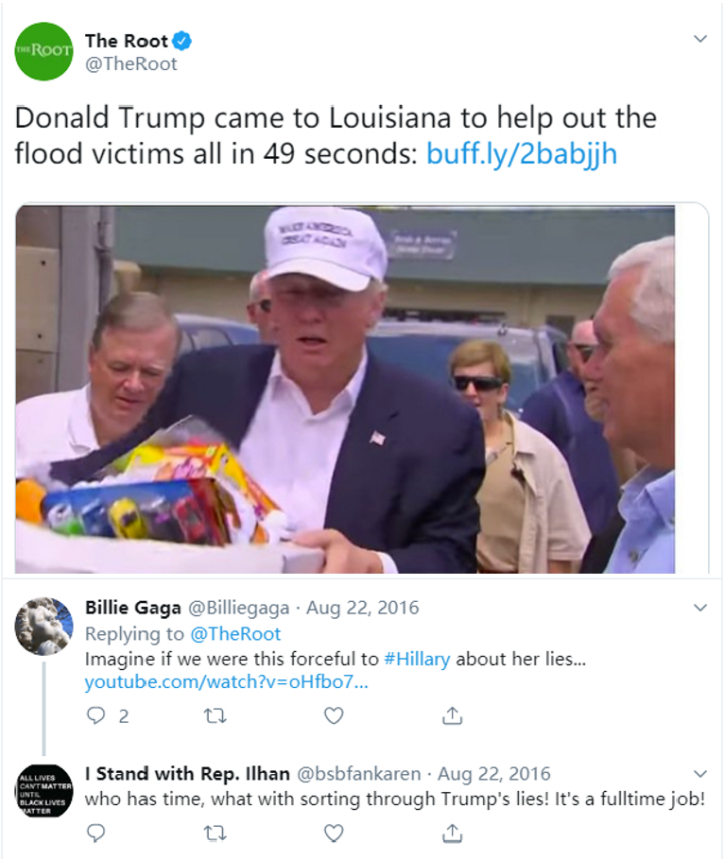


Figure 2.1: Example Twitter messages about Trump visiting Louisiana flood victims.

meaning of social media messages. For example, people tend to incorporate images and video in their messages for better engagement of other uses. There is also a clear conversation structure information, indicating who replies to who. Such non-text information is commonly used to study the interaction of social behaviors, which can help us to enrich the contextual information of social media messages.

Social media text provides large volume of social conversations and user interactions. A wide range of real-life applications are built on social media text data, for example, breaking event detection [74, 100, 138], real-time sentiment analysis [117, 132],

user profiling [22, 139], advertising [67] and social chatbot [114].

2.2 Topic Modeling

Topic model has achieved huge success in lots of areas in the last decades. It can automatically discover the pattern of words in the document, indicating the word clusters as the latent "topic" representations from texts. We brief introduce Latent Dirichlet Allocation (LDA) in Section 2.2.1. Then we will discuss the topic modeling in social media text in Section 2.2.2. At last, we will introduce the latest neural network based topic modeling in Section 2.2.3.

2.2.1 Conventional Topic Modeling

Topic modeling provides an effective way to analyse large scale of unlabeled text data. One of the most well-known topic models is Latent Dirichlet Allocation (LDA) [13].
The assumption of LDA is that each document is a mixture of topics, where a topic is a probabilistic distribution over words. In other words, LDA is a generative model, which specifies a probabilistic procedure for reconstructing the document. The generation procedure can be described as a writer to compose a story. The writing procedure is the repeat of the following steps: (1) The writer picks a topic z from the topic mixture θ_d of the document d. (2) from the word mixture β_z of topic z, the writer picks a word w and write it down.
LDA model each document d as a mixture of latent topics θ_d, following a multinomial distribution, each latent topic describes a multinomial distribution β_z over word vocabulary. The parameters of the multinomial for topics have a Dirichlet prior [13].

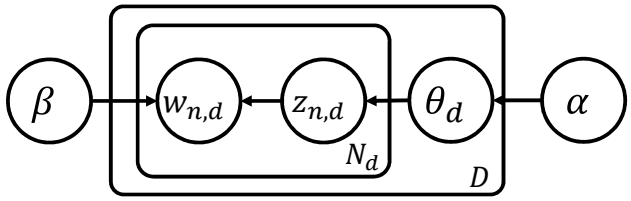


Figure 2.2: A graphical model representation of LDA.

Figure 2.2 shows the overall graphical representation of LDA. Therefore, the above writing process can be formulated into the following form:
For each document d:

- Draw a topic mixture $\theta_d \sim Dirichlet(\alpha)$
- For each word $w_{d,n}$ in d:
  - Draw a topic $z_{d,n} \sim Multinomial(\theta_d)$
  - Draw a word $w_{d,n} \sim Multinomial(\beta_{z_{d,n}})$

In the above generative process, the only observed variable are the given collection of words in documents, others are latent variables (θ and β) and hyper parameters (α). In order to estimate the latent variables and hyper parameters, we need to maximize the probability of seeing these words in documents W as follows:

$$p(W|\alpha,\beta) = \prod_{d=1}^{D} \int p(\theta_d|\alpha) \prod_{n=1}^{N_d} \sum_{z_{d,n}} p(z_{d,n}) p(w_{d,n}|z_{d,n},\beta) d\theta_d \quad (2.1)$$

where D is the number of documents, N_d is number of words in document d. The variables θ_d are document-level variables, sampled in each document. $z_{d,n}$,$w_{d,n}$ are word-level variables, sampled for each word in each document.
LDA is an excellent tool for the modeling distribution of potential topics for large corpora. Therefore, it is able to identify sub-topic in the text documents composed of many patents

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 2.3: Top: 15 most probable words for four selected topics. Bottom: a text document with words colored according to which topic they belong to. [13]

and represents each patent in a series of subject distributions. Figure 2.3 gives an example of what LDA can learn from text document. By using LDA, we can discover the hidden topics given set of documents. Here, each document is then treated as a probability distribution for a set of topics, and each topic is a probability distribution of for a set of words.
Typically, there are two ways for estimating the parameters of LDA, variational EM [13] and Gibbs sampling [122].

- Variational EM, is a type of Expectation-Maximization algorithm of variational inference. EM is a powerful method for estimating the parameter of graphical model in an unsupervised way. However, in LDA, the calculation of integral of marginal likelihood is intractable, variational

EM uses a parametric approximation (e.g. mean-field) to the posterior distributions and the latent variables. Therefore, the objection of Variational EM is to optimize the fit of variational approximation to the true posterior via KL-divergence. This method is also applied to solve the parameter inference in LDA-alike topic models [9, 23, 167].

- Gibbs sampling, is a Monte Carlo Markov-chain approach to generate a sample from a joint distribution, which is widely used in parameter estimation of topic models. Since LDA has the probabilistic conjugated pair in generative process, that is (α, θ). People also parameterize the topic word mixture β into the conjugated form (β, ϕ), by treating the topic word mixtures as ϕ, and β as the prior. Therefore, an efficient form of Gibbs sampling, collapsed Gibbs sampling, can be performed in LDA. When the Markov chain converges, we can infer the multinomial from the state of sampled latent variables. More details about conjugacy of Dirichlet and multinomial distributions, and the collapsed Gibbs sampling can be found in Mark Steyvers's tutorial [122].

LDA is a springboard for topic modeling, and lots of researches have proposed methods which are extended from LDA, such as Correlated Topic Model (CTM) [11], Author-Topic Model [111], Dynamic Topic Model [12], and Relational Topic Models (RTM) [23] etc. Besides "topic" modeling, it has also inspired modeling discourse [109] and sentiment [73] in unsupervised or weak supervised way, which is the foundation work of Chapter 3. In particular, Ritter [109] proposed to model the discourse structure of conversation through Hidden Markov Model (HMM). But none of them consider the joint effect of

topic and hard to be extended to other tasks, such as text classification, which is an issue that Chapter 3 tackles.

2.2.2 Topic Modeling in Social Media Text

Despite of the huge success achieved by the springboard topic models (e.g., pLSA [41] and LDA [13]), and their extensions [8, 111], the applications of these models have been limited to formal and well-edited documents, such as news reports [8] and scientific articles [111], attributed to their reliance on document-level word collocations. When processing social media texts, such as the messages on microblogs, it is likely that the performance of these models will be inevitably compromised, due to the severe data sparsity issue and informal texts.

To deal with such an issue, many prior work focuses on how to enrich the context of short messages. To this end, biterm topic model (BTM) [148] extends a message into a biterm set with all combinations of any two distinct words appearing in the message. On the contrary, our model allows the richer context in a conversation to be exploited, where word collocation patterns can be captured beyond a short message. In addition, there are many methods employing some heuristic rules to aggregate short messages into long pseudo-documents, such as those based on authorship [43, 165], that is aggregating the posts of the same user. However, such heuristic aggregation is something unnatural in social media scenario. For example, one person might has multiple interest covering a wide range of topics, or express distinct writing style in different time. Ramage et al. [106] and Mehrotra et al. [84] propose the aggregation strategy based on hashtags. However, there are just a small portion of Twitter messages that containing a hashtag, their

performance is inevitably compromised for the messages with the topics irrelevant to any hashtag, and this is very common in the social media, cause topics keep changing rapidly.
In another line of the research, many previous efforts incorporate the *external* representations, such as word embeddings and knowledge base. In particular, Nguyen et al. [91], Shi et al. [113] and Li et al. [68] propose to incorporate word embedding pre-trained on large-scale high-quality resources into conventional Dirichlet multinomial topic model and jointly model. For example, Latent Feature Dirichlet Mixture Model (LF-DMM) use the assumption of short text topic model DMM, all the words in a document share the same topic. Based on which, LF-DMM use a Bernoulli distribution as a latent factor to determine whether the Dirichlet moltinomial or latent feature (i.e., word embedding) component will be used to generate each word in document. Song et al. [115], Yang et al. [149], and He et al. [46] propose to use knowledge base to improve the performance of topic modeling, for example joint model the topic and entity in knowledge graph. Such kind of external knowledge based topic modeling require large-scale high-quality external resources, which is domain specific. Since the content of social media is keeping changing, innovative knowledge emerging every day, it is hard to maintain an up-to-date external knowledge resources.

2.2.3 Neural Topic Model

Recently, there are several attempts to attack the inference problem of topic model based on neural networks and neural variational inference. Miao et al. [88] introduced Neural Variational Document Model (NVDM) based on variational auto-

encoder (VAE) [62], which mapping the bag-of-words to a latent Gaussian distribution over topics. The topic-word distribution in NVDM is implicitly attained by averaging over the logit space of reconstruction network. Miao et al. [87] extended NVDM by proposing a family of priors: Gaussian Softmax, Gaussian Stick-Breaking and Recurrent Stick-Breaking constructions to parameterize topic distributions. Srivastava et al. [118] also proposed VAE-style neural variational inference method, called Autoencoding Variational Inference For Topic Models (AVITM). AVITM simulates the original LDA formulation by approximating the Dirichlet prior using a Gaussian and explicitly captures the topic assignments.

Before introducing the neural topic model, we would like the revisit the optimization objective and variational inference method of LDA. The basic idea of variational inference is to obtain an adjustable lower bound on the log likelihood. By introducing a variabional distribution $q(\theta, \boldsymbol{z})$, we can bound the log likelihood of a document $\boldsymbol{w}$ by using Jensen's inequality. We have:

$$\begin{aligned}
\log p(\boldsymbol{w}|\alpha,\beta) =& \log \int \sum_{\boldsymbol{z}} p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha,\beta) d\theta \\
=& \log \int \sum_{\boldsymbol{z}} \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha,\beta) q(\theta,\boldsymbol{z})}{q(\theta,\boldsymbol{z})} d\theta \\
\geq& \int \sum_{\boldsymbol{z}} q(\theta,\boldsymbol{z}) \log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha,\beta) d\theta \\
& - \int \sum_{\boldsymbol{z}} q(\theta,\boldsymbol{z}) \log q(\theta,\boldsymbol{z}) d\theta \\
=& \mathbb{E}_q[\log p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha,\beta)] - \mathbb{E}_q[\log q(\theta,\boldsymbol{z})].
\end{aligned} \tag{2.2}$$

We now have a lower bound on the log likelihood, which is called *evidence lower bound* (ELBO), denoted as L_{ELBO}. It can be easily verified that the diffenece of $\log p(\boldsymbol{w}|\alpha,\beta)$ of the

Equation 2.2 and L_{ELBO} is the KL divergence between the variational posterior distribution and the true posterior distribution. Therefore, the optimization problem is to maximize:

$$L_{ELBO} = \log p(\boldsymbol{w}|\alpha, \beta) - D_{KL}[q(\theta, \boldsymbol{z})||p(\theta, \boldsymbol{z}|\boldsymbol{w}, \alpha, \beta)] \quad (2.3)$$

Since the exact inference is intractable in LDA, a popular approximation for the efficient parameter inference for LDA is mean field variational inference, which breaks the coupling between θ and $\boldsymbol{z}$. This results an approximate variational posterior $q(\theta, \boldsymbol{z}) = q(\theta) \prod_{N_d} q(z_{d,n})$. The detail of mean field variational inference for LDA can be found in Blei's paper [9].
In neural topic model (NTM), with the help of neural variational inference, we can formulate topic model as variational auto-encoder framework, and directly solve the optimization problem of L_{ELBO} in topic model. In particular, we assume the latent variables in topic model as $\boldsymbol{z}$, and the prior over the latent variables is an isotropic multivariate Gaussian distribution $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbf{I})$. The variational approximate posterior is a multivariate Gaussian with a diagonal covariance matrix:

$$\log q(\boldsymbol{z}|\boldsymbol{w}) = \log \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) \quad (2.4)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the mean and variance of the posterior distribution. Here we can rewire the ELBO as:

$$L_{ELBO} = 0.5 \sum_{k=1}^{K} (1 + \log(\sigma_k^2) - \mu_k^2 - \sigma_k^2) + \frac{1}{N_d} \sum_{n=1}^{N_d} \log p(w_{d,n}|\boldsymbol{z}_d) \quad (2.5)$$

where K is number of topics, $\boldsymbol{z}_d = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The latent variables in NTM, such as $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ can be inferred via an encoding network and decoder network for $p(w_{d,n}|\boldsymbol{z}_d)$ part.

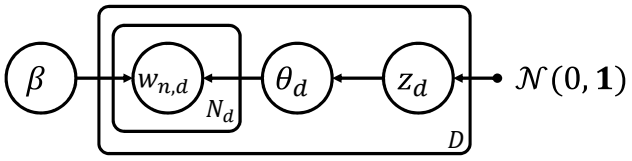


Figure 2.4: A graphical model representation of NTM.

NTM have a very similar generative process to LDA, as following:
For each document d:

- Draw latent variable $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
- $\theta = \text{softmax}(f_\theta(\mathbf{z}))$
- For each word $w_{d,n}$ in d:
  - $\beta = \text{softmax}(f_\phi(\theta))$
  - Draw word $w_{d,n} \sim Multinomial(\beta)$

NTM bring severals benefits when comparing with conventional topic models. (1) From the overview of graphical model representation of NTM, as shown in Figure. 2.4, NTM belongs to the probabilistic model, which enable the intractability of latent topic just like LDA. (2) The optimization of NTM can be solve by back-propagation, without introducing mean field assumption, which is easy to implement and more accurate than inference for conventional topic models. (3) NTM is neural network based model, enabling flexibility and extensible to other neural network structure. We extend the idea of NTM in Chapter 3, and use it to joint model topics and discourse for social media conversation.

2.3 Discourse Analysis

Discourse is the language signal reflecting semantic relations of the textual units and the architecture of dialog structure. We

review the prior research on conventional discourse analysis in Section 2.3.1 and discourse roles in conversation structure in Section 2.3.2.

2.3.1 Traditional View of Discourse

People have long realized that a coherent document, which gives readers continuity of senses [29] with the absence of non-sequiturs and gaps, is not simply a collection of independent sentences. The study of discourse can even be traced back to ancient Greece [6]. Hovy and Marier [44] depict the modern concept of discourse as the a structured collection of clauses, act as the connections between text units.

Rhetorical structure theory (RST) [81] is one of the most influential discourse theories, providing a systematic way for analyzing the natural text. According to its assumption, a coherent document can be represented by a hierarchical structure, consisted of different levels of units (e.g., relations, schemas, schema applications and structure). In particular, the most fundamental structural pattern defined by RST is the relations of the adjacent two text spans. Relations in RST, such as "backgound", "evidence", and "elaboration", is the specific asymmetric roles for one text unit to another.

Based on RST, early work employs manually pre-defined rules for automatic discourse analysis [83, 127]. With the appearing of large-scale discourse corpus, e.g., RST corpus [19], Graph Bank corpus [144], and Penn Discourse Treebank (PDTB) [101], people began to exploit supervised learning and data-driven based methods for discourse prediction or parsing [34, 57, 75, 116] and representative learning [50, 69].

2.3.2 Discourse Role in Conversation

With the development of Internet, our way of communication has been revolutionized by online social media platforms, e.g., discussion form, microblog. The flourish of social media bring a constant flood of information exchange, leads to a huge quantity of daily conversations among them. There is an increasing demand for automatically analysis the social media conversation. Although RST has been proved to be useful in analysis the formal and well-edited documents, discourse parsing on conversations is still a challenging problem [97], due to the complex structure and informal language. Previous research efforts mainly focus on the detection of dialogue acts (DA), which is defined in [124] as the shallow discourse role that captures illocutionary meanings of an utterance, e.g., "statement", "question", "agreement", etc. Automatic dialogue act taggers have been conventionally trained in a supervised way with predefined tag inventories and annotated data [25, 51, 123, 124].
However, the definition of DA is generally domain-specific and manually crafted by experts. The data annotation process is slow and expensive so that results in the limitation of available data for training DA classifiers. These issues become increasingly severe with the arrival of the Internet era where new domains of conversations and new types of dialogue act tags are boomed [56, 109]. For this reason, researchers have proposed unsupervised or weakly supervised dialogue act taggers that identify the discourse roles based on probabilistic graphical models [26, 51, 56, 109, 164].
For example, Ritter et al [109] is the first to study the unsupervised modeling of discourse in twitter conversation. They propose a probabilistic graphical model that model the topic

and discourse jointly. More specifically, they use hidden Markov model (HMM) to model the discourse sequence in utterance-level, model topic via LDA in conversation level. An latent factor with Bernoulli distribution is to control the current word generated from topic or discourse. Similar to the word cluster discovered by topic model, their model can also capture the indicative words cluster of discourse.
Although such latent discourse variables have been studied in previous work [51, 56, 109, 164], none of them explores the effects of latent discourse on the identification of conversation topic, which is a gap our work in Chapter 3 fills in.

2.4 Social Media Text Understanding

It has been long pointed out that a machine that can understand natural language, such as answers a question, executes commands, and accepts interactive information, is the core symbol of artificial intelligence (AI) [142]. Natural language understanding (NLU) is an important sub-field of natural language processing, with a wide range applications, such as question answering, text categorization, automated reasoning. Recently, with the development of social media, understanding the explosive social media text has received lots of research and commercial interests. Herein, we will introduce two applications of social media text understand, short text classification in Section 2.4.1 and argumentation mining in Section 2.4.2.

2.4.1 Short Text Classification

Social media such as Twitter and Weibo, have word limitation for each message. For example in Twitter, users are restricted to

post each tweet within limited 140 characters length. However the most common length of twitter is less than 30 characters, because most of users type their message through mobile phone. As a result, such short messages have become increasingly important in social media applications and there is a pressing need for analyzing and processing such short text. Among those techniques, text classification is a critical and fundamental one proven to be useful in various downstream applications, such as text summarization [45], recommendation [161], and sentiment analysis [24].

Although many classification models like support vector machines (SVMs) [136] and recently popular neural networks [58, 60, 146] have demonstrated their success in processing formal and well-edited texts, such as news articles [162], their performance is inevitably compromised when directly applied to short and informal online texts. This inferior performance is attributed to the severe data sparsity nature of short texts, which results in the noisy and limited features available for classifiers [99].

Most previous work focuses on alleviating the severe sparsity issues in short texts [148]. Some previous efforts encode knowledge from external resource [54, 78, 79, 133]. For example in [133], an external concept knowledge base trained from a large-scale of corpus, is utilized to obtain the relevant concepts of the short text message. For some specific classification tasks, such as sentiment analysis, manually-crafted features are designed to fit the target task [52, 94], which requires feature engineering process and thus hard to ensures its general applicability to diverse classification scenarios. There also exists work using representation learned from the internal text

structure (e.g., topic modeling). For example, pre-trained topic mixtures are leveraged as part of features fro training the short text classifier [24, 99, 107].

Recent research effort has focused on exploiting word embeddings or deep models for short text classification, due to the success of neural networks in many NLP tasks, such as semantic parsing and sequence labeling [17, 63, 148]. For example, Lee et al. [66] incorporate the preceding short texts based on recurrent neural network and convolutional neural network for classification. dos Santos et al. [31] exploit both character-level and sentence-level information. However, their work ignore the latent topics inherent from short texts, which could enrich the implicit representation of short text, and our work in Chapter 5 try to fill this gap by incorporating corpus-level implicit representation into a neural network framework for short text classification.

2.4.2 Argumentation Mining

Computational argumentation mining is a fast developing subfield in neural language understanding. Early work mainly focus on argumentation extraction, e.g., extract argumentation from legal text [89] and news [95], detecting argumentation structure, e.g., premise and conclusion [95, 120]. With the popularity of social media, online forum, such as idebate[3] and changemyview[4] provide a convenient platform for people to engage in argumentation. Researches have been paying increasing attention to analysis the argumentation in the online forum, for example, identification of convincing arguments [37, 137] and

[3] idebate.org

[4] reddit.com/cmv

viewpoints [39, 55] from social media discussions [126]. In this line, many existing studies focus on crafting hand-made features [126, 137], such as wordings and topic strengths [134, 158], semantic and syntactic rules [40, 98], participants' personality [131], argument interactions and structure [92], and so forth. These methods, however, require labor-intensive feature engineering process, and hence have limited generalization abilities to new domains. Recently, there have been some neural frameworks proposed to model the argument interactions and persuasiveness [39, 49, 55]. However, there is very few work that study the challenging problem, what changes the opinion holder's mind and how it happens during the argumentative conversation, which our work in Chapter 4 focuses on.

□ End of chapter.

Chapter 3

Joint Modeling of Topics and Discourse in Microblog Conversations

This chapter presents an unsupervised framework for jointly modeling topic content and discourse behaviors in microblog conversations. Concretely, we propose a neural model to discover word clusters indicating what a conversation concerns (i.e., topics) and those reflecting how participants voice their opinions (i.e., discourse). A key finding is that joint modeling of topics and discourse can yield both coherent topics and meaningful discourse behaviors. Also, out topic and discourse representations can benefit the classification of microblog messages, especially when they are jointly trained with the classifier. The main points of this chapter are as follows. (1) It proposes an unsupervised neural network built upon topic, enabling the joint exploration of word clusters to represent topic and discourse in microblog conversations. (2) It conducts an extensive empirical study on two large-scale Twitter datasets. (3) It presents how to extend the proposed network by easily combining with existing neural models for end-to-end training, such as CNN.

3.1 Introduction

The last decade has witnessed the revolution of communication, where the "kitchen table conversations" have been expanded to public discussions on online platforms. As a consequence, in our daily life, the exposure to new information and the exchange of personal opinions have been mediated through microblogs, one popular online platform genre [7]. The flourish of microblogs has also led to the sheer quantity of user-created conversations emerging every day, exposing individuals to superfluous information. Facing such unprecedented number of conversations relative to limited attention of individuals, how shall we automatically extract the critical points and make sense of these microblog conversations?

Towards key focus understanding of a conversation, previous work has shown the benefits of discourse structure [70, 72, 102], which shapes how messages interact with each other forming the discussion flow and can usefully reflect salient topics raised in the discussion process. After all, the topical content of a message naturally occurs in context of the conversation discourse and hence should not be modeled in isolation. On the other way around, the extracted topics can reveal the purpose of participants and further facilitate the understanding of their discourse behaviors [102].

To illustrate how the topics and discourse interplay in a conversation, Figure 3.1 displays a snippet of Twitter conversation. As can be seen, the content words reflecting the discussion topics (such as "*supreme court*" and "*gun rights*") appear in context of the discourse flow, where participants carry the conversation forward via making a statement, giving a comments, asking

...
M_1 [*Statement*]: Just watched **HRC** openly endorse a **gun-control measure** which will fail in front of the **Supreme Court**. This is a train wreck.
M_2 [*Comment*]: People said the same thing about **Obama**, and nothing took place. **Gun laws** just aren't being enforced like they should be. :/
M_3 [*Question*]: Okay, hold up. What do you think I'm referencing here? It's not what you're talking about.
M_4 [*Agreement*]: Thought it was about **gun control**. I'm in agreement that **gun rights** shouldn't be stripped.
...

Figure 3.1: A Twitter conversation snippet about the gun control issue in U.S. **Topic words** reflecting the conversation focus are in boldface. The *italic* words in [] are our interpretations of the messages' discourse roles.

a question, and so forth. Motivated by such observation, we assume that *a microblog conversation can be decomposed into two crucially different components: one for topical content and the other for discourse behaviors*. Here, the topic components indicate what a conversation is centered around and reflect the important discussion points put forward in the conversation process. The discourse components signal the **discourse roles** of messages, such as making a statement, asking a question, and other dialogue acts [56, 109], which further shape the discourse structure of a conversation.[1] To distinguish the above two components, we examine the conversation contexts and identify two types of words: **topic words**, indicating what a conversation focuses on, and **discourse words**, reflecting how the opinion is voiced in each message. For example, in Figure 3.1, the topic words "*gun*" and "*control*" indicate the conversation topic while the discourse word "*what*" and "*?*" signal the question in M_3.

Concretely, we propose a neural framework with the topic model fashion, enabling the joint exploration of word clusters to represent topic and discourse in microblog conversations. Different from the prior models trained on annotated data [70, 102], our model is fully unsupervised, not dependent on annotations for either topics or discourse, which ensures its immediate applicability in any domain or language. Moreover, taking advantages of the recent advances in neural topic models [87, 119], we are able to approximate Bayesian variational inference without requiring model-specific derivations, while most existing models [4, 56, 70, 72, 109] require the expertise involved to customize

[1] In this thesis, the discourse role refers to a certain type of dialogue act (e.g., *statement* or *question*) for each message. And the discourse structure refers to some combination of discourse roles in a conversation.

model inference algorithms. In addition, the neural nature of our model enables the end-to-end training of topic and discourse representations with other neural models for diverse tasks.
For model evaluation, we conduct an extensive empirical study on two large-scale Twitter datasets. The intrinsic results show that our model can produce latent topics and discourse roles with better interpretability than the state-of-the-art models from previous studies. The extrinsic evaluations on a tweet classification task exhibit our ability to capture useful representations for microblog messages. Particularly, our model enables the easy combination and the end-to-end training with other neural models, such as CNN, which is shown to perform better in classification than the pipeline approach without joint training.

3.2 Our Neural Model for Topics and Discourse in Conversations

This section introduces our neural model that jointly explores latent representations for topics and discourse in conversations.[2] We first present an overview of our model in Section 3.2.1, followed by the model generative process and inference procedure in Section 3.2.2 and 3.2.3, respectively.

3.2.1 Model Overview

In general, our model aims to learn coherent word clusters that reflect the latent topics and discourse roles embedded in the microblog conversations. To this end, we distinguish two latent components in the given collection: *topics* and *discourse*, each

[2]The code of our model will be released on Github after the anonymous review process.

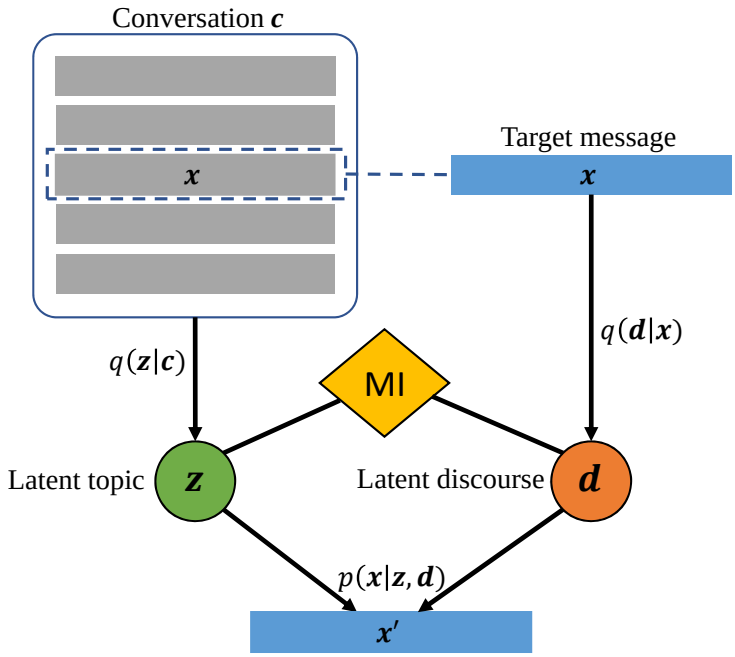


Figure 3.2: The architecture of our neural framework that jointly models latent topics and latent discourse.

represented by a certain type of word distribution (distributional word cluster). Specifically, at the corpus level, we assume there are K topics, represented by ϕ_k^T, $(k = 1, 2, \ldots, K)$, and D discourse roles, captured with ϕ_d^D, $(d = 1, 2, \ldots, D)$. ϕ^T and ϕ^D are all multinomial word distributions over the vocabulary size V. Inspired by the neural topic models in [87], our model encodes topic and discourse distributions (ϕ^T and ϕ^D) as latent variables in a neural network and learns the parameters via back propagation.

Before touching the details of our model, we first describe how we formulate the input. On microblogs, as a message might have multiple replies, messages in an entire conversation can be organized as a tree with replying relations [70, 72]. Though the recent progress in recursive models allows the representation learning from the tree-structured data, previous studies have pointed out that, in practice, sequence models serve

as a more simple yet robust alternative [71]. In this work, we follow the common practice in most conversation modeling research [56, 109, 164] to take a conversation as a sequence of turns. To this end, each conversation tree is flattened into root-to-leaf paths. Each one of such paths is hence considered as a conversation instance, and a message on the path corresponds to a conversation turn [21, 53, 151].

The overall architecture of our model is shown in Figure 3.2. Formally, we formulate a conversation $\mathbf{c}$ as a sequence of messages $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{M_c})$, where M_c denotes the number of messages in $\mathbf{c}$. In the conversation, each message $\mathbf{x}$, named as the **target message**, is fed into our model sequentially. Here we process the target message $\mathbf{x}$ as the bag-of-words (BoW) term vector $\mathbf{x}_{BoW} \in \mathbb{R}^V$, following the bag-of-words assumption in most topic models [8, 87]. The conversation, $\mathbf{c}$, where the target message $\mathbf{x}$ is involved, is considered as the context of $\mathbf{x}$. It is also encoded in the BoW form (denoted as $\mathbf{c}_{BoW} \in \mathbb{R}^V$) and fed into our model. In doing so, we ensure the context of the target message is incorporated while learning its latent representations. Following the previous practice in neural topic models [87, 119], we employ the variational auto-encoder (VAE) [62] to resemble the data generative process via two steps. First, given the target message $\mathbf{x}$ and its conversation $\mathbf{c}$, our model converts them into two latent variables: topic variable $\mathbf{z}$ and discourse variable $\mathbf{d}$. Then, using the intermediate representations captured by $\mathbf{z}$ and $\mathbf{d}$, we reconstruct the target message, $\mathbf{x}'$.

3.2.2 Generative Process

In this section, we first describe the two latent variables in our model: the topic variable $\mathbf{z}$ and the discourse variable $\mathbf{d}$. Then,

we present our data generative process from the latent variables.

Latent Topics. For latent topic learning, we examine the main discussion points in the context of a conversation. Our assumption is that messages in the same conversation tend to focus on similar topics [72, 156]. Concretely, we define the latent topic variable $\mathbf{z} \in \mathbb{R}^K$ at the *conversation* level and generate the topic mixture of $\mathbf{c}$, denoted as a K-dimentional distribution θ, via a softmax construction conditioned on $\mathbf{z}$ [87].

Latent Discourse. For modeling the discourse structure of conversations, we capture the *message*-level discourse roles reflecting the dialogue acts of each message, as is done in [109]. Concretely, given the target message $\mathbf{x}$, we employ a D-dimensional one-hot vector to represent the latent discourse variable $\mathbf{d}$, where the high bit indicates the index of discourse that can best express $\mathbf{x}$'s discourse role. In the generative process, $\mathbf{d}$ is drawn from a multinomial distribution with parameters estimated from the input data.

Data Generative Process As mentioned previously, our entire framework is based on VAE, which consists of an encoder and a decoder. The encoder maps a given input into latent topic and discourse representations and the decoder reconstructs the original input from the latent representations. In the following, we first describe the decoder followed by the encoder.
In general, our *decoder* is learned to reconstruct the words in the target message $\mathbf{x}$ (in the BoW form) from the latent topic $\mathbf{z}$ and latent discourse $\mathbf{d}$. We show the generative story that reflects the reconstruction process below:

- Draw the latent topic $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
- $\mathbf{c}$'s topic mixture $\theta = \text{softmax}(f_\theta(\mathbf{z}))$

- Draw the latent discourse $\mathbf{d} \sim Multi(\boldsymbol{\pi})$
- For the n-th word in $\mathbf{x}$
  - $\beta_n = \text{softmax}(f_{\phi^T}(\theta) + f_{\phi^D}(\mathbf{d}))$
  - Draw the word $w_n \sim Multi(\beta_n)$

where $f_*(\cdot)$ is a neural perceptron, with a linear transformation of inputs activated by a non-linear transformation. Here we use rectified linear units (ReLUs) [90] as the activate functions. In particular, the weight matrix of $f_{\phi^T}(\cdot)$ (after the softmax normalization) is considered as the topic-word distributions ϕ^T. The discourse-word distributions ϕ^D are similarly obtained from $f_{\phi^D}(\cdot)$.

For the *encoder*, we learn the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\pi}$ from the input $\mathbf{x}_{BoW}$ and $\mathbf{c}_{BoW}$ (the BoW form of the target message and its conversation), following the formula below:

$$\begin{aligned} \boldsymbol{\mu} = f_\mu(f_e(\mathbf{c}_{BoW})),\ \log \boldsymbol{\sigma} = f_\sigma(f_e(\mathbf{c}_{BoW})) \\ \boldsymbol{\pi} = \text{softmax}(f_\pi(\mathbf{x}_{BoW})) \end{aligned} \tag{3.1}$$

3.2.3 Model Inference

For the objective function of our entire framework, we take three aspects into account: the learning of latent topics and discourse, the reconstruction of the target messages, and the separation of topic-associated words and discourse-related words.

Learning Latent Topics and Discourse. For learning the latent topics/discourse in our model, we employ the variational inference [10] to approximate posterior distribution over the latent topic $\mathbf{z}$ and the latent discourse $\mathbf{d}$ given all the training data. To this end, we maximize the variational lower bound $\mathcal{L}_z$ for $\mathbf{z}$ and $\mathcal{L}_d$ for $\mathbf{d}$, each defined as following:

$$\begin{aligned}\mathcal{L}_z &= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{c})}[p(\mathbf{c}\,|\,\mathbf{z})] - D_{KL}(q(\mathbf{z}\,|\,\mathbf{c})\,||\,p(\mathbf{z})) \\ \mathcal{L}_d &= \mathbb{E}_{q(\mathbf{d}\,|\,\mathbf{x})}[p(\mathbf{x}\,|\,\mathbf{d})] - D_{KL}(q(\mathbf{d}\,|\,\mathbf{x})\,||\,p(\mathbf{d}))\end{aligned} \tag{3.2}$$

$q(\mathbf{z}\,|\,\mathbf{c})$ and $q(\mathbf{d}\,|\,\mathbf{x})$ are approximated posterior probabilities describing how the latent topic $\mathbf{z}$ and the latent discourse $\mathbf{d}$ are generated from the data. $p(\mathbf{c}\,|\,\mathbf{z})$ and $p(\mathbf{x}\,|\,\mathbf{d})$ represent the corpus likelihoods conditioned on the latent variables. $p(\mathbf{z})$ follows the standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\mathbf{d})$ is the uniform distribution $Unif(0, 1)$. D_{KL} refers to the Kullback-Leibler divergence that ensures the approximated posteriors to be close to the true ones. Due to the space limitation, we leave out the derivation details and refer the readers to [87].

Reconstructing target messages. From the latent variables $\mathbf{z}$ and $\mathbf{d}$, the goal of our model is to reconstruct the target message $\mathbf{x}$. The corresponding learning objective is to maximize $\mathcal{L}_x$ defined as:

$$\mathcal{L}_x = \mathbb{E}_{q(z\,|\,\mathbf{x})q(d\,|\,\mathbf{c})}[\log p(\mathbf{x}\,|\,\mathbf{z}, \mathbf{d})] \tag{3.3}$$

Here we design $\mathcal{L}_x$ to ensure that the learned latent topics and discourse can reconstruct $\mathbf{x}$.

Distinguishing Topics and Discourse. Our model aims to distinguish word distributions for topics (ϕ^T) and discourse (ϕ^D), which enables topics and discourse to capture different information in conversations. Concretely, we employ the mutual information, given below, to measure the mutual dependency between the latent topics $\mathbf{z}$ and the latent discourse $\mathbf{d}$. [3]

[3]The distributions in Eq. 3.4 are all conditional probability distributions given the target message $\mathbf{x}$ and its conversation $\mathbf{c}$. We omit the conditions for simplicity.

$$\mathbb{E}_{q(\mathbf{z})q(\mathbf{d})}[\log \frac{p(\mathbf{z},\mathbf{d})}{p(\mathbf{z})p(\mathbf{d})}] \tag{3.4}$$

Eq. 3.4 can be further derived as the Kullback-Leibler divergence of the conditional distribution, $p(\mathbf{d} \mid \mathbf{z})$, and marginal distribution, $p(\mathbf{d})$. The derived formula, defined as the mutual information loss ($\mathcal{L}_{MI}$) and shown in Eq. 3.5, is used to map $\mathbf{z}$ and $\mathbf{d}$ into the separated semantic space.

$$\mathcal{L}_{MI} = \mathbb{E}_{q(\mathbf{z})}[D_{KL}(p(\mathbf{d} \mid \mathbf{z})||p(\mathbf{d}))] \tag{3.5}$$

We can hence minimize $\mathcal{L}_{MI}$ for guiding our model to separate word distributions that represent topics and discourse.

The Final Objective. To capture the joint effects of the learning objectives described above ($\mathcal{L}_z$, $\mathcal{L}_d$, $\mathcal{L}_x$, and $\mathcal{L}_{MI}$), we design the final objective function for our entire framework as following:

$$\mathcal{L} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda\mathcal{L}_{MI} \tag{3.6}$$

where the hyperparameter λ is the trade-off parameter for balancing between the MI loss ($\mathcal{L}_{MI}$) and the other learning objectives. By maximizing the final objective $\mathcal{L}$ via back propagation, the word distributions of topics and discourse can be jointly learned from microblog conversations.[4]

[4]To smooth the gradients in implementation, for $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, we apply the reparameterization on $\mathbf{z}$ [62, 108], and for $\mathbf{d} \sim Multi(\boldsymbol{\pi})$, we adopt the Gumbel-Softmax trick [48, 80].

3.3 Experimental Setup

In this section, we describe how we set up the experiment for model evaluation.

3.3.1 Data Collection

For our experiments, we collected two microblog conversation datasets from Twitter. One is released by the TREC 2011 microblog track (henceforth **TREC**), containing conversations concerning a wide rage of topics.[5] The other is crawled from January to June 2016 with Twitter streaming API[6] (henceforth **TWT16**, short for Twitter 2016), following the way of building TREC dataset. During this period, there are a large volume of discussions centered around U.S. presidential election. In addition, for both datasets, we apply Twitter search API[7] to retrieve the missing tweets in the conversation history, as the Twitter streaming API (used to collect both datasets) only returns sampled tweets from the entire pool.
The statistics of the two experiment datasets are shown in Table 3.1. For model training and evaluation, we randomly sampled 80%, 10%, and 10% of the data to form the training, development, and test set, respectively.

3.3.2 Data Preprocessing

We preprocessed the data with the following steps. First, non-English tweets were filtered out. Then, hashtags, men-

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

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

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

Datasets	# of convs	Avg msgs per conv	Avg words per msg	|Vocab|
TREC	116,612	3.95	11.38	9,463
TWT16	29,502	8.67	14.70	7,544

Table 3.1: Statistics of the two datasets containing Twitter conversations.

tions (@username), and links were replaced with generic tags "HASH", "MENT", and "URL", respectively. Next, the natural languge toolkit (NLTK) was applied for tweet tokenization.[8] After that, all letters were normalized to lower cases. Finally, words occurred less than 20 times were filtered out from the data.

3.3.3 Parameter Setting

To ensure comparable results with [72] (the prior work focusing on the same task as ours), in the topic coherence evaluation, we follow their setting to report the results under two sets of K (the number of topics): $K = 50$ and $K = 100$, and with the number of discourse roles (D) set to 10. The parameter analysis of K and D will be further presented in Section 3.4.5. For all the other hyper-parameters, we tuned them on development set by grid search. The trade-off parameter λ (defined in Eq. 3.6), balancing the MI loss and the other objective functions, is set to 0.01. In model training, we use Adam optimizer [61] and run 100 epochs with early stop strategy adopted.

3.3.4 Baselines

In topic modeling experiments, we consider the five topic model baselines treating each tweet as a document: LDA [8],

[8] `https://www.nltk.org/`

BTM [148], LF-LDA, LF-DMM [91], and NTM [87]. In particular, BTM and LF-DMM are the state-of-the-art topic models for short texts. BTM explores the topics of all word pairs (biterms) in each message to alleviate data sparsity in short texts. LF-DMM incorporates word embeddings pre-trained on external data to expand semantic meanings of words, so does LF-LDA. In Nguyen et al. (2015)[91], LF-DMM, based on one-topic-per-document Dirichlet Multinomial Mixture (DMM) [93], was reported to perform better than LF-LDA, based on LDA. For LF-LDA and LF-DMM, we use GloVe Twitter embeddings [96] as the pre-trained word embeddings.[9]

For the discourse modeling experiments, we compare our results with LAED [164], a VAE-based representation learning model for conversation discourse. In addition, for both topic and discourse evaluation, we compare with Li et al.[72], a recently proposed model for microblog conversations, where topics and discourse are jointly explored with a *non-neural* framework. Besides the existing models from previous studies, we also compare with the variants of our model that only models topics (henceforth TOPIC ONLY) or discourse (henceforth DISC ONLY). Our joint model of topics and discourse is referred to as TOPIC+DISC.

In the preprocessing process for the baselines, we removed stop words and punctuation for topic models unable to learn discourse representations following the common practice in previous work [148, 87]. For the other models, stop words and punctuation were retained in the vocabulary considering their usefulness as discourse indicators [72].

[9] `https://nlp.stanford.edu/projects/glove/`

3.4 Experimental Results

In this section, we first report the topic coherence results in Section 3.4.1, followed by a discussion in Section 3.4.2 comparing the latent discourse roles discovered by our model with the manually annotated dialogue acts. Then, we study whether we can capture useful representations for microblog messages in a tweet classification task (in Section 3.4.3). A qualitative analysis, showing some example topics and discourse roles, is further provided in Section 3.4.4. Finally, in Section 3.4.5, we provide more discussions on our model.

3.4.1 Topic Coherence

For the topic coherence, we adopt the C_v scores measured via the open-source Palmetto toolkit as our evaluation metric.[10] C_v scores assume that the top N words in a coherent topics (ranked by likelihood) tend to co-occur in the same document and have shown comparable evaluation results to human judgments [110]. Table 3.2 shows the average C_v scores over the produced topics given $N = 5$ and $N = 10$. The values range from 0.0 to 1.0 and higher scores indicate better topic coherence. We can observe that:

• ***Models assuming a single topic for each message do not work well.*** It has long been pointed out that the one-topic-per-message assumption (each message contains only one topic) helps topic models alleviate the data sparsity issue in short texts on microblogs [72, 91, 105, 165]. However, we observe contradictory results since both LF-DMM and [72], following this assumption, achieve generally worse performance than the

[10] `https://github.com/dice-group/Palmetto`

Models	$K = 50$		$K = 100$	
	TREC	TWT16	TREC	TWT16
Baselines				
LDA	0.467	0.454	0.467	0.454
BTM	0.460	0.461	0.466	0.463
LF-DMM	0.456	0.448	0.463	0.466
LF-LDA	0.470	0.456	0.467	0.453
NTM	0.478	0.479	0.482	0.443
Li et al. [72]	0.463	0.433	0.464	0.435
Our models				
TOPIC ONLY	0.478	0.482	0.481	0.471
TOPIC+DISC	**0.485**	**0.487**	**0.496**	**0.480**

Table 3.2: C_v coherence scores for latent topics produced by different models. The best result in each column is highlighted in **bold**. Our joint model TOPIC+DISC achieves significantly better coherence scores than all the baselines ($p < 0.01$, paired test).

other models. This might be attributed to the large-scale data used in our experiments (each dataset has over 250K messages as shown in Table 3.1), which potentially provide richer word co-occurrence patterns and thus partially alleviate the data sparsity issue.

• ***Pre-trained word embeddings do not bring benefits.*** Comparing LF-LDA with LDA, we found that they give similar coherence scores. This shows that with sufficiently large training data, with or without using the pre-trained word embeddings do not make any difference in the topic coherence results.

• ***Neural models perform better than non-neural baselines.*** When comparing the results of neural models (NTM and our models) with the other baselines, we find the former yield topics with better coherence scores in most cases.

• ***Modeling topics in conversations is effective.*** Among

Models	Purity	Homogeneity	VI
Baselines			
LAED	0.505	0.022	6.418
Li et al. [72]	0.511	0.096	5.540
Our models			
DISC ONLY	0.510	0.112	5.532
TOPIC+DISC	**0.521**	**0.142**	**5.097**

Table 3.3: The purity, homogeneity, and variation of information (VI) scores for the latent discourse roles measured against the human-annotated dialogue acts. For purity and homogeneity, higher scores indicate better performance, while for VI scores, lower is better. In each column, the best results are in **boldface**. Our joint model TOPIC+DISC significantly outperforms all the baselines ($p < 0.01$, paired t-test).

neural models, we found our models outperform NTM (without exploiting conversation contexts). This shows that the conversations provide useful context and enables more coherent topics to be extracted from the entire conversation thread instead of a single short message.

• ***Modeling topics together with discourse helps produce more coherent topics.*** We can observe better results with the joint model TOPIC+DISC in comparison with the variant considering topics only. This shows that TOPIC+DISC, via the joint modeling of topic- and discourse-word distributions (reflecting non-topic information), can better separate topical words from non-topical ones, hence resulting in more coherent topics.

3.4.2 Discourse Interpretability

In this section, we evaluate whether our model can discover meaningful discourse representations. To this end, we train

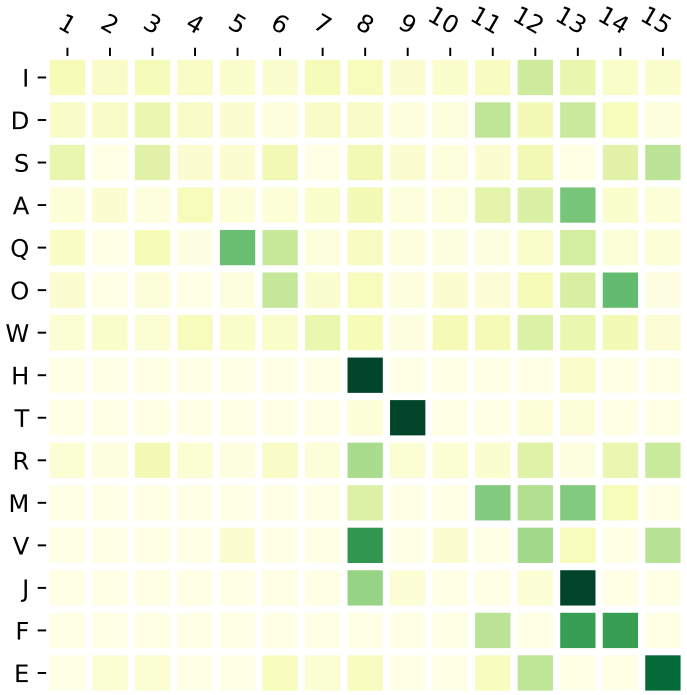
the comparison models for discourse modeling on the TREC dataset and test the learned latent discourse on a benchmark dataset released by [21]. The benchmark dataset consists of 2,217 microblog messages forming 505 conversations collected from Mastodon[11], a microblog platform exhibiting Twitter-like user behaviors [21]. For each message, there is a human-assigned discourse label, selected from one of the 15 dialogue acts, such as *question*, *answer*, *disagreement*, etc.

For discourse evaluation, we measure whether the model-produced discourse assignments are consistent with the human-annotated dialogue acts. Hence following Zhao et al. (2018)[164], we assume that an interpretable latent discourse role should cluster messages labeled with the same dialogue act. Therefore, we adopt purity [82], homogeneity [112], and variation of information (VI) [35, 86] as our automatic evaluation metrics. Here, we set $D = 15$ to ensure the number of latent discourse roles to be the same as the number of manually-labeled dialogue acts. Table 3.3 shows the comparison results of the average scores over the 15 latent discourse roles. Higher values indicate better performance for purity and homogeneity, while for VI, lower is better.

It can be observed that our models exhibit generally better performance, showing the effectiveness of our framework in inducing interpretable discourse roles. Particularly, we observe the best results achieved by our joint model TOPIC+DISC, which is learned to distinguish topic- and discourse-words, important in recognizing indicative words to reflect latent discourse.

To further analyze the consistency of varying latent discourse roles (produced by our TOPIC+DISC model) with the human-

[11] `https://mastodon.social`



I: statement, D: disagreement, S: suggest, A: agreement, Q: yes/no question, O: wh*/open question, W: open+choice answer, H: initial greetings, T: thanking, R: request, M: sympathy, V: explicit performance, J: exclamation, F: acknowledge, and E: offer.

Figure 3.3: A heatmap showing the alignments of the latent discourse roles and human-annotated dialogue act labels. Each line visualizes the distribution of messages with the corresponding dialogue act label over varying discourse roles (indexed from 1 to 15), where darker colors indicate higher values.

labeled dialogue acts, Figure 3.3 displays a heatmap, where each line visualizes how the messages with a dialogue act distribute over varying discourse roles. It is seen that among all dialogue acts, our model discovers more interpretable latent discourse for "*greetings*", "*thanking*", "*exclamation*", and "*offer*", where most messages are clustered into one or two dominant discourse roles. It may be because these dialogue acts can be relatively easier to detect based on their associated indicative words, such as the word "*thanks*" for "*thanking*", and the word "*wow*" for

"*exclamation*".

3.4.3 Message Representations

To further evaluate our ability to capture effective representations for microblog messages, we take tweet classification as an example and test the classification performance with the topic and discourse representations as features. Here the user-

Models	TREC		TWT16	
	Acc	Avg F1	Acc	Avg F1
Baselines				
BoW	0.120	0.026	0.132	0.030
LDA	0.128	0.041	0.146	0.046
BTM	0.123	0.035	0.167	0.054
LFDMM	0.158	0.072	0.162	0.052
NTM	0.138	0.042	0.186	0.068
Our model	**0.259**	**0.180**	**0.341**	**0.269**

Table 3.4: Evaluation of tweet classification results in accuracy (Acc) and average F1 (Avg F1). Representations learned by various models serve as the classification features. For our model, both the topic and discourse representations are fed into the classifier.

generated hashtags capturing the topics of online messages are used as the proxy class labels [70, 154]. We construct the classification dataset from TREC and TWT16 with the following steps. First, we removed the tweets without hashtags. Second, we ranked hashtags by their frequencies. Third, we manually removed the hashtags that are not topic-related (e.g. "*#fb*" for indicating the source of tweets from Facebook), and combined the hashtags referring to the same topic (e.g. "*#DonaldTrump*" and "*#Trump*"). Finally, we selected the top 50 frequent hashtags, and all tweets containing these hashtags

as our classification dataset. Here, we simply use the support vector machines (SVMs) as the classifier, since our focus is to compare the representations learned by various models.
Table 3.4 shows the classification results of accuracy and average F1 on the two datasets with the representations learned by various models serving as the classification features. We observe that our model outperforms other models with a large margin. The possible reasons are two folds. First, our model derives topics from conversation threads and thus potentially yields better message representations. Second, the discourse representations (only produced by our model) are indicative features for hashtags, because users will exhibit various discourse behaviors in discussing diverse topics (hashtags). For instance, we observe prominent "argument" discourse from tweets with "*#Trump*" and "*#Hillary*", attributed to the controversial opinions to the two candidates in the 2016 U.S. presidential election.

3.4.4 Example Topics and Discourse Roles

We have shown that jointly modeling of topics and discourse presents superior performance on quantitative measure. In this section, we qualitatively analyze the interpretability of our outputs via analyzing the word distributions of some example topics and discourse roles.

Example Topics. Table 3.5 lists the top 10 words of some example latent topics discovered by various models from the TWT16 dataset. According to the words shown, we can interpret the extracted topics as "gun control". We observe that LDA wrongly includes off-topic word "*flag*". From the outputs of BTM, LF-DMM, Li et al., 2018 [72], and our TOPIC ONLY

LDA	people trump police violence gun death protest guns flag shot
BTM	gun guns people police wrong right think law agree black
LF-DMM	gun police black said people guns killing ppl amendment laws
Li et al. [72]	wrong don trump gun understand laws agree guns doesn make
NTM	gun understand yes guns world dead real discrimination trump silence
TOPIC ONLY	shootings gun guns cops charges control mass commit know agreed
TOPIC+DISC	guns gun shootings chicago shooting cops firearm criminals commit laws

Table 3.5: Top 10 representative words of example latent topics discovered from the TWT16 dataset. We interpret the topics as "gun control" by the displayed words. Non-topic words are wave-underlined and in blue, while off-topic words are underlined and in red.

variant, though we do not find off-topic words, there are some non-topic words, such as "*said*" and "*understand*".[12] The output of our TOPIC+DISC model appears to be the most coherent, with words such as "*firearm*" and "*criminals*" included, which are clearly relevant to "gun control". Such results indicate the benefit of examining the conversation contexts and jointly exploring topics and discourse in them.

Example Discourse Roles. To qualitatively analyze whether our TOPIC+DISC model can discover interpretable discourse roles, we select the top 10 words from the distributions of some example discourse roles and list them in Table 3.6. It can be observed that there are some meaningful word clusters reflecting varying discourse roles found without any supervision. Interestingly, we observe that the latent discourse roles from TREC and TWT16, though learned separately, exhibit some notable overlap in their associated top 10 words, except for "argument", represented by very different words. The reason is that TWT16 contains a large volume of arguments centered around candidate Clinton and Trump, resulting in the frequent appearance of words like "*he*" and "*she*".

3.4.5 Further Discussions

In this section, we further present more discussions on our joint model: TOPIC+DISC .

Parameter Analysis. Here we study the two important hyper-parameters in our model, the number of topics (K) and the number of discourse roles (D). In Figure 3.4, we show the C_v

[12]Non-topic words do not clearly indicate the corresponding topic, while off-topic words are more likely to appear in other topics.

Discourse Roles	**TREC**	**TWT16**
Question	was what why is how that like ? ?? you	? why what MENT do does it the to did
Response	! love ha !! you saw lmao lol awesome !!!	doin uhhh ❗ awards yay joseph 😖 👋 muted
Agreement	okaay thankss wateva okayy txtd twitcam entertained havee goooood darlin	! you are agree re to they we with their
Quotation	& ' < > (... feat “ ” ")	» « (< ¦ ¦ < MENT .< ,- - ?< "
Statement	to will ! the be rt my in on and	will have if do be can want vote should ?
Argument	fuck damn rt lmfao hair girl thing lmao ass bitch	😂 he said him she her but wrong did never

Table 3.6: Top 10 representative words of example discourse roles learned from TREC and TWT16. The discourse roles of the word clusters are manually assigned according to their associated words.

topic coherence given varying K in (a) and the homogeneity measure given varying D in (b). As can be seen, the curves corresponding to the performance on topics and discourse are not monotonic. In particular, better topic coherence scores are achieved given relatively larger topic numbers for TREC with the best result observed at $K = 80$. On the contrary, the optimum topic number for TWT16 is $K = 20$, while increasing the number of topics results in worse C_v scores in general. This may be attributed to the relatively centralized topic concerning U.S. election in the TWT16 corpus. For discourse homogeneity, the best result is achieved given $D = 15$, with same the number of manually annotated dialogue acts in the benchmark.

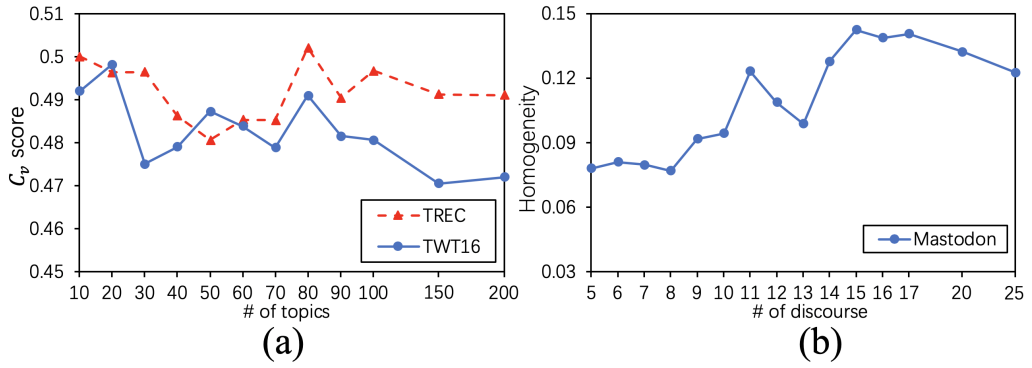


Figure 3.4: (a) The impact of topic numbers. The horizontal axis: the number of topics; The vertical axis: the C_v topic coherence. (b) The impact of discourse numbers. The horizontal axis: the number of discourse; The vertical axis: the homogeneity measure.

Case Study. To further understand why our model learns meaningful representations for topics and discourse, we present a case study based on the example conversation shown in Figure 3.1. Specifically, we visualize the topic words (with $p(w \mid \mathbf{z}) > p(w \mid \mathbf{d})$) in red and the rest words in blue to indicate discourse. Darker red indicates the higher topic likelihood ($p(w \mid \mathbf{z})$) while darker blue shows the higher discourse likelihood ($p(w \mid \mathbf{d})$). The results are shown in Figure 3.5. We can observe that topic and discourse words are well separated by our model, which explains why it can generate high-quality representations for both topics and discourse.

Model Extensibility. Recall that in the Introduction, we have mentioned that our neural-based model has an advantage to be easily combined with other neural network architectures and allows for the joint training of both models. Here we take message classification (with the setting in Section 3.4.3) as an example, and study whether joint training our model with convolutional neural network (CNN) [60], the widely-used

Figure 3.5: Visualization of the topic-discourse assignment of a twitter conversion from TWT16. The annotated blue words are pone to be discourse words, and the red are topic words. The shade is indicating the confidence of current assignment.

model on short text classification, can bring benefits to the classification performance. We set the embedding dimension to 200, with random initialization. The results are shown in Table 3.7, where we observe that joint training our model and the classifier can successfully boost the classification performance.

Error Analysis. We further analyze the errors in our outputs. For topics, taking a closer look at their word distributions, we found that our model sometimes mix sentiment words with topic words. For example, among the top 10 words of a topic "*win people illegal americans hate lt racism social tax wrong*", there are words "*hate*" and "*wrong*", expressing sentiment rather than conveying topic-related information. This is due to the prominent co-occurrences of topic words and sentiment words in

Models	TREC		TWT16	
	Acc	Avg F1	Acc	Avg F1
CNN only	0.199	0.167	0.334	0.311
Separate-Train	0.284	0.270	0.391	0.390
Joint-Train	**0.297**	**0.286**	**0.428**	**0.413**

Table 3.7: Accuracy (Acc) and average F1 (Avg F1) on tweet classification (hashtags as labels). CNN only: CNN without using our representations. Seperate-Train: CNN fed with our pre-trained representations. Joint-Train: Joint training CNN and our model.

our data, which results in the similar distributions for topics and sentiment. Future work could focus on the further separation of sentiment and topic words.

For discourse, we found that our model can induce some discourse roles beyond the 15 manually defined dialogue acts in the Mastodon dataset [21]. For example, as shown in Table 3.6, our model discover the "*quotation*" discourse from both TREC and TWT16, which is however not defined in the Mastodon dataset. This perhaps should not be considered as an error. We argue that it is not sensible to pre-define a fixed set of dialogue acts for diverse microblog conversations due to the rapid change and a wide variety of user behaviors in social media. Therefore, future work should involve a better alternative to evaluate the latent discourse without relying on manually defined dialogue acts. We also notice that our model sometimes fails to identify discourse behaviors requiring more in-depth semantic understanding, such as sarcasm, irony, and humor. This is because our model detects latent discourse purely based on the observed words, while the detection of sarcasm, irony, or humor requires deeper language understanding, which is beyond the capacity of our model.

3.5 Summary

We have presented a neural framework that jointly explores topic and discourse from microblog conversations. Our model, in an unsupervised manner, examines the conversation contexts and discovers word distributions that reflect latent topics and discourse roles. Results from extensive experiments show that our model can generate coherent topics and meaningful discourse roles. In addition, our model can be easily combined with other neural network architectures (such as CNN) and allows for joint training, which has presented better message classification results compared to the pipeline approach without joint training. Further discussion have shown the effectiveness of such representations for message classification, especially with joint training with CNN-based classifiers. Moreover, the source code of the proposed neural network is released for further study.

□ End of chapter.

Chapter 4

Topic Memory Networks for Short Text Classification

Many classification models work poorly on short segments of text due to data sparseness. To address this issue, in this chapter, we propose *topic memory networks* for short text classification with a novel topic memory mechanism to encode latent topical representations. Compared with existing work where most previous efforts focus on extending features with external knowledge or pre-trained topics, our model jointly explores topic inference and text classification with memory networks in an end-to-end manner. The main points of this chapter are as follows. (1) It presents the design of a topic memory network for short text classification. (2) It conducts experiments on four benchmark datasets for performance evaluation. (3) It further presents the analysis on generating coherent topics based on the proposed network. (4) It implements and open-source releases the model.

4.1 Introduction

Short texts have become an important form for individuals to voice opinions and share information on online platforms. A large body of daily-generated contents, such as tweets, web search snippets, news feeds, and forum messages, have far outpaced the reading and understanding capacity of individuals. As a consequence, there is a pressing need for automatic language understanding techniques for processing and analyzing such texts [163]. Among those techniques, text classification is a critical and fundamental one proven to be useful in various downstream applications, such as text summarization [45], recommendation [161], and sentiment analysis [24].

Although many classification models like support vector machines (SVMs) [136] and neural networks [58, 60, 146] have demonstrated their success in processing formal and well-edited texts, such as news articles [162], their performance is inevitably compromised when directly applied to short and informal online texts. This inferior performance is attributed to the severe data sparsity nature of short texts, which results in the limited features available for classifiers [99]. To alleviate the data sparsity problem, some approaches exploit knowledge from external resources like Wikipedia [54] and knowledge bases ([78, 133]). These approaches, however, rely on a large volume of high-quality external data, which may be unavailable to some specific domains or languages ([68]).

To illustrate the difficulties in classifying short texts, we take the tweet classification in Table 4.1 as an example. In the test instance S, only given the 11 words it contains, it is difficult to understand why its label is New.Music.Live. Without richer

Training instances
R_1: [SuperBowl] I'll do anything to see the **Steelers** win.
R_2: [New.Music.Live] Please give **wristbands**, she have major **Bieber** Fever.
Test instance
S: [New.Music.Live] I will do anything for **wristbands**, gonna tweet till I win.

Table 4.1: Tweet examples for classification. R_i denotes the i-th training instance; S denotes a test instance. [class] is the ground-truth label. **Bold** words are indicative of an instance's class label.

context, classifiers are likely to classify S into the same category as the training instance R_1, which happens to share many words with S, in spite of the different categories they belong to,[1] rather than R_2, which only shares the word "*wristbands*" with S. Under this circumstance, how might we enrich the context of these short texts? If looking at R_2, we can observe that the semantic meaning of "*wristbands*" can be extended from its co-occurrence with "*Bieber*", which is highly indicative of New.Music.Live.[2] Such relation can further help in recognizing the word "*wristbands*" to be important when classifying the test instance S.

Motivated by the above-mentioned observations, we present a novel neural framework, named as *topic memory networks* (TMN), for short text classification that does not rely on external knowledge. Our model can identify the indicative words for classification, e.g., "*wristbands*" in S, via jointly exploiting the

[1] R_1 is about SuperBowl, the annual championship game of the National Football League. R_2 and S are both about New.Music.Live, the flagship live music show.

[2] Justine Bieber was on New.Music.Live in 2011. There was a business activity for this event that gave free wristbands to fans if they supported Bieber on Twitter.

document-level word co-occurrence patterns, e.g., "*wristbands*" and "*Bieber*" in R_2. To be more specific, built upon the success of neural topic models [87, 118], our model is capable of discovering *latent topics*[3], which can capture the co-occurrence of words in document level. To employ the *latent topics* for short text classification, we propose a novel *topic memory mechanism*, which is inspired by memory networks [36, 140], that allows the model to put attention upon the indicative latent topics useful to classification. With such corpus-level latent topic representations, each short text instance is enriched, which thus helps alleviate the data sparsity issues.

In prior research, though the effects of topic models for short text classification have been explored [99, 107], existing methods tend to use pre-trained topics as features. To the best of our knowledge, our model is the first to encode latent topic representations via memory networks for short text classification, which allows joint inference of latent topics.

To evaluate our model, we experiment and compare it with existing methods on four benchmark datasets. Experimental results indicate that our model outperforms state-of-the-art counterparts on short text classification. The quantitative and qualitative analysis illustrate the capability of our model in generating topic representations that are meaningful and indicative of different categories.

[3] Latent topics are the distributional clusters of words that frequently co-occur in some of the instances instead of widely appearing throughout the corpus [14].

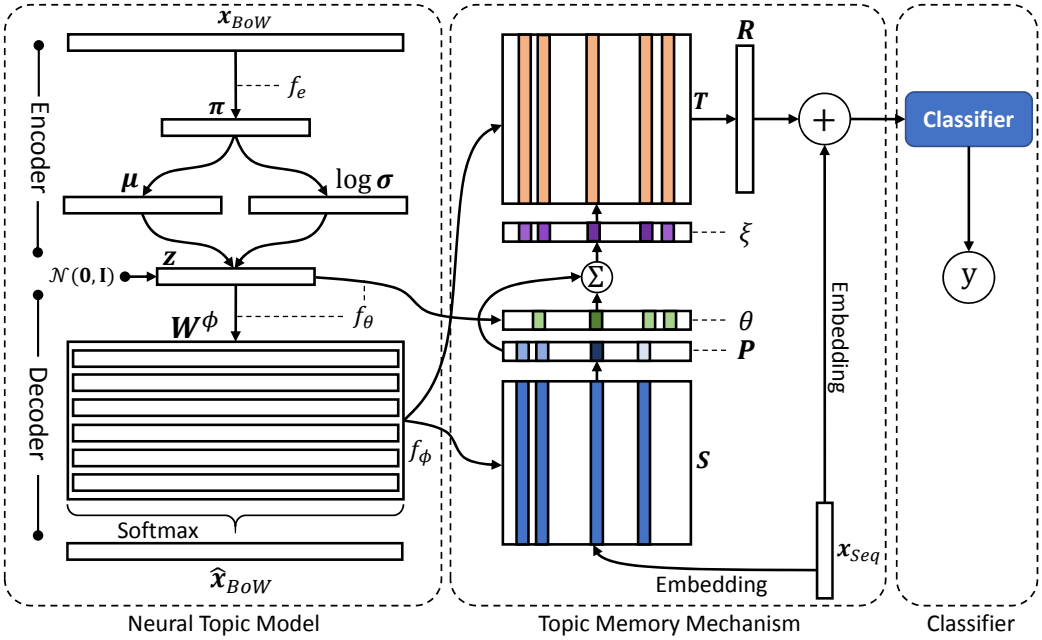


Figure 4.1: The overall framework of our topic memory networks. The dotted boxes from left to right show the neural topic model, the topic memory mechanism, and the classifier. Here the classifier allows multiple options and the details are left out.

4.2 Topic Memory Networks

In this section, we describe our topic memory networks (TMN), whose overall architecture is shown in Figure 4.1. There are three major components: (1) a neural topic model (NTM) to induce latent topics (described in Section 4.2.1), (2) a topic memory mechanism that maps the inferred latent topics to classification features (described in Section 4.2.2), and (3) a text classifier, which produces the final classification labels for instances. These three components can be updated simultaneously via a joint learning process, which is introduced in Section 4.2.3. In particular, for the classifier, our TMN framework allows the combination of multiple options, e.g., CNN and RNN, which can be determined by the specific application

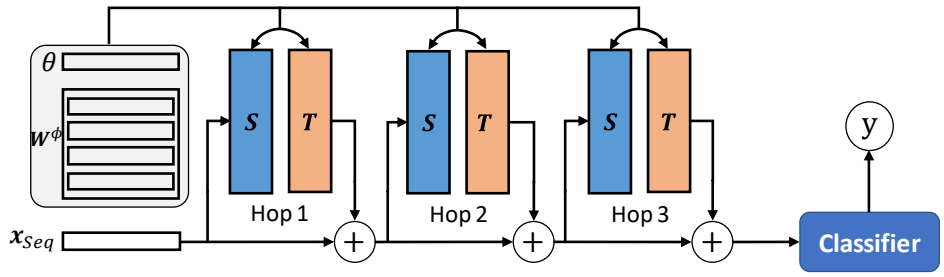


Figure 4.2: Topic memory network with three hops.

scenario.

Formally, given $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^M\}$ as the input with M short text instances, each instance $\mathbf{x}$ is processed into two representations: bag-of-words (BoW) term vector $\mathbf{x}_{BoW} \in \mathbb{R}^V$ and word index sequence vector $\mathbf{x}_{Seq} \in \mathbb{R}^L$, where V is the vocabulary size and L is the sequence length. $\mathbf{x}_{BoW}$ is fed into the neural topic model to induce latent topics. Such topics are further matched with the embedded $\mathbf{x}_{Seq}$ to learn classification features in the topic memory mechanism. Then, the classifier concatenates the representations produced by the topic memory mechanism and the embedded $\mathbf{x}_{Seq}$ to predict the classification label y for $\mathbf{x}$.

4.2.1 Neural Topic Model

Our topic model is inspired by neural topic model (NTM) [87, 118] that induces latent topics in neural networks. NTM is based on variational auto-encoder (VAE) [62], involved with a continuous latent variable $\mathbf{z}$ as an intermediate representation. Here in NTM, the latent variable $\mathbf{z} \in \mathbb{R}^K$, where K denotes the number of topics. In the following, we describe the generation and the inference of the model in turn.

NTM Generation. Similar to LDA-style topic models, we assume $\mathbf{x}$ having a topic mixture θ represented as a K-dimensional distribution, which is generated via Gaussian softmax construction [87]. Each topic k is represented by a word distribution ϕ_k over the vocabulary. Specifically, the generation story for $\mathbf{x}$ is:

- Draw latent variable $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
- $\theta = \text{softmax}(f_\theta(\mathbf{z}))$
- For the n-th word in $\mathbf{x}$:
  - Draw word $w_n \sim \text{softmax}(f_\phi(\theta))$

where $f_*(\cdot)$ is a neural perceptron that linearly transforms inputs, activated by a non-linear transformation. Here we use rectified linear units (ReLUs) [90] as activate functions. The prior parameters of $\mathbf{z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, are estimated from the input data and defined as:

$$\boldsymbol{\mu} = f_\mu(f_e(\mathbf{x}_{BoW})),\ \log \boldsymbol{\sigma} = f_\sigma(f_e(\mathbf{x}_{BoW})) \tag{4.1}$$

Note that NTM is based on VAE, where an encoder estimates the prior parameters and a decoder describes the generation story. Compared with the basic VAE, NTM includes the additional distributional vectors θ and ϕ, which can yield latent topic representations and thus ensuring their better interpretability in learning process [87].

NTM Inference. In NTM, we use variational inference [10] to approximate a posterior distribution over $\mathbf{z}$ given all the instances. The loss function of NTM is defined as

$$\mathcal{L}_{NTM} = D_{KL}(q(\mathbf{z}) \,||\, p(\mathbf{z} \,|\, \mathbf{x})) - \mathbb{E}_{q(\mathbf{z})}[p(\mathbf{x} \,|\, \mathbf{z})] \tag{4.2}$$

the negative of variational lower bound, where $q(\mathbf{z})$ is a standard Normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. $p(\mathbf{z} \,|\, \mathbf{x})$ and $p(\mathbf{x} \,|\, \mathbf{z})$ are probabilities to

describe encoding and decoding processes, respectively.[4] Due to the space limitation, we leave out the derivation details and refer the readers to [87].

4.2.2 Topic Memory Mechanism

We exploit a topic memory mechanism to map the latent topics produced by NTM (described in Section 4.2.1) to the features for classification. Inspired by memory networks [125, 140], we design two memory matrices, a source memory $\boldsymbol{S}$ and a target memory $\boldsymbol{T}$, both of which are in $K \times E$ size (K for the number of topics and E for the pre-defined size of word embeddings). $\boldsymbol{S}$ and $\boldsymbol{T}$ are produced by two ReLU-actived neural perceptrons, both taking the topic-word weight matrix $\boldsymbol{W}^{\phi} \in \mathbb{R}^{K \times V}$ as inputs. Recall that in NTM, we use $f_{\phi}(\cdot)$ to compute the word distributions given θ. $\boldsymbol{W}^{\phi}$ is the kernel weight matrix of $f_{\phi}(\cdot)$, where $\boldsymbol{W}^{\phi}_{k,v}$ represents the importance of the v-th word in reflecting the k-th topic. Assuming $\boldsymbol{U}$ as the embedded $\mathbf{x}_{Seq}$ (word sequence form of $\mathbf{x}$), in source memory, we compute the match between the k-th topic and the embedding of the l-th word in $\mathbf{x}_{Seq}$ by

$$\boldsymbol{P}_{k,l} = \text{sigmoid}(\mathbf{W}^{s}[\boldsymbol{S}_{k}; \boldsymbol{U}_{l}] + b^{s}) \tag{4.3}$$

where $[\mathbf{x}; \mathbf{y}]$ denotes the merge of $\mathbf{x}$ and $\mathbf{y}$, and we use concatenation operation here [32, 24]. $\mathbf{W}^{s}$ and b^{s} are parameters to be learned. To further combine the instance-topic mixture θ with $\boldsymbol{P}$, we define the integrated memory weights as

$$\xi_{k} = \theta_{k} + \gamma \sum_{l} \boldsymbol{P}_{k,l} \tag{4.4}$$

[4] In implementation, to smooth the gradients, we apply reparameterization on $\mathbf{z}$ following previous work [62, 108].

where γ is the pre-defined coefficient. Then, in target memory, via weighting target memory matrix $\boldsymbol{T}$ with ξ, we obtain the output representation $\boldsymbol{R}$ of the topic memory mechanism:

$$\boldsymbol{R}_k = \xi_k \boldsymbol{T}_k \tag{4.5}$$

The concatenation of $\boldsymbol{R}$ and $\boldsymbol{U}$ (embedded $\mathbf{x}_{Seq}$) further serves as the features for classification.

In particular, similar to the memory networks in prior research [24, 125], our model can be extended to handle multiple computation layers (hops). As shown in Figure 4.2, each hop contains a source matrix and a target matrix, and different hops are stacked following the way presented in [125].

4.2.3 Joint Learning

The entire TMN model integrates the three modules in Figure 4.1, i.e., the neural topic model, the topic memory mechanism, and the classifier, which can be updated simultaneously in one framework. In doing so, we jointly tackle topic modeling and classification, and define the loss function of the overall framework to combine the two effects as following:

$$\mathcal{L} = \mathcal{L}_{NTM} + \lambda \mathcal{L}_{CLS} \tag{4.6}$$

where $\mathcal{L}_{NTM}$ represents the loss of NTM and $\mathcal{L}_{CLS}$ is the cross entropy reflecting classification loss. λ is the trade-off parameter controlling the balance between topic model and classification.

4.3 Experiment Setup

4.3.1 Datasets

We conduct experiments on four short text datasets, namely, Snippets, TagMyNews, Twitter, and Weibo. Their details are described as follows.

Snippets. This dataset contains Google search snippets released by [99]. There are eight ground-truth labels, e.g., *health* and *sport*.

TagMyNews. We use the news titles as instances from the benchmark classification dataset released by [130].[5] This dataset contains English news from really simple syndication (RSS) feeds. Each news feed (with its title) is annotated with one from seven labels, e.g., *sci-tech*.

Twitter. This dataset is used to evaluate tweet topic classification, which is built on the dataset released by TREC2011 microblog track.[6] Following previous settings [68, 148], hashtags, i.e., user-annotated topic labels in each tweet such as "*#Trump*" and "*#SuperBowl*", serve as our ground-truth class labels. Specifically, we construct the dataset with the following steps. First, we remove the tweets without hashtags. Second, we rank hashtags by their frequencies. Third, we manually remove the hashtags that cannot mark topics, such as "*#fb*" for indicating the source of tweets from Facebook, and combine the hashtags referring to the same topic, such as "*#DonaldTrump*"

[5] http://acube.di.unipi.it/tmn-dataset/

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

Dataset	# of labels	# of docs	Avg len per doc	Vocab size
Snippets	8	12,332	17	7,334
TagMyNews	7	32,567	8	9,433
Twitter	50	15,056	5	6,962
Weibo	50	21,944	6	10,121

Table 4.2: Statistics of the experimental datasets. Labels refers to class labels. Avg len per doc refers to the average count of words in each document instance.

and "*#Trump*". Finally, we select the top 50 frequent hashtags, and all tweets containing these hashtags.

Weibo. To evaluate our model on a different language other than English, we employ a Chinese dataset with short segments of text for topic classification. This dataset is released by [70] with a collection of messages posted in June 2014 on Weibo, a popular Twitter alike platform in China.[7] Similar to Twitter, Weibo allows up to 140 Chinese characters in its messages. In this Weibo dataset, each Weibo message is labeled with a hashtag as its category, and there are 50 distinct hashtag labels in total, following the same procedure performed for the Twitter dataset.

Table 4.2 shows the statistic information of the four datasets. Each dataset is randomly split into 80% for training and 20% for test. 20% of randomly selected training instances are used to form development set. We preprocess our English datasets, i.e., Snippets, TagMyNews, and Twitter, with *gensim tokenizer*[8] for tokenization. As to the Chinese Weibo dataset, we use

[7]The original dataset contains conversations to enrich the context of Weibo posts, which are not considered here.

[8]`https://radimrehurek.com/gensim/utils.html`

FudanNLP toolkit [104][9] for word segmentation. In addition, for each dataset, we maintain a vocabulary built based on the training set with removal of stop words[10] and words occurring less than 3 times. The inputs of topic models $\mathbf{x}_{BoW}$ are constructed based on this vocabulary following common topic model settings [14, 88]. Differently, we use the raw word sequence (without words removal) for the inputs of classification $\mathbf{x}_{Seq}$ as is done in previous work of text classification [60, 77].

4.3.2 Model Settings

We use pre-trained embeddings to initialize all word embeddings. For Snippets and TagMyNews datasets, we use pre-trained `GloVe` embeddings [96][11]. For Twitter and Weibo datasets, we pre-train embeddings on large-scale external data with 99M tweets and 467M Weibo messages, respectively. For the number of topics, we follow previous settings [28, 30, 148] to set $K = 50$. For all the other hyperparameters, we tune them on the development set by grid search. For our classifier, we employ CNN in experiment because of its better performance in short text classification than its counterparts such as RNN [133]. The hidden size of CNN is set as 500. The dimension of word embedding $E = 200$. $\gamma = 0.8$ for trading off θ and $\boldsymbol{P}$, and $\lambda = 1.0$ for controlling the effects of topic model and classification. In the learning process, we run our model for 800 epochs with early-stop strategy applied [20].

[9] `https://github.com/FudanNLP/fnlp`

[10] `https://radimrehurek.com/gensim/parsing/preprocessing.html`

[11] `http://nlp.stanford.edu/data/glove.6B.zip` (200d)

4.3.3 Comparison Models

For comparison, we consider a weak baseline of majority vote, which assigns the major class labels in training set to all test instances. We further compare with the widely-used baseline SVM+BOW, SVM with unigram features [136]. We also consider other SVM-based baselines: SVM+LDA, SVM+BTM, SVM+NTM, whose features are topic distributions for instances learned by LDA [14], BTM [148], and NTM [87], respectively. In particular, BTM is one of the state-of-the-art topic models for short texts. To compare with neural classifiers, we test bidirectional long short-term memory with attention (AttBiLSTM) [160] and convolutional neural network (CNN) classifiers [60]. No topic representation is encoded in these two classifiers. We also compare with the state-of-the-art short-text classifier CNN+TEWE [107], i.e., CNN classifier with topic-enriched word embeddings (TEWE), where the word embeddings are enriched by pre-trained NTM-inferred topic models. Moreover, to investigate the effectiveness of our proposed topic memory mechanism, we compare with CNN+NTM, which concatenates the representations learned by CNN and topics induced by NTM as classification features. In addition, we compare with our variant, TMN (*Separate TM Inference*), where topics are induced separately before classification, and only used for initializing the topic memory. To be consistent, our model with a joint learning process for topic modeling and classification, described in Section 4.2.3, is named as TMN (*Joint TM Inference*). Note that the comparison CNN-based models share the same settings as our model, and the hidden size for each direction of BiLSTM is set to 100.

Models	Snippets		TagMyNews		Twitter		Weibo	
	Acc	Avg F1	Acc	Avg F1	Acc	Avg F1	Acc	Avg F1
Comparison models								
Majority Vote	0.202	0.068	0.247	0.098	0.073	0.010	0.102	0.019
SVM+BOW [136]	0.210	0.080	0.259	0.058	0.070	0.009	0.116	0.039
SVM+LDA [14]	0.689	0.694	0.616	0.593	0.159	0.111	0.192	0.147
SVM+BTM [148]	0.772	0.772	0.686	0.677	0.232	0.164	0.331	0.277
SVM+NTM [87]	0.779	0.776	0.664	0.654	0.261	0.177	0.379	0.348
AttBiLSTM [157]	0.943	0.943	0.838	0.828	0.375	0.348	0.547	0.547
CNN [60]	0.944	0.944	0.843	0.843	0.381	0.362	0.553	0.550
CNN+TEWE [107]	0.944	0.944	0.846	0.846	0.385	0.368	0.537	0.532
CNN+NTM	0.945	0.945	0.844	0.844	0.382	0.365	0.556	0.556
Our models								
TMN (*Separate TM Inference*)	0.961	0.961	0.848	0.847	0.394	**0.386**	0.568	0.569
TMN (*Joint TM Inference*)	**0.964**	**0.964**	**0.851**	**0.851**	**0.397**	0.375	**0.591**	**0.589**

Table 4.3: Comparisons of accuracy (Acc) and average F1 (Avg F1) on four benchmark datasets. Our TMN, either with separate or joint TM inference, performs significantly better than all the comparisons ($p < 0.05$, paired t-test).

4.4 Experimental Results

4.4.1 Classification Comparison

Table 4.3 shows the comparison on classification results, where the accuracy and average F1 scores on different classes labels are reported. We have the following observations.

• ***Topic representations are indicative features.*** On all four datasets, simply by combining topic representations into features, SVM models produce better results than the models without exploiting topic features (*i.e.*, SVM+BOW). This observation indicates that latent topic representations captured at corpus level are helpful to alleviate the data sparsity problem in short text classification.

Model	Snippets	TagMyNews	Twitter
LDA	0.436	0.449	0.436
BTM	0.435	0.463	0.435
NTM	0.463	0.468	0.463
TMN	**0.487**	**0.499**	**0.468**

Table 4.4: C_V coherence scores for topics generated by various models. Higher is better. The best result in each column is in **bold**.

• ***Neural network models are effective.*** It is seen that neural models based on either CNN or AttBiLSTM yield better results than SVM. This observation shows the effectiveness of representation learning in neural networks for short texts.

• ***CNN serves as a better classifier for short texts than AttBiLSTM.*** In comparison of CNN and AttBiLSTM without taking topic features, we observe that CNN yields generally better results on all the four datasets. This is consistent with the discovery in [133], where CNN can better encode short texts than sequential models.

• ***Topic memory is useful to classification.*** By exploring topic representations in memory mechanisms, our TMN model, inferring topic models either separately or jointly with classification, significantly outperform the best comparison models on each of the four datasets. Particularly, when compared with CNN+TEWE and CNN+NTM, both concatenating topics as part of the features, the results yielded by TMN are better. This demonstrates the effectiveness of topic memory to learn indicative topic representations for short text classification.

• ***Jointly inferring latent topics is effective to text classification.***

In comparison between two TMN variants, TMN (*Joint TM Inference*) produces better classification results, though large margin improvements are not observed on the three English datasets, i.e., TagMyNews, Snippets, and Twitter. This may be because the classifiers do not rely too much on high-quality latent topics, since other features may be sufficient to indicate the labels, e.g., word positions in the instance. As a result, better topic models, learned via jointly induced with classification, may not provide richer information for classification. Nevertheless, we notice that on Chinese Weibo dataset, the jointly trained topic model improves the accuracy and average F1 by 2.3% and 2.0%, respectively. It may result from the prevalence of word order misuse in informal Weibo messages. This mis-order phenomenon is common in Chinese and generally does not affect understanding. The rich information conveyed by Chinese characters are capable of indicating semantic meanings of words even without correct orders [103, 135]. As a result, the CNN classifier, which encodes orders of words, may also bring such mis-order noise to classification. For these instances with mis-ordered words, a better topic model that learns text instances as unordered words, provides useful representations that compensate the loss of information in word orders and in turn improves the performance of text classification.

4.4.2 Topic Coherence Comparison

In Section 4.4.1, we find that TMN can significantly outperform comparison models on short text classification. In this section, we study whether jointly learning topic models and classification can be helpful in producing coherent and meaningful topics.

LDA	mubarak bring run obama democracy speech believe regime power bowl
BTM	mubarak egypt push internet people government phone hosni need son
NTM	mubarak people egyptian egypt stay tomorrow protest news phone protester
TMN	mubarak protest protester tahrir square egyptian al jazeera repo cairo

Table 4.5: Top 10 representative terms of the sample latent topics discovered by various topic models from Twitter dataset. We interpret the topics as "*Egyptian revolution of 2011*" according to their word distributions. Non-topic words are wave-underlined and in blue, and off-topic words are underlined and in red.

We use the C_V metric [110] computed by Palmetto toolkit[12] to evaluate the topic coherence, which has been shown to give the closest scores to human evaluation compared to other widely-used topic coherence metrics like NPMI [16]. Table 4.4 shows the comparison results of LDA, BTM, NTM, and TMN on the three English datasets.[13] Note that we do not report C_V scores for Chinese Weibo dataset as the Palmetto toolkit cannot process Chinese topics.

As can be seen, TMN yields higher C_V scores by large margins than all others in comparison. This indicates that jointly exploring classification would be effective in producing coherent topics. The reason is that the supervision from classification labels can

[12] `https://github.com/dice-group/Palmetto`

[13] Without otherwise indicated, TMN is used as a short form for TMN (*Joint TM Inference*).

# of Hops	Snippets	TagMyNews	Twitter	Weibo
TMN-1H	0.958	0.841	0.382	0.568
TMN-2H	**0.964**	0.843	0.383	0.578
TMN-3H	0.962	0.845	0.384	0.581
TMN-4H	0.961	0.846	0.389	0.582
TMN-5H	0.960	**0.851**	**0.397**	**0.591**
TMN-6H	0.958	0.848	0.388	0.579

Table 4.6: The impact of the # of hops on accuracy.

guide unsupervised topic models in discovering meaningful and interpretable topics. We also observe that NTM produces better results than LDA and BTM, which implies the effectiveness of inducing topic models by neural networks.

To further analyze the quality of yielded topics, Table 4.5 shows the top 10 words of the sample latent topics reflecting "*Egyptian revolution of 2011*" discovered by various models. We find that LDA yields off-topic word "*bowl*". For the results of BTM and NTM, though we do not find off-topic words, non-topic words like "*need*" and "*stay*" are included.[14] The topic generated by TMN appears to be the best, which presents indicative words like "*tahrir*" and "*cairo*", for the event.

4.4.3 Results with Varying Hyperparameters

We further study the impact of two important hyperparameters in TMN, i.e., the hop number and the topic number, which will be discussed in turn.

Impact of Hop Numbers. Recall that Figure 4.2 shows the capacity of TMN in combining multiple hops. Here we analyze

[14]Off-topic words are more likely to be interpreted to reflect other topics. Non-topic words cannot clearly indicate the corresponding topic.

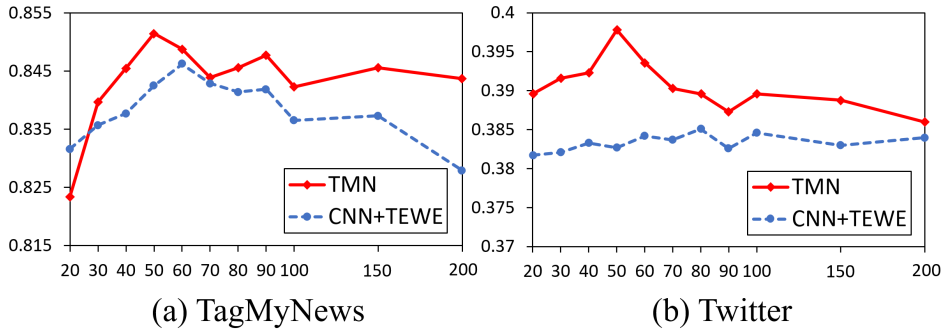
the effects of hop numbers on the accuracy of TMN. Table 4.6 reports the results, where NH refers to using N hops ($N = 1, 2, ..., 6$). As can be seen, generally, TMN with 5 hops achieves the best accuracy on most datasets except for Snippets dataset. We also observe that, although within a particular range, more hops can produce better accuracy, the increasing trends are not always monotonic. For example, TMN-6H always exhibits lower accuracy than TMN-5H. This observation implies that the overall representation ability of TMN is enhanced as the increasing complexity of the model via combining more hops. However, this enhancement will reach saturation when the hop number exceeds a threshold, which is 5 hops for most datasets in our experiment.

Impact of Topic Numbers. Figure 4.3 shows the accuracy of TMN and CNN+TEWE (the best comparison model in Table 4.3) given varying K, the number of topics on TagMyNews and Twitter datasets.[15] As we can see, the curves of all the models are not monotonic and the best accuracy is achieved given a particular number of topics, e.g., K=50 for TMN on TagMyNews dataset. When comparing different curves, we observe that TMN yields consistently better accuracy than CNN+TEWE, a comparison model shown in Table 4.3, which demonstrates the robust performance of TMN over varying number of topics.

4.4.4 A Case Study on Topic Memory

Section 4.4.1 demonstrates the effectiveness of using topic memory on short text classification. To further understand why,

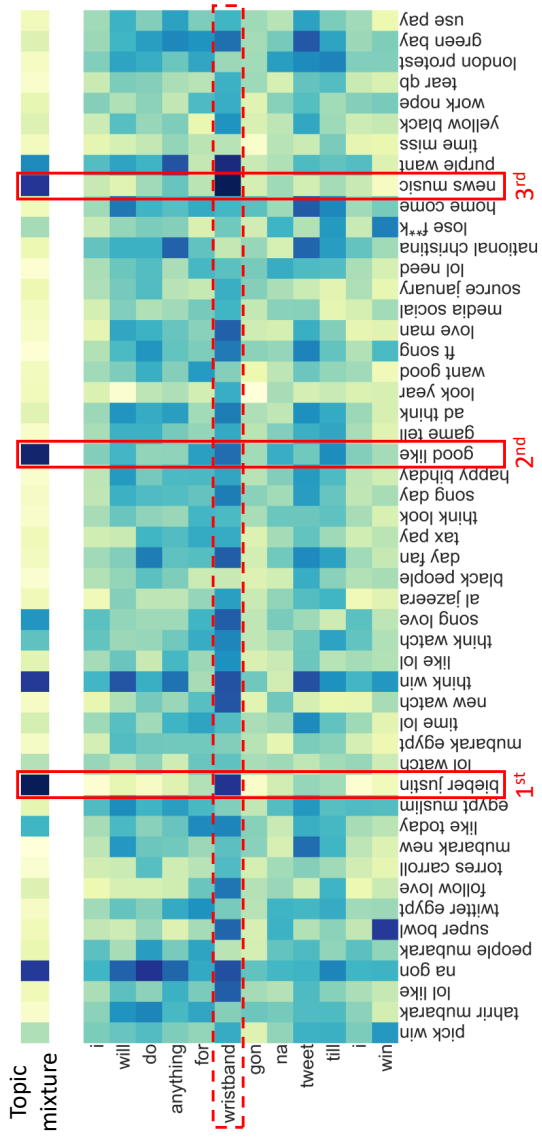
[15]We observe similar distributions on Snippets and Weibo.



(a) TagMyNews (b) Twitter

Figure 4.3: The impact of topic numbers, where the horizontal axis shows the number of topics and the vertical axis shows the accuracy.

in this section, we use the test instance S in Table 4.1 to analyze what the information captured by topic memory is indicative of class labels. Recall that the label of S, which should be New.Music.Live, can be indicated by containing word "*wristbands*" and the collocation of "*wristbands*" and "*Bieber*" in training instance R_2 labeled New.Music.Live. Figure 4.4 shows the heatmaps of the weight matrix $\boldsymbol{P}$ in topic memory and the topic mixture θ captured by NTM for instance S. As can be seen, the top 3 words for the latent topic with the largest value in θ are "*bieber*", "*justine*", and "*tuesday*", which can effectively indicate the class label of S to be *New.Music.Live* because Justine Bieber was there on Tuesday. Interestingly, S contains none of the top three words. The latent semantic relations of S and these words are purely uncovered by the co-occurrence of words in S with other instances in the corpus, which further shows the benefit of using latent topics for alleviating the sparsity in short texts. We also observe that topic memory learns different representations for topical word "*wristband*", highly indicating instance label, and background words, such as "*i*" and "*for*". This explains why topic memory is effective to classification.



(a)

1st Topic	2nd Topic	3rd Topic
bieber	good	news
justin	like	music
tuesday	fan	forever
newmusiclive	look	change
think	tweet	right
thought	thing	attack
jb	know	time
new	rumble	wing
music	na	aistdirect
live	right	films

(b)

Figure 4.4: Topic memory visualization for test instance S shown in Table 4.1. (a) Heatmaps of topic mixture θ (the upper one) and topic memory weight matrix $\boldsymbol{P}$ (the lower one) illustrating the relevance between the words of S (left) and the learned topics (bottom, with top-2 words displayed). The red dotted rectangle indicates the representation for "*wristband*", the topical word in S. The red rectangles with solid frames indicates the 3 most relevant topics ordered by θ. (b) Top-10 words of these topics indicated by ϕ.

4.4.5 Error Analysis

In this section, we take our classification results on TagMyNews dataset as an example to analyze our errors. We observe that one major type of incorrect prediction should be ascribed to the polysemy phenomenon. For example, the instance "*NBC gives 'the voice' post super bowl slot*" should be categorized as *entertainment*. However, failing to understand the particular meaning of "*the voice*" here as the name of a television singing competition, our model mistakenly categorizes this instance as sport because of the occurrence "*super bowl*". In future work, we would exploit context-sensitive topical word embeddings [143], which is able to distinguish the meanings of the same word in different contexts. Another main error type comes from the failure to capture phrase-level semantics. Taking "*On the merits of face time and living small*" as an example, without understanding "*face time*" as a phrase, our model wrongly predicts its category as business instead of its correct label as sci_tech. Such errors can be reduced by enhancing our NTM to phrase discovery topic models [38, 76], which is worthy exploring in future work.

4.5 Summary

In this chapter, we have presented topic memory networks that exploit corpus-level topic representations with a topic memory mechanism for short text classification. The model alleviates data sparsity issues via jointly learning latent topics and text categories. Empirical comparisons with state-of-the-art models on four benchmark datasets have demonstrated the validity and effectiveness of our model, where better results have been

achieved on both short text classification and topic coherence evaluation.

□ **End of chapter.**

Chapter 5

The Roles of Dynamic Topics and Discourse in Argumentation Process

This chapter presents a study that automatically analyzes the key factors in argument persuasiveness, beyond simply predicting who will win the debate. The key notion is that we propose a novel neural model which is able to dynamically track the changes of latent topics and discourse in argumentative conversations, allowing the investigation of their roles in influencing the outcomes of persuasion. The main points of this chapter are as follows. (1) It presents a novel neural model that can analyze the changes of both topics and discourse in argumentative conversations. (2) It conducts extensive experiments on argumentative conversations on both social media and supreme court. (3) It presents the effects of topics and discourse on persuasiveness, and achieves that they are both useful. (4) It draws some findings from our empirical results, which will help people better engage in future persuasive conversations.

5.1 Introduction

In our current world with full of uncertainty, arguments play a central role in making decisions, constructing knowledge, and bringing truths and better ideas to life [55]. The understanding of these argumentation processes will help individuals better engage with conflicting stances and open up their minds to pros and cons [65]. It collides different ideas to form thoughts and knowledge, contributing to advance science and society forward [141]. However, making sense of argumentative conversations is a daunting task for human readers, mostly due to the varied viewpoints and evidence continuously put forward and the complicated interaction structure therein; not to mention huge volume of data containing argumentation processes appeared on online platforms every day.

We hence study how to automatically understand argumentation processes, predicting who will persuade whom and figuring out why it happens. To date, much progress made in persuasiveness prediction has focused on individual arguments, the wordings therein [37, 137] and how they locally connect with other arguments [40, 49]. On the contrary, we examine the context and the dynamic progress of argumentative conversations, which is beyond the studies of argument-level persuasiveness. Some research works analyze argument interactions [39, 55, 126, 134] to predict who will win. Most of them focus on the outcome of argumentation instead of diving deep into the persuasion process. The latter, however, is arguably the essence of argumentation, revealing how participants collaborate to reshape and refine ideas.

In light of these missing points, we track the argumentation

...
A_1 [*Evidence*]: ... There is research that indicates "that *those who spoke two or more languages had significantly better cognitive abilities compared to what would have been expected from their baseline test*." ⟨url⟩. ... Another study found that " *the language-learning participants ended up with increased density in their grey matter and that their white matter tissue had been strengthened.* " ⟨url⟩
A_2 [*Metaphor*]: The common comparison is made to learning music, as /u/awesomeosprey has pointed out. I did some research into the matter. It seems that *learning a musical instrument does have long-lasting benefits* (⟨url⟩) *that relate to "higher-order aspects of cognition.*"
...
A_4 [*Reference*] ... But a quick search and I have other sources: ⟨digit⟩ ⟨url⟩, ⟨digit⟩ ⟨url⟩, ⟨digit⟩ ⟨url⟩. The most interesting study is this one (⟨url⟩), but I can't find a complete version of it, sorry. /n/nNote: Study ⟨digit⟩ has an exceptionally small sample size. It's still interesting reading.

Figure 5.1: A ChangeMyView conversation snippet of challengers' arguments against "*learning a second language isn't worth it anymore for most people*" (raised by an opinion holder). The red and italic words indicate the key points resulting in the challengers' victory. The words in [] are our interpretations of the arguments' discourse styles.

process and explicitly explore the dynamic patterns of what a discussion is centered around (henceforth **topics**) and how the participants voice their opinion in arguments (henceforth **discourse**), as well as how they affect the persuasion results. To illustrate the interplay of topics and discourse in argument persuasiveness, Figure 5.1 shows a Reddit conversation snippet from ChangeMyView subreddit.[1] It is formed with challengers' arguments against "*learning a second language isn't worth it for most people anymore*", which was raised by an opinion holder. It is seen that the challengers successfully persuaded the opinion holder to change their view in the aforementioned example. The probable reasons are two fold. First, there are strong evidences (reflected by topic words) put forward, such as the research findings on cognitive abilities. Second, they deploy skillful debating styles (captured by discourse words), such as the metaphors with learning music (in A_2) and the reference to external information (in A_4).

Motivated with these observations, we propose a novel neural framework that explicitly models how the change of discussion topic and discourse styles affect persuasion effectiveness. Our model first explores latent topic and discourse in arguments with word clusters. Furthermore, it tracks topic change and discourse flow in argumentation processes and automatically interpret the key factors indicating the success or failure of the persuasion. Coupling the advantages of neural topic models [87, 152, 155] and dynamic memory networks [64, 147, 159], we are able to explore dynamic topic and discourse representations indicative of persuasiveness in an end-to-end manner with the persuasion

[1] On ChangeMyView (`https://www.reddit.com/r/changemyview/`), an opinion holder first raises a viewpoint, followed by challengers' arguments attempting to change the opinion holder's mind.

outcome prediction. To the best of our knowledge, we are the first to explicitly model topics and discourse in argumentation processes, and investigate how they contribution to argument persuasiveness.

We carry out extensive experiments on argumentative conversations from both social media and supreme court. The results show that our model can effectively identify persuasive arguments, significantly outperforming state-of-the-art methods on both datasets. For example, we achieve 83.3% accuracy when predicting winners in supreme court debates, compared with 63.1% obtained by logistic regression without explicitly exploring dynamic topics and discourse features in argumentation processes. Based on the produced topics and discourse, we further analyze how they affect persuasiveness. It is indicated that topics (such as evidence and viewpoints) statistically contribute more on persuasion success while skillful discourse style may sometimes lead to victory. In addition, we summarize the key findings from our empirical results, which will help individuals better engage in future persuasions.

5.2 Problem and Data Description

In this section, we first introduce how we formulate our problem, followed by a description about data collection and analysis.

Problem Formulation. As discussed above, our work studies argument persuasiveness, which however relies on subjective judgement. After all, human performance on persuasiveness judgement is still close to random guess [126]. We hence adopt the pairwise comparison settings following [126] to take a pair of

Datasets	# of moots	# of convs	# of turns	avg. words per turn	vocab size
CMV	2,396	10,341	39,644	96.2	13,541
Court	204	655	17,599	46.1	6,260

Table 5.1: Statistics of the ChangeMyView (CMV) and the Supreme Court (Court) datasets. Here a moot refers to an original post in CMV and a case in Court.

persuasion conversations as input. Our goal is to analyze the key factors in their argumentation processes and predict which one has a better chance to win. Furthermore, to ensure the input pair to be comparable, its two conversations should be under the same discussion subject (i.e., **moot**). For example, in court debate, we take arguments from both sides concerning the same case as input.

Data Description. Our problem setting can be fit in various applications to learn what a good persuasion should be. Here we conduct our study on two scenarios — social media arguments, which tend to use colloquial and informal languages, and supreme court debates, exhibiting a more formal language style. The social media arguments are from the ChangeMyView subreddit, where the participants engaged in discussion both attempting to change opinion holder's view. We aim to predict which conversation has a better chance to achieve a Δ, awarded by opinion holders to indicate successful persuasion. For the supreme court debates, we aim to predict whether the petitioner or respondent will win the case, given their corresponding conversational exchanges with the justices.

The ChangMyView social media dataset (henceforth **CMV**) is built based on a corpus released by [126] with argumentative

conversations held from Jan 2013 to May 2015. Each discussion in CMV can be organized in tree structure with in-reply-to relations, its root is an original post with opinion holder's view. When formulating our pairwise input, we extract two conversation paths, one awarded with Δ and other not.[2] Following [126], we also make filtering to acquire high quality arguments when constructing our dataset.

For the supreme court debate dataset (abbreviated as **Court**), it is gathered by [27] with U.S. supreme court dialogues.[3] Here given the arguments delivered by both sides' lawyers to justices, we predict the one more likely to win justices' favor. The statistics of our two datasets are shown in Table 5.1. As can be seen from Table 5.1, there are more conversations in CMV than Court. However, the Court debates involve more turns (26.9 vs. 3.8 turns on average per conversation). It might be because court debates usually result in a back-and-forth fashion while social media discussions may end soon.

Here we do not feed the words either form opinion holder or justices to avoid their possible influence on persuasiveness computing. It allows us to focus on linguistic features in participants' arguments that lead to good persuasion. Besides, in doing so, our setting can be adapted to scenarios without the third-party engagement (e.g., opinion holders and justices).

[2] In this chapter, without otherwise specified a **conversation** is used as the short form of a conversation path on ChangeMyView.

[3] `http://www.supremecourt.gov/oral_arguments/`

5.3 DTDMN: Dynamic Topic-Discourse Memory Networks for Persuasiveness

This section presents our model that predicts persuasiveness, and dynamically discovers the key topic and discourse factors therein to explain the reasons behind. Our model, named as dynamic topic-discourse memory networks (DTDMN), consists of three modules — one to learn latent topic and discourse factors from each argument (henceforth **argument factor encoder**), one to explore the change of topic and discourse factors in argumentation flows (henceforth **dynamic process encoder**), and the last one to identify the more persuasive conversation from the input pair (henceforth **persuasiveness predictor**). The model architecture is shown in Figure 5.2 with an overview presented in Section 5.3.1. Then in Section 5.3.2, 5.3.3 and 5.3.4, we describe our three modules in turn, followed by our our learning objective discussed in Section 5.3.5.

5.3.1 Model Overview

As described in Section 5.2, our model takes pairwise conversations as input. In training, we feed $\langle \boldsymbol{C}^+; \boldsymbol{C}^- \rangle$ into our model, where $\boldsymbol{C}^+$ is a positive instance referring to a persuasive conversation. Likewise, $\boldsymbol{C}^-$, the negative instance, denotes a failed persuasion. During the testing, given two conversations, our model will recognize the one which is more persuasive. Each conversation $\boldsymbol{C}$ is formed with a sequence of argumentative turns (henceforth **arguments**): $\boldsymbol{C} = \langle \boldsymbol{x}^1, \ldots, \boldsymbol{x}^T \rangle$, where T denotes the number of arguments in $\boldsymbol{C}$.

For the t-th argument $\boldsymbol{x}^t$, we capture argument-level representations, $\boldsymbol{z}^t \in \mathbb{R}^K$ for topic factor and $\boldsymbol{d}^t \in \mathbb{R}^D$ for discourse factor,

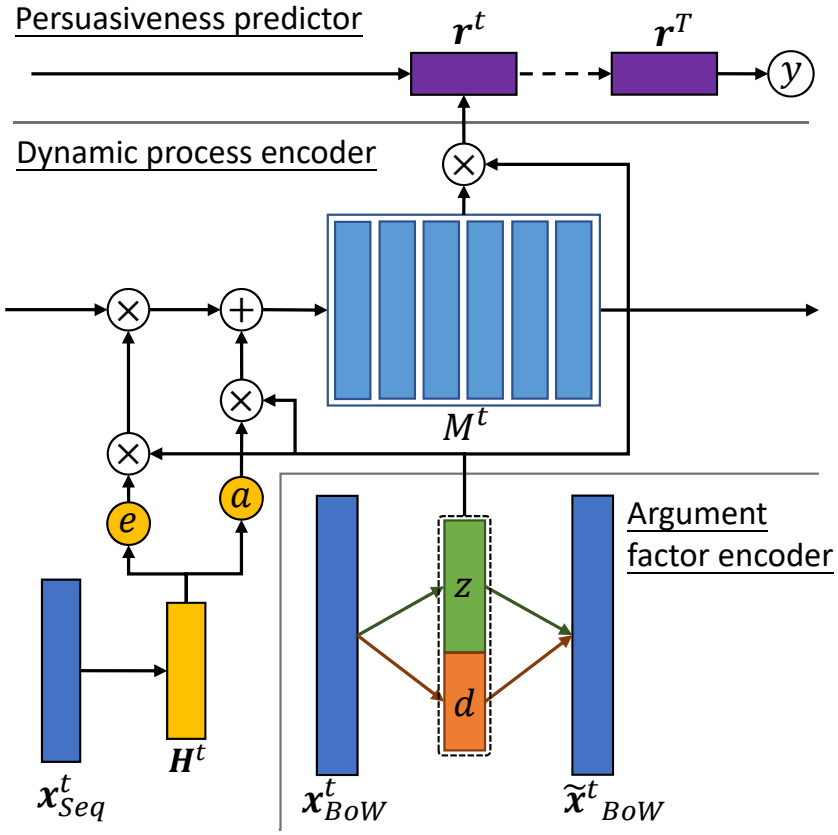


Figure 5.2: The architecture of our dynamic topic-discourse memory networks (DTDMN) for persuasiveness prediction.

from the input of bag-of-words vector $\boldsymbol{x}^t_{BoW} \in \mathbb{R}^V$, where K and D is the number of topics and discourse, respectively, V is the vocabulary size. Then, $\boldsymbol{z}^t$ and $\boldsymbol{d}^t$ are fed into the dynamic memory, together with the word index sequence $\boldsymbol{x}^t_{Seq} \in \mathbb{R}^L$, to update the memory state, where L is the sequence length. The output of the dynamic memory network is used to predict the persuasiveness score y for each conversation, where higher scores indicate better persuasiveness. Our training target is to have $y^+ > y^-$ for $\boldsymbol{C}^+$ and $\boldsymbol{C}^-$.

5.3.2 Argument Factor Encoder

This section presents how we capture topic and discourse factors at the argument level. The superscript t is omitted for simplicity. As mentioned in Section 5.3.1, we employ latent variables $\boldsymbol{z}$ for argument topic factor representation, and $\boldsymbol{d}$ for discourse. The modeling process is inspired by [153] and based on variational auto-encoder (VAE) [62] to reconstruct a given argument in the BoW form, $\boldsymbol{x}_{BoW}$, conditioned on $\boldsymbol{z}$ and $\boldsymbol{d}$. Here $\boldsymbol{z}$ is the topic mixture and $\boldsymbol{d}$ is a one-hot vector denoting the discourse style.[4] Specifically, the generation process for each word $w_n \in \boldsymbol{x}_{BoW}$ is defined as:

$$\begin{aligned} &\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2), \quad \boldsymbol{z} = \text{softmax}(f_z(\boldsymbol{\epsilon})), \quad \boldsymbol{d} \sim Multi(\boldsymbol{\pi}), \\ &\beta_n = \text{softmax}(f_{\phi^T}(\boldsymbol{z}) + f_{\phi^D}(\boldsymbol{d})), \quad w_n \sim Multi(\beta_n), \end{aligned} \tag{5.1}$$

where $f_*(\cdot)$ is a neural perceptron that linearly transforms inputs. For both latent topic and discourse factors, we employ word distributions to represent them. Here we consider the weight matrix of $f_{\phi^T}(\cdot)$ (after the softmax normalization) as topic-word distributions, ϕ^T. Likewise, $f_{\phi^D}(\cdot)$'s weight matrix is used to compute the discourse-word distributions, ϕ^D.
For the other parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\pi}$, they can be learned from the input $\boldsymbol{x}_{BoW}$ following the formula below:

$$\begin{aligned} \boldsymbol{\mu} = f_\mu(\tanh(f_e(\boldsymbol{x}_{BoW}))), \log \boldsymbol{\sigma} = f_\sigma(\tanh(f_e(\boldsymbol{x}_{BoW}))) \\ \boldsymbol{\pi} = \text{softmax}(f_\pi(\boldsymbol{x}_{BoW})). \end{aligned} \tag{5.2}$$

5.3.3 Dynamic Process Encoder

Based on the topic and discourse factors learned at the argument level, here we discuss how to capture their dynamic patterns

[4] We follow the setting of [153], and apply Gumbel-Softmax relaxation for $\boldsymbol{d}$.

in the persuasion process. Our dynamic process encoder is inspired by dynamic memory network (DMN) [64, 147, 159] and topic memory mechanism [155], where we capture the indicative dynamic topic and discourse factors to interpret why a conversation can result in successful persuasion.

To be more specific, memory weight $\boldsymbol{w}^t \in \mathbb{R}^{(K+D)}$ is defined as the concatenation of latent aspects $\boldsymbol{z}^t$ and $\boldsymbol{d}^t$:

$$\boldsymbol{w}^t = [\boldsymbol{z}^t; \boldsymbol{d}^t] \tag{5.3}$$

where $[\cdot;\cdot]$ represents the concatenation. Once we have the memory weight, DTDMN will retrieve and update the memory according to the memory weight and input argument. Here we employ an attentive RNN to encode the the index sequence vector input $\boldsymbol{x}^t_{Seq}$ into the hidden state $\boldsymbol{H}^t$. Similar to [159], we employ a forget gate to erase the retrieved memory. The erase vector is denoted as $\boldsymbol{e}^t \in \mathbb{R}^E$, where E is the dimension of memory embeddings. Afterwards, an augment gate is used to strengthen the retrieved memory. The augment vector is denoted as $\boldsymbol{a}^t \in \mathbb{R}^E$. The overall update formulae for episodic memory are:

$$\begin{aligned} \boldsymbol{M}^t_i &= \boldsymbol{M}^{t-1}_i[\mathbf{1} - w^t_i \boldsymbol{e}^t] + w^t_i \boldsymbol{a}^t \\ \boldsymbol{e}^t &= \text{sigmoid}(f_e(\boldsymbol{E}^t)), \quad \boldsymbol{a}^t = \tanh(f_a(\boldsymbol{E}^t)) \end{aligned} \tag{5.4}$$

where $\boldsymbol{M}^t_i \in \mathbb{R}^E$ is the i-th row of the memory matrix $\boldsymbol{M}^t$, $\mathbf{1}$ is a row-vector of all 1s. The read content $\boldsymbol{r}^t \in \mathbb{R}^E$ of the episodic memory $\boldsymbol{M}_t$ is the weighted sum of the memory matrix:

$$\boldsymbol{r}^t = \sum_{i=1}^{K+D} w^t_i \boldsymbol{M}^t_i \tag{5.5}$$

5.3.4 Persuasiveness Predictor

For each conversation, DTDMN summarizes the read contents of all the arguments in a conversation $\{\boldsymbol{r}^t\}$, $t = 1, \ldots, T$ into $\boldsymbol{r}$ via an attentive RNN. Then we map $\boldsymbol{r}$ to a score y via a neural perceptron layer.

$$y = f_r(\boldsymbol{r}) \tag{5.6}$$

5.3.5 Learning Objective

Argument Factor Learning. To model topic and discourse factors, in learning, we maximize the variational lower bound $\mathcal{L}_z$ for $\mathbf{z}$ and $\mathcal{L}_d$ for $\mathbf{d}$. The corresponding functions are defined as:

$$\begin{aligned} \mathcal{L}_z &= \mathbb{E}_{q(\boldsymbol{z}\,|\,\boldsymbol{x})}[p(\boldsymbol{x}\,|\,\boldsymbol{z})] - D_{KL}(q(\boldsymbol{z}\,|\,\boldsymbol{x})\,||\,p(\boldsymbol{z})) \\ \mathcal{L}_d &= \mathbb{E}_{q(\boldsymbol{d}\,|\,\boldsymbol{x})}[p(\boldsymbol{x}\,|\,\boldsymbol{d})] - D_{KL}(q(\boldsymbol{d}\,|\,\boldsymbol{x})\,||\,p(\boldsymbol{d})) \end{aligned} \tag{5.7}$$

where $p(\boldsymbol{z})$ is the standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\boldsymbol{d})$ the uniform distribution $Unif(0, 1)$. $q(\boldsymbol{z}\,|\,\boldsymbol{x})$ and $q(\boldsymbol{d}\,|\,\boldsymbol{x})$ are posterior probabilities to approximate how $\boldsymbol{z}$ and $\boldsymbol{d}$ are generated from the arguments. $p(\boldsymbol{x}\,|\,\boldsymbol{z})$ and $p(\boldsymbol{x}\,|\,\boldsymbol{d})$ represent the corpus likelihoods conditioned on these topic and discourse factors.
The overall argument factor learning objective is to maxmize:

$$\mathcal{L}_{Factor} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda \mathcal{L}_{MI} \tag{5.8}$$

where $\mathcal{L}_x$ is for reconstructing the argument $\boldsymbol{x}$ from $\boldsymbol{z}$ and $\boldsymbol{d}$, $\mathcal{L}_{MI}$ is the mutual information (MI) penalty. The hyperparameter λ is the trade-off parameter for balancing between the $\mathcal{L}_{MI}$ and the other learning objectives. We leave out the details and refer the readers to [153].

Persuasiveness Prediction Learning. In our setting, we aim to identify which conversation is more persuasive given an input

of two conversations. Therefore, our goal is to have $\boldsymbol{C}^+$ scored higher than $\boldsymbol{C}^-$. We apply the pairwise cross-entropy loss to maximize the score margin of y^+ and y^- for $\boldsymbol{C}^+$ and $\boldsymbol{C}^-$, which is equivalent to minimize:

$$\mathcal{L}_{Pred} = \log(1 + \exp(y^- - y^+)) \tag{5.9}$$

Overall learning Objective. The three components of our model can be jointly optimized by minimizing the following objective function:

$$\mathcal{L} = \mathcal{L}_{Pred} - \sum_t (\mathcal{L}^t_{Factor}) \tag{5.10}$$

where $\mathcal{L}^t_{Factor}$ is for argument level.

5.4 Experimental Setup

Data Preprocessing. We randomly split the dataset with 80% for training and 20% for test. Then, 20% of the training data is randomly selected for validation. For preprocessing, we take the the following steps. First, non-English terms were filtered out. Then, quotations, digits and links were replaced with generic tags '⟨quote⟩', '⟨digit⟩', and '⟨url⟩', respectively. Next, we employed the natural language toolkit (NLTK) for tokenization[5]. After that, all letters were converted to lowercase. Finally, words occurred less than 10 times were filtered out from the data.

Parameter Setting. We use Gated Recurrent Unit (GRU) as the RNN cell. The hidden size of GRU is set to 512 with the word dropout rate of 0.2. The dimensions of word embeddings and memory embeddings are both set to 200. $\lambda = 0.01$ following

[5] `https://www.nltk.org/`

the setting of [153] for balancing the MI loss. For all the other hyperparameters, we tune them on the development set by grid search. Optimization is performed using Adam [61]. In the learning process, we alternatively update the parameters of the argument factor encoder and the rest of our model. We run our model for 80 epochs with early-stop strategy applied [20].

Comparison Baselines. LR-TFIDF [126] uses logistic regression with Bag-of-Words features in the pairwise pervasiveness prediction tasks, achieving good performance when compared with most of the handcrafted features. Here we implemented logistic regression with Tfidf-weighted n-grams features. Similar to [126], we adopt ℓ_1 regularization on the training stage to avoid overfitting. Joint topic-discourse model (JTDM) extracts topics and discourse features in an unsupervised way and can be used to place our argument factor encoder. We use the mean of each argument's topic-discourse mixture as the feature of an input conversation without considering the dynamics. Hierarchical attention recursive neural network (HATT-RNN) [150] uses bi-directional GRU as sequence encoder, including two levels of attention mechanisms (i.e., word level and argument level) while constructing the representation of a conversation. Dynamic memory network (DMN) [64] is a neural sequence model that can encode the contextual history into the episodic memory component. Dynamic key-value memory network (DKVMN) [159] improves upon DMN using one static matrix as key to compute the memory reading weights and one dynamic matrix as value for updating the memory states.

Models	CMV		Court	
	Acc.	F1	Acc.	F1
Baselines				
LR-TFIDF	0.571	0.727	0.631	0.773
JTDM	0.615	0.762	0.642	0.782
HATT-RNN	0.828	0.890	0.559	0.717
DMN	0.858	0.893	0.662	0.755
DKVMN	0.896	0.911	0.726	0.841
Our models				
W/O TOPIC	0.871	0.931	0.797	0.887
W/O DISCOURSE	0.922	0.959	0.821	0.902
W/O MEMORY	0.885	0.918	0.761	0.864
FULL MODEL	**0.939**	**0.968**	**0.833**	**0.909**

Table 5.2: Pairwise classification results on persuasiveness prediction. Best results in **bold**. Our FULL MODEL achieves significantly better results than all the baselines ($p < 0.01$, paired test).

5.5 Experimental Results

This section presents the how models perform on persuasiveness prediction. The further discussions on the effects of topics and discourse will be given in Section 5.6.

5.5.1 Persuasiveness Prediction Comparison

We follow Tan et al. [126] to conduct pairwise classification. For the CMV dataset, we predict which conversation can win Δ, and for the Court dataset, which side will win the case. In Table 5.2, we report the pairwise accuracy and F1 scores. For our models, we also display results without considering topic, discourse, and memory structures, respectively. It is observed that:

• ***Topic and discourse factors are useful.*** By exploiting pre-learned latent topic and discourse factors, JTDM outper-

forms LR-TFIDF baseline on both datasets. It even performs better than HATT-RNN on Court debates. This observation implies that topic and discourse factors can be indicative of persuasiveness arguments.

• ***Neural models generally outperform the non-neural baselines.*** This indicates that neural models are able to learn deep persuasiveness features. We also find that the improvement upon non-neural models is less significant on the Court dataset compared to CMV. This may be partly attributed to the sparse training instances in the Court dataset as shown in Table 5.1, which may result in overfitting. Nevertheless, our models can well alleviate such sparsity and achieve significantly better performance on both datasets.

• ***Processing modeling is important to predict argument persuasiveness.*** We observe that LR-TFIDF and JTDM, with only word features encoded, perform worse compared to other methods that explore dynamic patterns in argumentation process. This shows that persuasion outcomes are also dependent on a dynamic process beyond word features.

• ***Dynamic memory mechanism is effective.*** Our FULL MODEL obtains better results than its W/O MEMORY variant. Also, DMN and DKVMN outperform other baselines without dynamic memory mechanism. The above observations indicate that dynamic memory mechanism can help to capture indicative signals from the persuasion progress.

• ***Both dynamic topic and discourse factors contribute to argument persuasiveness.*** It is observed that our FULL MODEL achieves better results than the W/O TOPIC and W/O DISCOURSE ablation, which considers only dynamic discourse or

topic factors. Though the slightly better performance of W/O DISCOURSE than W/O TOPIC shows that topic factors might contribute more to argument persuasiveness, coupling the topics and discourse exhibiting the best performance.

5.5.2 Parameter Analysis

Here we study how the two important hyper-parameters in our model, the number of topics (K) and the number of discourse roles (D) affect our model performance. In Figure 5.3, we show the persuasiveness prediction accuracy given varying K in (a) and varying D in (b).

As can be seen, for both topic and discourse, the curves corresponding model performance are not monotonic. In particular, better accuracies are achieved given relatively larger topic numbers for CMV with the best result observed at $K = 50$. While for Court, the optimum topic number is $K = 20$. This may be due to the relatively more centralized topics in Court debates, whereas wider range of topics discussed in social media, CMV. For discourse, we observed a similar trend in both CMV and Court datasets. The best score is achieved when $D = 10$ for CMV and $D = 8$ for Court dataset. This implies that the discourse styles used in both CMV and Court are somewhat limited. We will later summarize the discourse styles useful to persuasiveness in Section 5.6.2.

5.6 Discussion on Topics and Discourse

In Section 5.5, we show the superior performance of our proposed model on persuasiveness prediction. Here we investigate the reasons behind.

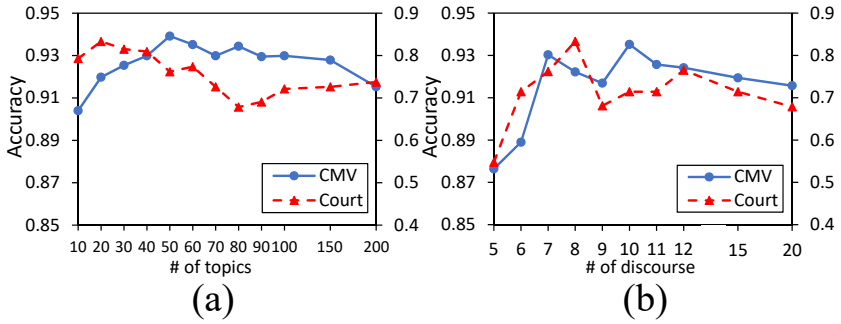


Figure 5.3: The impact of topic number (a) and discourse number (b) on our model for persuasiveness prediction. For both (a) and (b), the blue and solid line shows the results on CMV with left vertical axis, and the red and dashed line Court with the right vertical axis.

5.6.1 Analysis of the Persuasiveness Process

As discussed before, our output can be used to figure out what results in a good persuasion. Here we take the CMV conversation in Figure 5.1 as an example to look into its persuasion process. Recall that the challengers put forward viewpoints centered around "*the advantage to learn a second language*", and they successfully change the opinion holder's mind with good arguments delivered. In Figure 5.4(a), we visualize the dynamic memory weights $\boldsymbol{w}^t$ (see Eq. 5.3) for each turn. It is observed that our model highlights the '*cognition*' topic factor, which suggests the cognitive research evidence (e.g., learning a musical instrument) might help challengers win. For discourse, the model highlights latent factors represented by words like '⟨url⟩', '⟨digits⟩', and '*more*'. This shows effective discourse styles, such as reference to external URLs ('⟨url⟩') and statistics ('⟨digits⟩'), may also play an important role in persuasiveness.

To further study how each topic and discourse alone contributes to this example's persuasion, we disable the effects from other

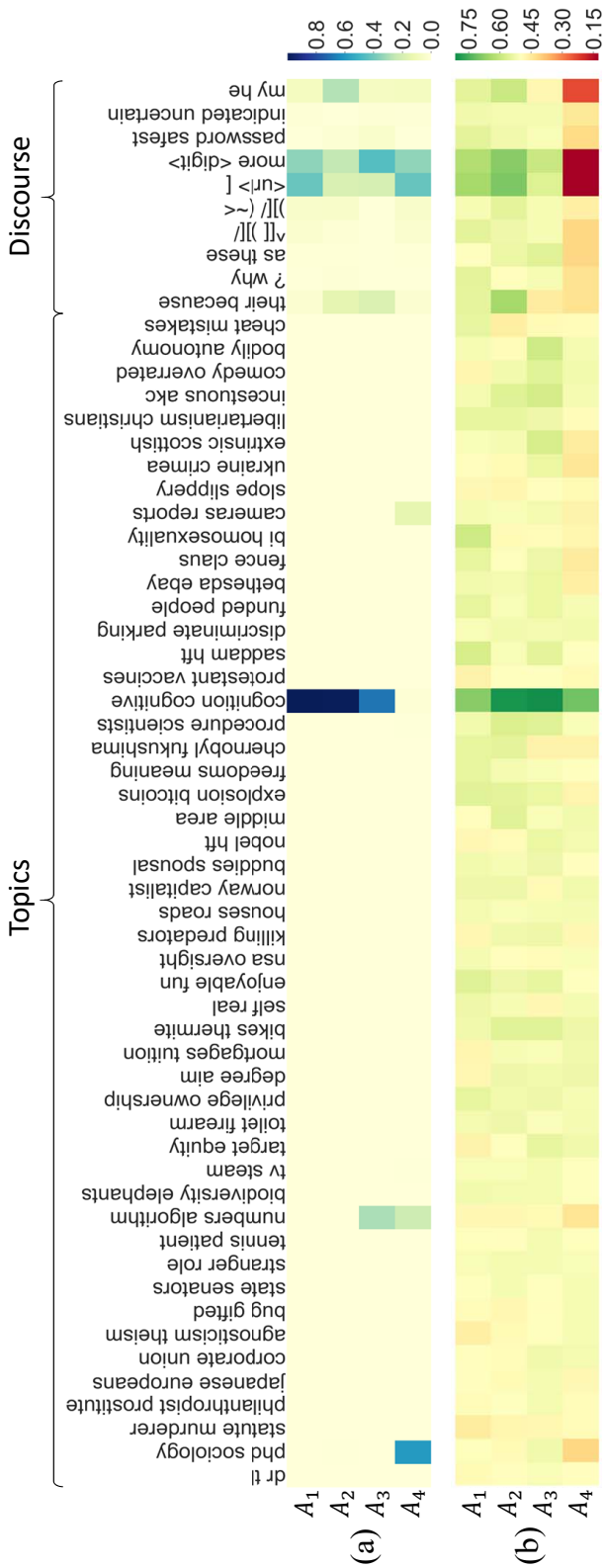


Figure 5.4: The heatmap visualizing the dynamic memory weights and persuasiveness on topic and discourse factors for the conversation in Figure 5.1. The vertical axis shows the turn id (from A_1 to A_4), and horizontal axis shows the latent topics and discourse displayed with their top 2 words. (a) Dynamic memory weights $\boldsymbol{w}^t$ that indicate topics shift and discourse flow. (b) Persuasiveness effect from each topic or discourse. For (a) and (b), darker colors indicate higher impacts. For (b), green indicate positive impacts while red negative.

topics and discourse via masking $\boldsymbol{w}^t$, and map the prediction score y in Eq. 5.6 to $[0, 1]$ range. We visualize the prediction scores in Figure 5.4(b) to depict the effect of persuasiveness from each topic and discourse. We observe that the "*cognition*" topic is still highlighted for all turns. It implies our model still recognize this topic to be important, without taking the discourse effects into account. For discourse, we notice that the *reference* and *statistics* behaviors considered useful for the first few turns, whose impacts however later changed to be negative. It might be because people tend to be tired of excessive URL links and statistics without providing more insightful opinions or content.

5.6.2 Roles of Topics and Discourse

In Section 5.5.1, we have shown that topics contribute slightly more on persuasiveness than discourse. Here we further analyze their roles in affecting persuasion outcome and similar trends are observed on both datasets. Due to space limitation, we only discuss the results on CMV dataset.

To investigate topic effects, we follow [134] to identify *strong* argument topics when the topic likelihood is larger than a pre-defined threshold (set to 0.2 here).[6] Then in Figure 5.5(a) we show how the number of strong argument topics distribute on winning arguments compared with the losing ones. For discourse, we show the discourse factor distribution on winning and losing arguments in Figure 5.5(b). Here we display the discourse factors with our interpretations on the discourse styles

[6]To set the threshold, we first sample 100 arguments and manually align them to the strongly related latent topic. Then we set the threshold resulting in the smallest errors compared with human annotations.

according to their word distributions. In the following we discuss the findings from topic and discourse in turn.

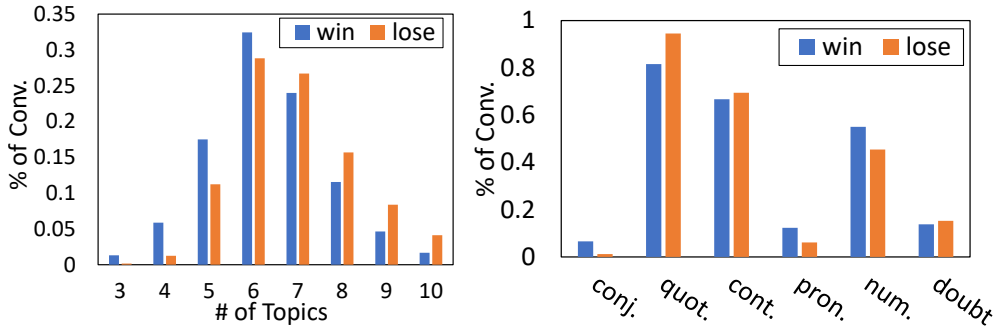
Topics. As can be seen in Figure 5.5(a), the winning side tends to put forward fewer topics in the argumentative process. This indicates that strong and focused argument points are better than diverse topics, since arguing with too many perspectives might overwhelm the opinion holder, which may lead to the persuasion failure.

Discourse. From Figure 5.5(b), we can see discourse styles vary in their contributions on the persuasiveness results. Specifically, personal pronoun and numbers are more likely to appear in the winning side than the losing side. Their positive effects have also been previously reported [126, 134]. Moreover, we find that conjunction, though not widely used, is obviously more endorsed by winning sides. The benefit of conjunction may result from the better logic it renders. For the losing side, they are more in favor of the quotation discourse, which is used in CMV to quote and attack others' weak points. People may dislike such criticism, which renders the negative impact on persuasiveness.

5.6.3 Implications

From this work, we distill some general suggestions on argumentation which are beyond specific task and approach.

1) Topics are more important than discourse styles. In an argumentative conversation, opponents attempt to establish the validity of two positions by convincing each other and trying to win points in the debate [121]. Our study shows that topics contribute slightly more on persuasiveness than discourse. This



(a) Topic (b) Discourse

Figure 5.5: Distributions of winning and losing persuasion over the number of *strong* argument topics involved in (a) and varying discourse factors in (b). For (b), we display the discourse factors with our interpretation on them ("conj."-conjunction, "quot."-quotation mark, "cont."-contrast, "pron."-personal pronoun, and "num."-number). Two-sided Mann-Whitney rank test shows that the two distributions shown here are significantly different for both sides ($p < 0.01$).

phenomenon has also been mentioned by [129], where they find that discourse strategy is not the dominant factor in the debate: Topicality must be considered.

2) Strong and focused argument points are better than diverse topics. Strong arguments which are well-supported with evidence and/or reasoning, generally deliver more persuasive messages to audience. Our study reveals that successful argumentation usually convey fewer and focused topics. Diverse topics could only distract audience and expose more vulnerable points to the opponent.

3) Well organize the points and address them in a modest and concrete way. Discourse Argument discourse represents the cultural and situational realities of human reasoning, and is more sensitive to audience in conversational debates [33]. [5] also claims that argumentativity constitutes an inherent feature of discourse.

5.7 Summary

In this chapter, we propose to dynamically track both topics and discourse factors in conversational argumentation for persuasiveness prediction. The proposed neural model not only identifies persuasive arguments more accurately, but also provides insights into the usefulness of topics and discourse for a successful persuasion. The findings concluded in this chapter can facilitate future argument persuasiveness analysis.

□ End of chapter.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The growing popularity of social media results in large volume of user-created text data. Understanding such social media text data is very important and critical to many real-life applications, e.g., event detection, user profiling. However, analyzing such large volume of short and noisy text data is very challenging for many language understanding tasks. In this thesis, we propose to model the latent topics and discourse on social media text in an unsupervised way. By modeling the latent variables, we design models for two social media text understating tasks, short text classification and argumentation mining, and demonstrate their superior performance when compared with conventional methods.

In particular, in Chapter 3, we presented a neural framework that jointly explores topic and discourse from microblog conversations. Our model, in an unsupervised manner, examines the conversation contexts and discovers word distributions that reflect latent topics and discourse roles. Results from extensive experiments show that our model can generate coherent topics

and meaningful discourse roles. In addition, our model can be easily combined with other neural network architectures (such as CNN) and allows for joint training, which has presented better message classification results compared to the pipeline approach without joint training.

In Chapter 4, we present topic memory networks that exploit corpus-level topic representations with a topic memory mechanism for short text classification. The model alleviates data sparsity issues via jointly learning latent topics and text categories. Empirical comparisons with state-of-the-art models on four benchmark datasets have demonstrated the validity and effectiveness of our model, where better results have been achieved on both short text classification and topic coherence evaluation.

In Chapter 5, we propose a neural model to dynamically extract and track both topics and discourse factors in conversational argumentation for persuasiveness prediction. The proposed neural model not only identifies persuasive arguments more accurately, but also provides insights into the usefulness of topics and discourse for a successful persuasion. The findings concluded can facilitate future argument persuasiveness analysis.

In summary, we design novel techniques to automatically analysis social media text by modeling the latent topic and latent discourse. We are the first to leverage neural network to jointly model topics and discourse in conversation, in an unsupervised manner. Our method can be generalizable, for example in short text classification without the conversation information, our integration model demonstrates a superior performance by exploiting corpus-level topic representations. Moreover, our method can be extensible, for example in the challenging task of

persuasiveness tracing. The integration model can dynamically extract and track both topics and discourse factors in conversational argumentation for persuasiveness prediction, and achieve a significant performance gain compared with previous methods.

6.2 Future Work

System reliability management via automatic runtime data analysis has been widely studied in recent years, and it is a promising research topic. Although we have proposed a number of novel techniques that advance the state-of-the-art solutions, there are still many interesting research directions which are considered as future work.

6.2.1 Joint Modeling Topic, Discourse and Sentiment in Microblog Conversation

Sentiment analysis aims to study the sentiment polarity, such as "positive" or "negative", over a piece of text. Sentiment analysis has lots of applications in social media, such as public opinion tracing and user profiling. Traditional sentiment analysis heavily relies on high-quality labeled corpus, which is not always easy to obtain and often domain-specific in practical applications. Much work has been done for addressing the above issues through unsupervised or weak-supervised modeling for sentiment analysis in various granularity (e.g., word/phrase, aspect, sentence and document) [128, 59, 18]. Intuitively, sentiment polarities are not independent with the topics and discourse in the message. For example, the adjective word "wonderful", typically thought as positive orientation, might have negative orientation in messages about Trump with sar-

casm tone. Therefore, jointly modeling sentiment with topics and discourse in microblog conversation can bring benefits for understanding both factors.

To fill the gap, we plan to explore joint modeling topic, discourse, and sentiment in microblog conversation in an unsupervised neural framework. We will try to design a mechanism that can separate the representation of sentiment, topics, and discourse in an unsupervised framework. We can also incorporate the conversation tree and hashtag into our framework for better modeling the latent variables. For the evaluation, we can use the SemEval dataset, which provides large quantity of labeled sentiment orientation Twitter messages.

6.2.2 Unsupervised Microblog Conversation Summarization

Text summarization techniques have been widely applied to many real-life applications, like Baidu Baike [1]. With the flourish of social media, microblog such as Twitter, Weibo, has become an important channel for people to acquire the latest information. For users of microblog, there is a pressing need for automatically summarizing the key points of microblog conversations. However, due to the data sparsity issues and the lack of ground-truth labeled data of social media text, most of the existing summarization system works poorly for microblog messages.

Previous work for unsupervised text summarization employs graph-based or integer programming methods to maximize the "coverage" of original text [85, 145], which did not consider the semantic effects of topics and discourse in microblog conver-

[1] baike.baidu.com

sation. In our future work, we will explore the unsupervised microblog conversation summarization through modeling latent topics and discourse.

6.2.3 Topic and Discourse-Aware Social Chatbot

Nowadays, there is a surgent interest in developing intelligent open-domain dialog systems, i.e., social chatbot, due to the availability of large volume of conversational data and advanced neural network techniques. Commercial social chatbots, such as Microsoft XiaoIce, Amazon Alexa, have attracted millions of users and can converse with users on various topics for hours [114]. However, developing neural-based social chatbot still faces the challenges of understanding users' intentions and providing interactive responses [47]. In this thesis, we build up model that can captures topic and discourse representations embedded in conversations, which is useful for developing social chatbots [166]. By explicitly modeling "*what you say*" and "*how you say*", our model can be adapted to track the change of topics and user behaviors in conversation context, helpful to determine "*what to say and how to say*" in the next turn.

□ End of chapter.

Chapter 7

List of Publications

1. **Jichuan Zeng**, Jing Li, Yulan He, Cuiyun Gao, Michael R. Lyu, and Irwin King. *What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations.* IEEE Transactions of the Association for Computational Linguistics (TACL), orally presented in ACL, 2019.

2. Cuiyun Gao, **Jichuan Zeng**, Xin Xia, David Lo, Michael R. Lyu, and Irwin King. *Automating App Review Response Generation.* The 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2019.

3. Cuiyun Gao, Wujie Zheng, Yuetang Deng, David Lo, **Jichuan Zeng**, Michael R. Lyu, and Irwin King. *Emerging app issue identification from user feedback: Experience on WeChat.* The 41th ACM/IEEE International Conference on Software Engineering (ICSE), 2019.

4. **Jichuan Zeng**, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. *Topic Memory Network for Short Text Classification.* Empirical Methods in Natural Language Processing (EMNLP), oral long paper, 2018.

5. Cuiyun Gao, **Jichuan Zeng**, David Lo, Chin-yew Lin, Michael R. Lyu, and Irwin King. *INFAR: Insight Extraction from App Reviews.* The 26th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), demo track, 2018.

6. Cuiyun Gao, **Jichuan Zeng**, Federica Sarro, Michael R. lyu, and Irwin King. *Exploring the Effects of Ad Schemes on the Performance Cost of Mobile Phones*. The 1st International Workshop on Advances in Mobile App Analysis (A-Mobile), co-located with ASE, 2018.

7. Cuiyun Gao, **Jichuan Zeng**, Michael R. Lyu, and Irwin King. *Online App Review Analysis for Identifying Emerging Issues*. The 40th International Conference on Software Engineering (ICSE), 2018.

8. **Jichuan Zeng**, Haiqin Yang, Irwin King, and Michael R. Lyu. *A Comparison of Lasso-type Algorithms on Distributed Parallel Machine Learning Platforms*. Distributed Machine Learning and Matrix Computations Workshop, Annual Conference on Neural Information Processing Systems, 2014.

□ **End of chapter.**

Bibliography

[1] 2018 weibo user development report. https://data.weibo.com/report/reportDetail?id=433.

[2] Know your limit: The ideal length of every social media post. https://sproutsocial.com/insights/social-media-character-counter/.

[3] Sina weibo ends 140-character limit ahead of twitter. https://www.bbc.com/news/technology-35361157.

[4] D. Alvarez-Melis and M. Saveski. Topic modeling in twitter: Aggregating tweets by conversations. In *Proceedings of the Tenth International Conference on Web and Social Media, May 17-20, 2016.*, pages 519–522, Cologne, Germany, 2016.

[5] R. Amossy. Argumentation in discourse: A socio-discursive approach to arguments. *Informal Logic*, 29(3):252–267, 2009.

[6] S. J. Bakker and G. C. Wakker. *Discourse cohesion in ancient Greek*, volume 16. Brill, 2009.

[7] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[8] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems, NIPS 2003, December 8-13, 2003*, pages 17–24, Vancouver and Whistler, British Columbia, Canada, 2003.

[9] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28 - August 1, 2003, Toronto, Canada*, pages 127–134, 2003.

[10] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *CoRR*, abs/1601.00670, 2016.

[11] D. M. Blei and J. D. Lafferty. Correlated Topic Models. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005]*, Vancouver, British Columbia, Canada, pages 147–154, 2005.

[12] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006)*, Pittsburgh, Pennsylvania, USA, pages 113–120, 2006.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001]*,

pages 601–608, Vancouver, British Columbia, Canada, 2001.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[15] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *J. Comput. Science*, 2(1):1–8, 2011.

[16] G. Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[17] S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts. A Fast Unified Model for Parsing and Sentence Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*, Berlin, Germany, 2016.

[18] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 804–812, 2010.

[19] L. Carlson, D. Marcu, and M. E. Okurovsky. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1,*

2001 to Sunday, September 2, 2001, Aalborg, Denmark, 2001.

[20] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, an"d early stopping. In *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems, NIPS 2000*, Denver, CO, USA, pages 402–408, 2000.

[21] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. T. Le. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, August 20-26, 2018*, pages 745–754, Santa Fe, New Mexico, USA, 2018.

[22] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*, 2010.

[23] J. Chang and D. M. Blei. Relational topic models for document networks. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, pages 81–88, 2009.

[24] P. Chen, Z. Sun, L. Bing, and W. Yang. Recurrent Attention Network on Memory for Aspect Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, pages 452–461, 2017.

[25] W. W. Cohen, V. R. Carvalho, and T. M. Mitchell. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 309–316, 2004.

[26] N. T. Crook, R. Granell, and S. G. Pulman. Unsupervised classification of dialogue acts using a dirichlet process mixture model. In *Proceedings of the SIGDIAL 2009 Conference, The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 11-12 September 2009, London, UK*, pages 341–348, 2009.

[27] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. M. Kleinberg. Echoes of power: language effects and power differences in social interaction. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 699–708, 2012.

[28] R. Das, M. Zaheer, and C. Dyer. Gaussian LDA for Topic Models with Word Embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*, Beijing, China, pages 795–804, 2015.

[29] R. De Beaugrande and W. U. Dressler. *Introduction to text linguistics*. Routledge, 1981.

[30] A. B. Dieng, C. Wang, J. Gao, and J. W. Paisley. TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. *CoRR*, abs/1611.01702, 2016.

[31] C. N. dos Santos and M. Gatti. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, COLING 2014*, Dublin, Ireland, pages 69–78, 2014.

[32] Z. Dou. Capturing User and Product Information for Document Level Sentiment Analysis with Deep Memory Network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, pages 521–526, 2017.

[33] D. G. Ellis. Argument discourse. *The International Encyclopedia of Language and Social Interaction*, pages 1–6, 2015.

[34] V. W. Feng and G. Hirst. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 511–521, 2014.

[35] S. Goldwater and T. L. Griffiths. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association*

for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic, 2007.

[36] A. Graves, G. Wayne, and I. Danihelka. Neural Turing Machines. *CoRR*, abs/1410.5401, 2014.

[37] I. Habernal and I. Gurevych. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[38] Y. He. Extracting Topical Phrases from Clinical Documents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, pages 2957–2963, 2016.

[39] C. Hidey and K. R. McKeown. Persuasive influence detection: The role of argument sequencing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5173–5180, 2018.

[40] C. Hidey, E. Musi, A. Hwang, S. Muresan, and K. McKeown. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 11–21, 2017.

[41] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999*, pages 50–57, Berkeley, CA, USA, 1999.

[42] B. Hollerit, M. Kröll, and M. Strohmaier. Towards linking buyers and sellers: detecting commercial intent on twitter. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 629–632, 2013.

[43] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNAKDD 2009, June 28*, pages 80–88, Paris, France, 2010.

[44] E. H. Hovy and E. Maier. Parsimonious or profligate: How many and which discourse structure relations. *Discourse Processes*, 1997.

[45] B. Hu, Q. Chen, and F. Zhu. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, Lisbon, Portugal, pages 1967–1972, 2015.

[46] Z. Hu, G. Luo, M. Sachan, E. P. Xing, and Z. Nie. Grounding topic models with knowledge bases. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, 9-15 July 2016*, pages 1578–1584, New York, NY, USA, 2016.

[47] M. Huang, X. Zhu, and J. Gao. Challenges in building intelligent open-domain dialog systems. *CoRR*, abs/1905.05709, 2019.

[48] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *CoRR*, abs/1611.01144, 2016.

[49] L. Ji, Z. Wei, X. Hu, Y. Liu, Q. Zhang, and X. Huang. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3703–3714, 2018.

[50] Y. Ji and J. Eisenstein. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 13–24, 2014.

[51] Y. Ji, G. Haffari, and J. Eisenstein. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016, June 12-17*, pages 332–342, San Diego California, USA, 2016.

[52] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter Sentiment Classification. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings*

of the Conference, Portland, Oregon, USA, pages 151–160, 2011.

[53] Y. Jiao, C. Li, F. Wu, and Q. Mei. Find the conversation killers: A predictive study of thread-ending posts. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, April 23-27*, pages 1145–1154, Lyon, France, 2018.

[54] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang. Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011*, Glasgow, United Kingdom, pages 775–784, 2011.

[55] Y. Jo, S. Poddar, B. Jeon, Q. Shen, C. P. Rosé, and G. Neubig. Attentive interaction model: Modeling changes in view in argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 103–116, 2018.

[56] S. R. Joty, G. Carenini, and C. Lin. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011, July 16-22*, pages 1807–1813, Barcelona, Catalonia, Spain, 2011.

[57] S. R. Joty, G. Carenini, and R. T. Ng. A novel discriminative framework for sentence-level discourse analysis. In

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea, pages 904–915, 2012.

[58] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, April 2017.

[59] S. Kim and E. H. Hovy. Determining the sentiment of opinions. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

[60] Y. Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751, Doha, Qatar, 2014.

[61] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[62] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[63] J. Krishnamurthy, P. Dasigi, and M. Gardner. Neural Semantic Parsing with Type Constraints for Semi-Structured Tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, pages 1516–1526, 2017.

[64] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1378–1387, 2016.

[65] M. R. Leary, K. J. Diebels, E. K. Davisson, K. P. Jongman-Sereno, J. C. Isherwood, K. T. Raimi, S. A. Deffler, and R. H. Hoyle. Cognitive and interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, 43(6):793–813, 2017.

[66] J. Y. Lee and F. Dernoncourt. Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016*, San Diego California, USA, pages 515–520, 2016.

[67] C. Li, Y. Lu, Q. Mei, D. Wang, and S. Pandey. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1959–1968, 2015.

[68] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information*

Retrieval, SIGIR 2016, July 17-21, 2016, pages 165–174, Pisa, Italy, 2016.

[69] J. Li, R. Li, and E. H. Hovy. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 2061–2069, 2014.

[70] J. Li, M. Liao, W. Gao, Y. He, and K. Wong. Topic extraction from microblog posts using conversation structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Volume 1: Long Papers*, Berlin, Germany, 2016.

[71] J. Li, T. Luong, D. Jurafsky, and E. H. Hovy. When are tree structures necessary for deep learning of representations? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, September 17-21, 2015*, pages 2304–2314, Lisbon, Portugal, 2015.

[72] J. Li, Y. Song, Z. Wei, and K.-F. Wong. A joint model of conversational discourse and latent topics on microblogs. *Computational Linguistics*, 44(4):719–754, 2018.

[73] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 375–384, 2009.

[74] C. X. Lin, B. Zhao, Q. Mei, and J. Han. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, pages 929–938, 2010.

[75] Z. Lin, M. Kan, and H. T. Ng. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 343–351, 2009.

[76] R. V. Lindsey, W. Headden, and M. Stipicevic. A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012*, Jeju Island, Korea, pages 214–222, 2012.

[77] P. Liu, X. Qiu, and X. Huang. Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*, Vancouver, Canada, pages 1–10, 2017.

[78] W. Lucia and E. Ferrari. EgoCentric: Ego Networks for Knowledge-based Short Text Classification. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014*, Shanghai, China, pages 1079–1088, 2014.

[79] Y. Ma, H. Peng, and E. Cambria. Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, USA, 2018.

[80] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712, 2016.

[81] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988.

[82] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[83] D. Marcu. The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics*, pages 96–103, 1997.

[84] R. Mehrotra, S. Sanner, W. L. Buntine, and L. Xie. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13*, Dublin, Ireland, pages 889–892, 2013.

[85] Q. Mei, J. Guo, and D. R. Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010, pages 1009–1018, 2010.

[86] M. Meila. Comparing clusterings by the variation of information. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel, August 24-27, 2003, Proceedings*, pages 173–187, Washington, DC, USA, 2003.

[87] Y. Miao, E. Grefenstette, and P. Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, 6-11 August 2017*, pages 2410–2419, Sydney, NSW, Australia, 2017.

[88] Y. Miao, L. Yu, and P. Blunsom. Neural Variational Inference for Text Processing. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, New York City, NY, USA, pages 1727–1736, 2016.

[89] M. Moens, E. Boiy, R. M. Palau, and C. Reed. Automatic detection of arguments in legal texts. In *The Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*, pages 225–230, 2007.

[90] V. Nair and G. E. Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 807–814, Haifa, Israel, 2010.

[91] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, TACL*, 3:299–313, 2015.

[92] V. Niculae, J. Park, and C. Cardie. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 985–995, 2017.

[93] K. Nigam, A. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.

[94] A. Pak and P. Paroubek. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010*, Uppsala University, Uppsala, Sweden, pages 436–439, 2010.

[95] R. M. Palau and M. Moens. Argumentation mining: the detection, classification and structure of arguments in text. In *The 12th International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 8-12, 2009, Barcelona, Spain*, pages 98–107, 2009.

[96] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543, Doha, Qatar, 2014.

[97] J. Perret, S. D. Afantenos, N. Asher, and M. Morey. Integer linear programming for discourse parsing. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 99–109, 2016.

[98] I. Persing and V. Ng. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4082–4088, 2017.

[99] X. H. Phan, M. L. Nguyen, and S. Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *Proceedings of the 17th International Conference on World Wide Web, WWW 2008*, Beijing, China, pages 91–100, 2008.

[100] A. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1873–1876, 2010.

[101] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. The penn discourse treebank 2.0. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[102] K. Qin, L. Wang, and J. Kim. Joint modeling of content and discourse relations in dialogues. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, July 30 - August 4, Volume 1: Long Papers, pages 974–984, Vancouver, Canada, 2017.

[103] Z. Qin, Y. Cong, and T. Wan. Topic Modeling of Chinese Language beyond a Bag-of-Words. *Computer Speech & Language*, 40:60–78, 2016.

[104] X. Qiu, Q. Zhang, and X. Huang. FudanNLP: A Toolkit for Chinese Natural Language Processing. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2013.

[105] X. Quan, C. Kit, Y. Ge, and S. J. Pan. Short and Sparse Text Topic Modeling via Self-Aggregation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, Buenos Aires, Argentina, pages 2270–2276, 2015.

[106] D. Ramage, S. T. Dumais, and D. J. Liebling. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, May 23-26*, Washington, DC, USA, 2010.

[107] Y. Ren, Y. Zhang, M. Zhang, and D. Ji. Improving Twitter Sentiment Classification Using Topic-Enriched Multi-Prototype Word Embeddings. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, USA, pages 3038–3044, 2016.

[108] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International*

Conference on Machine Learning, ICML 2014, 21-26 June 2014, pages 1278–1286, Beijing, China, 2014.

[109] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010,*, pages 172–180, Los Angeles, California, USA, 2010.

[110] M. Röder, A. Both, and A. Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015*, pages 399–408, Shanghai, China, 2015.

[111] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, July 7-11, 2004*, pages 487–494, Banff, Canada, 2004.

[112] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2007, June 28-30*, pages 410–420, Prague, Czech Republic, 2007.

[113] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information*

Retrieval, August 7-11, 2017, pages 375–384, Shinjuku, Tokyo, Japan, 2017.

[114] H. Shum, X. He, and D. Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of IT & EE*, 19(1):10–26, 2018.

[115] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011, July 16-22*, pages 2330–2336, Barcelona, Catalonia, Spain, 2011.

[116] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003*, 2003.

[117] J. Spencer and G. Uchyigit. Sentimentor: Sentiment analysis of twitter data. In *Proceedings of the 1st International Workshop on Sentiment Discovery from Affective Data, SDAD@ECML/PKDD 2012, Bristol, UK, September 28, 2012*, pages 56–66, 2012.

[118] A. Srivastava and C. Sutton. Autoencoding Variational Inference For Topic Models. *arXiv preprint arXiv:1703.01488*, 2017.

[119] A. Srivastava and C. Sutton. Autoencoding variational inference for topic models. In *Proceedings of the Fifth International Conference on Learning Representations, ICLR 2017, 24-26, April 2017*, Toulon, France, 2017.

[120] C. Stab and I. Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 46–56, 2014.

[121] N. L. Stein and E. R. Albro. The origins and nature of arguments: Studies in conflict understanding, emotion, and negotiation. *Discourse processes*, 32(2-3):113–133, 2001.

[122] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[123] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. V. Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3), 2000.

[124] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. A. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR*, cs.CL/0006023, 2000.

[125] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, NIPS 2015*, Montreal, Quebec, Canada, pages 2440–2448, 2015.

[126] C. Tan, V. Niculae, C. Danescu-Niculescu-Mizil, and L. Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624, 2016.

[127] H. L. Thanh, G. Abeysinghe, and C. R. Huyck. Generating discourse structures for written text. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*, 2004.

[128] P. D. Turney and M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *CoRR*, cs.LG/0212012, 2002.

[129] T. A. Van Dijk, W. Kintsch, and T. A. Van Dijk. Strategies of discourse comprehension. 1983.

[130] D. Vitale, P. Ferragina, and U. Scaiella. Classification of Short Texts by Deploying Topical Annotations. In *Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012*, Barcelona, Spain, pages 376–387, 2012.

[131] M. A. Walker, P. Anand, S. M. Lukin, and S. Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 742–753, 2017.

[132] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 115–120, 2012.

[133] J. Wang, Z. Wang, D. Zhang, and J. Yan. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, Melbourne, Australia, pages 2915–2921, 2017.

[134] L. Wang, N. Beauchamp, S. Shugars, and K. Qin. Winning on the merits: The joint effects of content and style on debate outcomes. *TACL*, 5:219–232, 2017.

[135] S. Wang, J. Zhang, and C. Zong. Exploiting Word Internal Structures for Generic Chinese Sentence Representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, pages 298–303, 2017.

[136] S. I. Wang and C. D. Manning. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, ACL 2012, Volume 2: Short Papers*, Jeju Island, Korea, pages 90–94, 2012.

[137] Z. Wei, Y. Liu, and Y. Li. Is this post persuasive? ranking argumentative comments in online forum. In

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers, 2016.

[138] J. Weng and B. Lee. Event detection in twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*, 2011.

[139] J. Weng, E. Lim, J. Jiang, and Q. He. TwitterRank: Finding Topic-Sensitive Influential Twitterers. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010*, New York, NY, USA, pages 261–270, 2010.

[140] J. Weston, S. Chopra, and A. Bordes. Memory Networks. *CoRR*, abs/1410.3916, 2014.

[141] C. A. Willard. *Liberalism and the problem of knowledge: A new rhetoric for modern democracy.* University of Chicago Press, 1996.

[142] T. Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

[143] N. Witt, C. Seifert, and M. Granitzer. Explaining Topical Distances Using Word Embeddings. In *27th International Workshop on Database and Expert Systems Applications, DEXA 2016 Workshops*, Porto, Portugal, pages 212–217, 2016.

[144] F. Wolf and E. Gibson. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287, 2005.

[145] K. Woodsend and M. Lapata. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 233–243, 2012.

[146] Y. Xiao and K. Cho. Efficient Character-level Document Classification by Combining Convolution and Recurrent Layers. *CoRR*, abs/1602.00367, 2016.

[147] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2397–2406, 2016.

[148] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *22nd International World Wide Web Conference, WWW '13, May 13-17, 2013*, pages 1445–1456, Rio de Janeiro, Brazil, 2013.

[149] Y. Yang, D. Downey, and J. L. Boyd-Graber. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, September 17-21, 2015*, pages 308–317, Lisbon, Portugal, 2015.

[150] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy. Hierarchical attention networks for document classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 1480–1489, 2016.

[151] E. Zarisheva and T. Scheffler. Dialog act annotation for twitter conversations. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015*, pages 114–123, Prague, Czech Republic, 2015.

[152] J. Zeng, J. Li, Y. He, C. Gao, M. R. Lyu, and I. King. What you say and how you say it: Joint modeling of topics and discourse in microblog conversations. *CoRR*, abs/1903.07319, 2019.

[153] J. Zeng, J. Li, Y. He, C. Gao, M. R. Lyu, and I. King. What you say and how you say it: Joint modeling of topics and discourse in microblog conversations. *Arxiv Preprint*, abs/1903.07319, 2019.

[154] J. Zeng, J. Li, Y. Song, C. Gao, M. R. Lyu, and I. King. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, November 31-4, 2018*, Brussels, Belgium, 2018.

[155] J. Zeng, J. Li, Y. Song, C. Gao, M. R. Lyu, and I. King. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3120–3131, 2018.

[156] X. Zeng, J. Li, L. Wang, N. Beauchamp, S. Shugars, and K. Wong. Microblog conversation recommendation via

joint modeling of topics and discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, June 1-6, 2018, Volume 1 (Long Papers)*, pages 375–385, New Orleans, Louisiana, USA, 2018.

[157] D. Zhang and D. Wang. Relation Classification via Recurrent Neural Network. *CoRR*, abs/1508.01006, 2015.

[158] J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil. Conversational flow in oxford-style debates. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 136–141, 2016.

[159] J. Zhang, X. Shi, I. King, and D. Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 765–774, 2017.

[160] S. Zhang, D. Zheng, X. Hu, and M. Yang. Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29*, Shanghai, China, 2015.

[161] W. Zhang, D. Wang, G. Xue, and H. Zha. Advertising Keywords Recommendation for Short-Text Web Pages Using Wikipedia. *ACM TIST*, 3(2):36:1–36:25, 2012.

[162] X. Zhang, J. J. Zhao, and Y. LeCun. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, NIPS 2015*, Montreal, Quebec, Canada, pages 649–657, 2015.

[163] Y. Zhang, J. Li, Y. Song, and C. Zhang. Encoding conversation context for neural keyphrase extraction from microblog posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, New Orleans, Louisiana, USA, pages 1676–1686, 2018.

[164] T. Zhao, K. Lee, and M. Eskénazi. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, July 15-20, 2018, Volume 1: Long Papers*, pages 1098–1107, Melbourne, Australia, 2018.

[165] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, April 18-21, 2011. Proceedings*, pages 338–349, Dublin, Ireland, 2011.

[166] L. Zhou, J. Gao, D. Li, and H. Shum. The design and implementation of xiaoice, an empathetic social chatbot. *CoRR*, abs/1812.08989, 2018.

[167] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 1257–1264, 2009.