学号：__2015302580288__

密级：_____

武汉大学本科毕业论文

基于情感分析的社交媒体虚假信息检测

院(系)名 称：　计算机学院

专 业 名 称：　软件工程

学 生 姓 名：　张雪遥

指 导 教 师：　贾向阳

二〇一九年五月

郑 重 声 明

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名: _____ 日期: _____

摘　　要

微博以其便利性和低成本，已逐渐发展成为人们发布、分享和寻求信息的热门社交平台。然而，微博平台中发布的大量的新闻事件，其中也包含了很多虚假信息。虚假新闻的产生与传播，极有可能导致十分严重的社会后果，因此在微博等社交媒体平台上检测虚假新闻对于社会利益非常重要。

已有大量的社会心理学研究表明，情感因素对于社交平台中信息的传播具有极其显著的作用。在本文中，我们首先定义出社交媒体中情感信号所呈现出的两种表现形式，其包括发布者的情感与社交群体的情感，并对这两种情感形式产生的原因作出了一定的阐释。在此假设的基础上，本研究采用真实环境中的微博数据集，从三个角度分析了虚假新闻的情感信号的分布特点：情感种类的分布，情感强度的特性，以及情感词的表达。实验结果表明：虚假新闻中所含的愤怒、悲伤、质疑等负面情绪更多、情感强度更为激烈；且社交媒体用户在表达同一种情绪时，对待虚假新闻的言语和用词常常更为夸张与极端。

针对上述数据挖掘工作的发现，本研究提出了一种新型的基于情感的虚假信息检测模型（Emotion-based Fake News Detection Framework，EFN），它可以同步地挖掘利用新闻文本内容、用户评论内容中的情感信号，具体地：（1）模型借鉴了情感分析领域最先进的情感嵌入表示（Emotion Embedding）技术，并通过预训练任务的设计，使情感嵌入向量的编码富含情感种类的抽象信息；（2）模型注重了对于情感强度信号的建模，构建了程度词、否定词、问号、感叹号等统计情感特征，来提高对文本情感强度的利用；（3）模型融合语义编码和情感编码时，通过门机制（Gate Mechanism）的思想，使模型能够自适应地学到对这两种编码的利用倾向。

最后，本研究详细评估了此模型在真实环境的微博数据集下的性能体现。结果表明：本文提出的 EFN 模型，比虚假信息检测领域最先进的深度模型拥有 4% 的 F1 值提高。除此之外，本文还设计了多角度的评估实验，验证了 EFN 模型所使用的两种关键性技术（情感嵌入表示、门融合机制）的有效性。

关键词：虚假信息；社交媒体计算；文本挖掘；情感分析

ABSTRACT

Weibo has gradually developed into a popular social platform for people to publish, share and seek information. However, a large number of news events on *Weibo* also contain a lot of false information. The dissemination of false news is very likely to lead to very serious social consequences. Therefore, detecting fake news on social media platforms such as Weibo is very essential to public interests.

Longstanding research of social psychology has shown that emotion plays an extremely significant role in the dissemination of information. In this study we first summarizes two forms of emotional expression in social platforms, *publisher emotion* and *social emotion*. Based on that, we analyze the distribution characteristics of the emotional signals of false news by emotional categories, emotional intensity, and emotional expression. The results show that fake news contains more negative emotions such as anger, sadness and doubt, and their emotional intensity is more intense and extreme. Besides, users tend to express more exaggerated and fierce words towards fake news.

The discovery of the above motivates us to propose *Emotion-based Fake News Detection Framework (EFN)*, which can synchronously exploit the emotional and semantic signals in the content of news and user comments. Specifically, (1) *EFN* adopt *Emotion Embedding* to obtain emotional embedded vector that is encoded with high-level information rich in emotional categories through the design of the pre-trained task; (2) *EFN* focuses on emotional intensity and constructs statistical emotion features based on such as degree words, negative words, question marks and exclamation marks. (3) *Gate Mechanism* is applied in *EFN* to fuse semantic encoder and emotional coder, which enables the model to adaptively learn the tendency to use the two respectively.

Finally, we evaluates *EFN* with the real-world *Weibo*'s data sets. The results show that *EFN* increases F1-score by 4% comparing to the state-of-art deep models. In addition, other evaluation experiments verify the effectiveness of the two key technologies (*Emotion Embedding* and *Gate Mechanism*) used in *EFN* model.

Key words: Fake News; Social Media Computing; Text Mining; Sentiment Analysis

目　　录

1 绪论

1.1 研究背景与意义

以互联网为代表的信息技术的发展，颠覆性地改变了人们的各项生活方式。现如今，人们查阅信息、阅读新闻的渠道，已越来越倾向于在线社交媒体。比起传统的新闻媒体，社交媒体中的新闻更容易在线制作，在其天然的社交媒介属性之上，人们可以借助平台的转发、评论等功能，更自由地进行信息分享与讨论。以中文的社交媒体新浪微博为例，其注册用户数量已超过 5 亿，且其中有 30 万以上的认证用户、超过 13 万家机构账户[1]，新型社交媒体的影响力可见一斑。

然而，由于社交媒体中的大多数新闻篇幅短小、制作容易，其内容的发布常常缺少专业的第三方过滤与事实核查[2]，这也为低质量新闻，甚至虚假新闻的扩散与传播创造了有利条件。例如，2017 年的“红黄蓝幼儿园事件”中，有不法分子利用公众的同情心大肆造谣，带来了极坏的社会影响[3]。不仅如此，虚假新闻还广泛遍布于政治选举、科学发现、自然灾害、恐怖袭击、股市投资等多个领域[4, 5]。

社交媒体上虚假信息的产生与传播，是多方面因素共同促进的结果。一方面，很多虚假信息的发布者旨在影响公众舆论，激发社会的不安情绪，为公众在特定领域中的认识营造偏见[2]。另一方面，大量具有煽动性的虚假新闻更能够吸引读者，因此文章的点击率得以提高，这为社交媒体平台带来了大量的广告营收。除此之外，社会心理学的相关研究也表明，人类的“认知失调”、“选择性感知”与“回音室效应”等心理因素，也是推动虚假信息传播的重要原因[6]。

在如此严峻的形势面前，对社交媒体虚假信息的有效检测，成为学术界与工业界共同关注的重要议题。然而此任务难度巨大，面临着严峻挑战：（1）首先，绝大多数虚假信息都经过了精心编造，其写作风格、行文手法都极力模仿真实新闻，这为二者的区分带来了极大的困难[7, 8]；（2）其次，由于互联网产生与传播信息的速度快、体量大，社交媒体中的待审核新闻规模海量，且其更新迭代速度与日俱增，因此这也宣告着向人类专家发送文章、寻求人工核查的传统检测方法可发挥的作用极低，而需要自动化的数据驱动的检测框架[7]；（3）然而，由于大多数社交媒体平台对于公共数据搜集的限制[9]，现有的含有虚假信息标注的数据集较为有限，这也为数据挖掘模型的设计增大了难度。

1.2 国内外研究现状

1.2.1 相关术语

虚假信息　目前，领域内对于虚假信息还未达成一个公认的准确概念，广义上讲，其代表着所有具有欺骗性质的信息[6]，可以看作是错误信息（Misinformation）、假情报（Disinformation）、假新闻（Fake News, False News）、谣言（Rumor）等多个概念的统称。在本文中，虚假信息指的是**已被证伪的新闻事实类消息**。

社交媒体　社交媒体（Social Media）指的是在社交网络基础上的信息传播平台。目前，国内外在虚假信息检测领域所研究的社交媒体对象，多为 Facebook、Twitter、新浪微博等。在本文中主要的研究对象为中文的社交媒体**新浪微博**。

社交媒体虚假信息检测　在国内外的相关研究中，社交媒体虚假信息检测指的是**判断信息是否虚假的二分类（Binary Classification）检测任务**，其一般采用有监督学习（Supervised Learning）的方法，利用已被标注的数据集训练出机器学习模型，从而对未来信息进行二分类预测。

1.2.2 社交媒体虚假信息检测方法综述

在社交媒体上的虚假信息检测，最早可追溯到 2011 年 Castillo 等人对于 Twitter 平台所发布的新闻可信度的评估[10]。自此之后的虚假信息检测领域，其研究方法大致可归为三类：（1）基于特征构造的传统机器学习检测方法；（2）基于可信度传播网络的检测方法；（3）基于神经网络的深度学习方法。

基于特征构造的传统机器学习方法，其研究重难点是提取各式各样的新闻特征，从而提高机器学习分类器的精度。[10] 首次把社交媒体中的新闻特征归结四类：发布者的用户特征，新闻文本的内容特征，新闻 Tweet 在社交平台上的传播特征，以及新闻所属的主题特征。[11] 等人也在之后强调了新闻发布的地理特征以及用户特征的重要性。

基于可信度传播网络的方法，旨在对同一新闻事件中的多种实体及其关系进行图模型的构建。[12] 率先通过构建关系图的方法，引入了可信度传播网络的模型。[13, 14] 则利用了层次化的网络结构，对新闻事件中所蕴含的核心报道、子事件、消息进行了传播网络的建模。

深度学习的快速发展，也把多种基于深度神经网络的方法带到了虚假信息检测领域。[15] 首次利用了循环神经网络（RNN）模型，将每个新闻事件中所出现的

微博，按照时间先后次序作为 RNN 输入的多个时间步。[16] 则借鉴了注意力机制（Attention）的思想，提出了基于社交注意力机制的层次化模型。

1.3 研究思路

1.3.1 现有方法对情感信息利用的不足

正如 1.1节中所提到的，虚假信息的发布者在撰写假新闻时，会尽力使其编纂的内容足够耸人听闻，吸引眼球，以便促使人们的传播，从而进一步的扩大谣言的影响力。而人的情感因素，正是能够促进虚假信息传播的重要手段。

已有很多的社会心理学研究表明，对于社交媒体平台而言，一条具有较强感染性与煽动性的、能够激发公众（恐惧、愤怒或烦躁）情绪的新闻，更能被有效传播[17, 18]。除此之外，也已有大量虚假信息检测的相关研究，显露出情感信号的重要性。例如，在本领域的开山之作 [10] 中，作者在分析各类特征对检测模型的重要性贡献时，提出新闻内容中与情感相关的特征（如感叹号、问号、正负情感词等），比其他特征更靠近决策树模型的根结点，其信息增益更高；在 [19] 中也提出，对于质疑类情绪的捕捉与利用，能够有效促进虚假信息在早期发展阶段的检测。

然而，现有研究对于情感信息的利用十分有限。首先，领域内尚缺乏全面的对虚假信息情感信号的挖掘工作，我们对于情感在虚假信息中的具体表现与分布特点仍不够清楚；其次，大多数方法对情感信息的利用，仍停留在根据已有的情感词典进行情感词匹配，而现有的情感词典多根据严肃的文学语料构建，由于新型社交媒体中的行文与用词常常具有互联网风格，因此通过情感词匹配所构建的特征表示，往往与真实的情感信号差距很大。

1.3.2 本文的主要贡献

正如1.3.1节中所提到的：一方面，我们需要完整的虚假信息情感信号的挖掘工作，来全面认识情感在虚假信息中的表现特点；另一方面，在利用情感信号进行建模表示时，不能仅仅采用简单的情感词典匹配的方法，而需要更先进的编码方式。

针对现有研究中存在的这两点劣势，具体地，本文的创新性贡献如下：

- 提出一种挖掘社交媒体信息中情感信号（发布者情感与社交群体情感）的方法，并在此基础上，用多个角度阐述了虚假信息与真实信息在情感分布上的

差异。

- 在情感信号的建模中，（1）既利用了传统的情感词典匹配的方法，（2）还把情感分析领域中较为先进的情感嵌入表示（Emotion Embedding）应用到虚假信息检测领域中，并通过大量互联网微博语料的预训练，得到了可靠的情感编码。
- 设计出基于情感分析的社交媒体虚假信息检测模型（Emotion-based Fake News Detection framework, EFN），该模型能够利用深度神经网络的方法，同步并自适应地利用虚假信息中的语义信号与情感信号。
- 在现实场景（新浪微博）的数据集上进行了详尽的实验，结果表明本文所提出的 EFN 框架超越了领域内现有的最优模型。除此之外，在实验的评估部分，还对两种情感信号的建模方式做出详细比较，结果验证了情感嵌入表示方法的有效性。

1.4 论文组织结构

本文的第一章为绪论，主要介绍了整个社交媒体虚假信息检测领域的研究背景与意义，概括性地对此领域的国内外研究现状作出说明，特别地，还对本文所涉及到的相关术语作出说明，并阐述了本研究所针对的单微博虚假信息检测问题，并指出了现有研究对情感信息利用的不足，并详细说明了本研究的主要工作与贡献。

第二章简要地介绍了本研究涉及到的相关理论与技术，包括情感分析、多模态融合。

第三章是全面性的虚假信息情感信号的挖掘工作，分别从情感种类的分布、情感强度的特性、情感词的表达这三个方面呈现出虚假信息中情感信号的分布特点。

第四章详细介绍了本研究提出的基于情感的虚假信息检测模型，展示了整个框架的具体架构，并对其中的新闻文本模态、社交评论模态作出细节性说明。

第五章是模型的实验与评估部分，说明本研究所使用的数据集、所比较的领域内基准模型，以及本文所提出的情感嵌入表示、门融合机制两种技术的性能评估。

第六章则对本文的整体工作进行了总结，并对未来工作作出了一定的构思与规划。

2 相关理论与技术

在本章中，主要介绍了本研究中所涉及到的关键理论与技术，其主要包括情感分析，以及多模态融合。在2.1节中，首先介绍了情感分析领域中对于情感、情绪的定义，并对各种情感信号的建模方法作出介绍，并详细说明了在本研究中使用情感嵌入表示方法的原因。在2.2节中，阐述了多模态融合领域的三种融合策略，并具体解释了本文中所应用的基于门机制的融合策略的技术背景。

2.1 情感分析

2.1.1 情感与情绪的定义

情感分析（Sentiment Analysis），或观点挖掘（Opinion Mining）是对人们关于某实体（如某个特定的问题或事件）的意见、评价、态度或情感进行计算的研究[20]。随着社交媒体的爆炸性发展，一些个人或组织将越来越多地利用这些媒体中的公众意见来做出决策。例如，某些商业机构将利用社交平台中用户对于其产品的评论，挖掘出消费者的相关意见[20]。在过去的数十年中，学术界在该领域已经做了大量的研究[21, 22]。

通常意义上讲，情感（Sentiment）或观点（Opinion）指的是对某实体的一种正面或负面的态度、情绪或评价，其中：正面（Positive）、负面（Negative）和中立（Neutral）被称作情感倾向（Sentiment Orientation）或极性（Polarity）。因此大部分研究对于情感的划分，都会归结为正面、负面（或中立）这两（三）类。

除此之外，在本领域还有另一个与情感（Sentiment）相近的概念：情绪（Emotion），它指的是我们主观的感受或想法[20]。根据 [23] 的相关研究，人们的情绪可分为爱（Love）、乐（Joy）、惊（Surprise）、怒（Anger）、悲（Sadness）、惧（Fear）等六大类。本文中的第 3 章在挖掘虚假新闻的情感信息时，就采用了情绪作为观测指标。

2.1.2 情感的计算方式

在情感分析领域中，情感与情绪的表示（Representation）形式大致可归为以下三种：（1）基于情感词典的情感值计算；（2）基于深度情感分类模型的情感值预测；（3）通过大规模语料预训练得到的情感词嵌入表示（Embedding）。

基于情感词典的情感值计算，是领域内的表示情感值的经典方法，它通过简单的硬匹配方法，根据词典判定句子中的某个词是否为情感词。因此，该方法大大依赖于专家编撰的情感词库。领域内较为权威的情感词典有 HowNet ①（中文）、NTUSD[24]（中文）、WordNet[25]（英文）与 MPQA[26]（英文）等。

基于深度情感分类模型的情感值预测方法，有赖于近年来取得重大进展的深度学习技术。该方法首先利用大规模的情感语料训练一个情感分类器，之后便可用分类器对句子的情感倾向进行预测打分。行业内，腾讯情感分析平台 ②与百度情感倾向平台 ③都开放了相关 API 接口供用户调用。

情感词嵌入表示（Embedding）的方法，借鉴了自然语言处理领域中词向量（Word2Vec）的思想。词向量的预训练结果，获得了每个词语在语义空间上的高维特征表示[27]；而情感词嵌入表示的预训练，也能够类似获得每个词语在情感空间上的高维向量表示。如 [28] 等人就利用了社交平台 Tweet 中的数据，训练得到了情感词嵌入表示。[29] 则通过亚马逊平台上用户的产品评论数据，训练得到蕴含丰富情感含义的词向量。

以上三种情感的计算方式在应用中各有优劣。基于情感词典的方法使用简单，但它在新型互联网上的社交媒体中应用时，常常存在着词语覆盖率低、难以处理语义歧义等问题；基于深度情感分类器的方法精度相对较高，但分类器的预训练需要大量的情感语料，而且由于模型的输出多是一个标量的情感值，因此只能处理句子级别的情感识别问题；情感词的嵌入表示方法较为灵活，能够获得词语级别的情感向量表示，因此可被广泛应用于情感分析相关的各类下游任务。本文在第 4 章设计基于情感的虚假信息检测模型时，主要利用了情感词嵌入表示的方法。

2.2 多模态融合

模态是指事物发生或存在的方式，或人接受信息的特定方式[30]，例如，人在面对多媒体数据时，往往涉及到同时对文字信息、视觉信息和听觉信息的处理与接收。模态间融合是多模态研究领域的一个基本问题，其目的在于充分整合与利用各模态间的不同信息，从而提高模型的精度与鲁棒性。根据采用融合的阶段的

①http://www.keenage.com/html/e_index.html
②https://ai.qq.com/product/nlpemo.shtml
③http://ai.baidu.com/tech/nlp/sentiment_classify

不同，多模态融合通常具有前融合（Early Fusion）、后融合（Late Fusion）、混合融合（Hybrid Fusion）三种方式。

最直观的融合方法就是合并（Concatenation），其通过简单拼接的方法把不同模态的向量组合起来。[31] 实现了一个堆叠的自编码器，将语义的嵌入表示和视觉信号的输入映射到了共同的向量空间。然而，这样简单合并的融合方式对各模态间的信息等同对待，不够切合现实场景（不同模态常常具有不同的贡献）。在 [32] 的研究中，作者在处理视觉模态和文字模态时，就采用了动态融合的方法，其借鉴了注意力机制（Attention）的思想，把处理后的视觉表示作为外部的权重向量，从而自适应地学习到了文本模态的重要性。

除此之外，门机制（Gate mechanism）也被广泛应用于融合学习的领域。LSTM[33] 和 GRU[34] 就是其中的经典模型，其通过不同门的设置（更新门、遗忘门等），计算得到上一个时间步对当前输入的隐层单元，从而解决了传统的循环神经网络（RNN）的长距离依赖问题。[35] 也利用了门机制的思想，融合了词语在不同模态中的表示。在本文中，在（1）融合每个词在语义空间和情感空间的嵌入表示，以及（2）融合新闻文本模态与评论文本模态时，就采用了门机制的动态融合方法。

3 虚假信息的情感信号挖掘

为了全面地认识社交媒体虚假信息中的情感表现，本研究首先进行了基础的数据分析工作：分别挖掘虚假新闻、真实新闻的新闻内容、用户评论中的情感信号。在数据收集上，本文抓取了 7880 条虚假新闻、7907 条真实新闻以及每条新闻所对应的用户评论（共约 15 万），此处的分析数据集与后续的检测实验数据集相同，在5.1节中会更详细介绍数据集的相关细节。

在本章中，我们首先定义出社交媒体中情感信号所呈现出的两种表现形式，其包括发布者的情感与社交群体的情感，并对这两种情感形式产生的原因作出了一定的阐释。之后，本章主要从三个角度挖掘出虚假新闻的情感信号的分布特点：情感种类的分布，情感强度的特性，以及情感词的表达。实验结果表明：虚假新闻中所含的愤怒、悲伤、质疑等负面情绪更多、情感强度更为激烈；且社交媒体用户在表达同一种情绪时，对待虚假新闻的言语和用词常常更为夸张与极端。

3.1 社交媒体中情感的表现形式

为了更清晰地证明情感信号对于虚假信息检测的重要性，本文首先定义出社交媒体新闻事件中，情感信息所呈现出的两种形式：（1）**发布者的情感（Publisher Emotion）**，即：信息发布者在撰写新闻时，其微博文本内容中所表现出的情感信号；（2）**社交群体的情感（Social Emotion）**，即：新闻微博的读者在讨论该新闻事件时所表现出的情感，其通常会在用户对新闻微博的评论、转发中的文本内容所呈现。

图3.1和3.2分别呈现了两条虚假新闻及用户对其的相应评论，我们以此为例来说明这两种情感的表现形式，并简要指出虚假信息在情感表现上的直观特点。在图3.1中，假新闻的发布者使用了多个情绪化的词语（如“我的天啊!”、“太可惜了!”），企图增加内容的煽动性，以达到“以假乱真”的目的；而图3.2中假新闻的情感表现则显得大相径庭，其新闻内容的写作风格平铺直叙，并未使用任何情感词语，但却勾勒出一个耸人听闻的事实性消息（“中国的国民素质排名位列世界倒数第二位”），从而引起了公众的不满与轰动，于是微博的评论区便充斥着愤怒的情绪化信息（“这是认真的吗?!”、“联合国是有病吗”）。由此可见，为了达到有效的社交媒体虚假信息检测，必须同时利用发布者情感和社交群体情感这两方面的

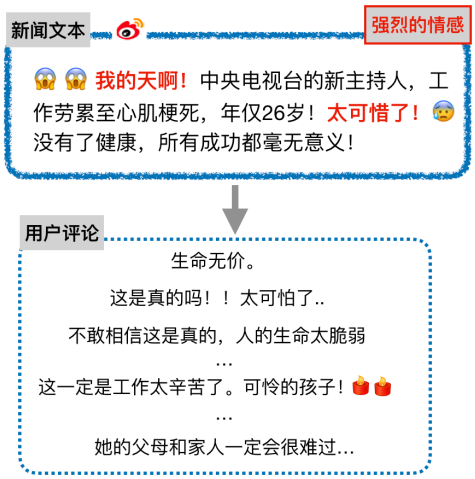

图 3.1　强烈的发布者情感

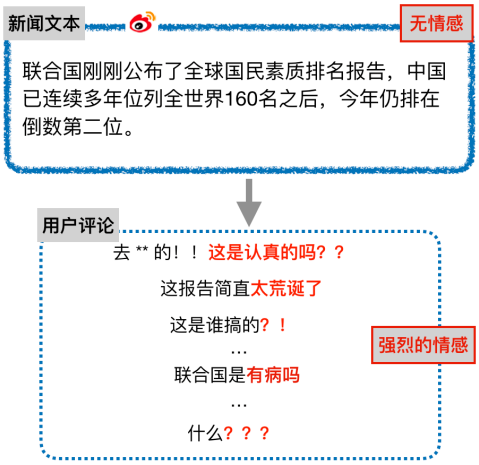

图 3.2　强烈的社交群体情感

信息。因此，本文的工作中对于情感信息的分析与利用，都将围绕情感的这两种表现形式展开。

3.2　情感种类的分布

直观感受上，虚假新闻常常耸人听闻且极具煽动性，很容易激起用户的某些情绪，如质疑、焦虑或震惊。因此，本文选择了 5 种细粒度的情绪分类来刻画虚假信息与真实信息的情感分布：愤怒（anger），悲伤（sadness），质疑（doubt），开心（happiness），以及无情绪（none，即不具备情感信号）。此处利用了4.2.1节中的情感分类器来对分析数据集进行了情感种类的标注。

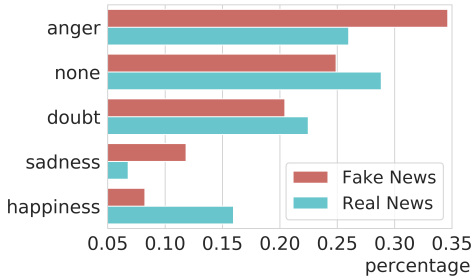

图 3.3　新闻文本

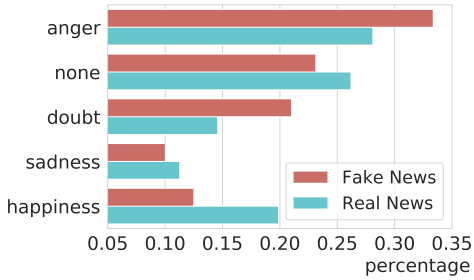

图 3.4　用户评论

图3.3和3.4分别展示了新闻文本和用户评论中的情感种类分布。相较于真实新闻，虚假新闻的文本内容中，愤怒的情感占比高了 9%，而开心的情感低了 8%。用户评论的情感种类分布也有类似的特点。而且对于虚假新闻，其新闻文本内容中的悲伤情绪以及其用户评论中的质疑情绪，都显著高于真实新闻。这样的结果表明：无论是假新闻的编撰者，还是参与讨论假新闻的用户，都会在言辞中表达出

更多的愤怒、悲伤、质疑等负面情绪。

3.3 情感强度的特性

除了情感种类外，每个句子还拥有在这类情感上的情感强度特性[20]。例如，“我今天非常高兴”比“我今天挺高兴的”拥有更强的快乐情绪。基于此，本文进一步分析了虚假新闻与真实新闻中的情感强度特性。

在情感强度的计算上，本研究依然利用了4.2.1节中的情感分类器。具体地，将分类器模型 softmax 层的输出值（代表样本对每种情感的分类概率值），作为对应种类情感的强度值（一个范围在 0 到 1 内的连续值，越大的值代表越强的情感）。

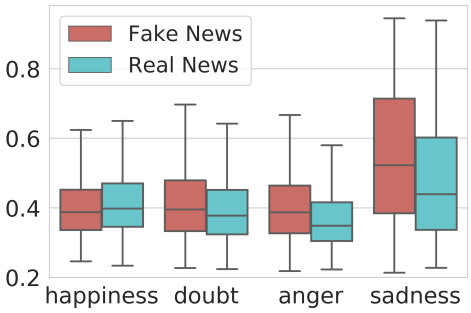

图 3.5 新闻文本

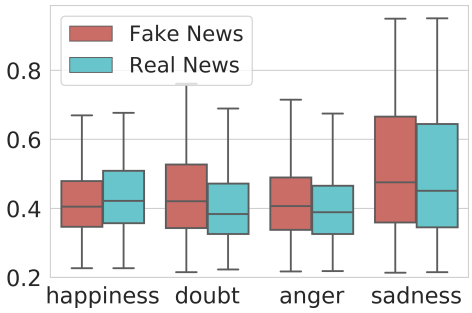

图 3.6 用户评论

在图3.5和图3.6中我们可以看到：无论对于新闻文本还是用户评论，虚假信息中愤怒、悲伤和质疑的情感所蕴含的强度都更为激烈（这样的差异，在新闻文本中表现得尤为明显）。因此，新闻发布者与社交媒体中的用户，都会对于虚假新闻表现出更强烈的负面情绪。

3.4 情感词的表达

不同的人在表达其情绪时，常常会使用不同的言辞。例如，有些人喜欢用平实的词语表达心理感受，而有些则倾向更为夸张的表达方式。因此，为了分析人们对于虚假新闻在情感言辞表达方式上的特性，本研究还进一步分析了其新闻内容、用户评论内容中情绪化词语的使用。

本文采用了 [36] 所提出的被广泛应用的方法，其能够计算出文本数据集中每个词对于不同情感种类的表达程度（权重）。此处以愤怒的情感表达为例，通过该算法计算出每个词语的“愤怒权重”，图3.7和图3.8分别列举了虚假新闻和真实新闻中“愤怒权重”最高的 30 个词语。

图 3.7　虚假新闻

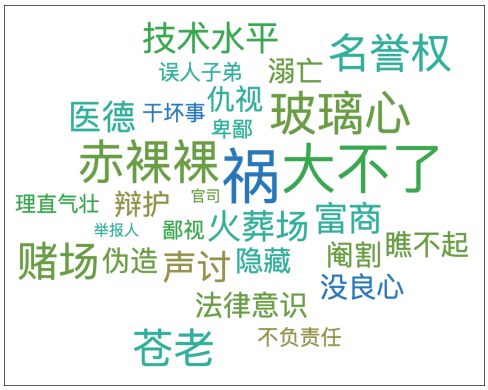

图 3.8　真实新闻

我们能够看到，在虚假新闻中，人们表达愤怒时会选用更暴躁更极端的用词，如“他妈的”、“婊子”等。相似的结论同样存在于其他的负面情感中。由此我们可知，人们对于真实信息与虚假信息会采用不同的情感言辞，这也进一步揭示着利用情感信号的重要性。

3.5　小结

通过对虚假信息情感种类分布、情感强度特性、情感词表达的挖掘，我们能够得到以下结论：(1) 虚假新闻的发布者与讨论者，均会更多地表现出负面情绪（如愤怒、悲伤、质疑）；(2) 即使同样是表达负面情绪，虚假新闻的参与者所表现出的情感强度也更为激烈；(3) 在表达同种情感时，虚假新闻与真实新闻的用户在言辞选择上依然有很大差异，虚假新闻的参与者尤其会使用更为夸张和耸动的用词。

这样的结论印证了情感信号对于虚假信息检测任务的重要性，也为接下来检测模型的设计带来了一定的启示。

4 基于情感的虚假信息检测模型

由上一章中对情感信号分布的挖掘可知，基于情感的虚假信息检测模型的设计，需要解决三个关键问题：（1）虚假信息与真实信息中的情感种类分布差异明显，因此模型需要拥有能够有效捕捉情感种类信号的特征；（2）虚假信息中的情感更浓烈与极端，因此模型需要对情感强度等信号进行特征的构建；（3）虚假信息与真实信息在表达情感时的行文用词不同，因此模型对文本数据的语义信号和情感信号需要进行有效的融合。

针对如上的三个问题，本研究所提出的检测模型的解决方案如下：（1）借鉴情感分析领域最先进的情感嵌入表示（Emotion Embedding）技术，具体地：在情感嵌入向量的预训练过程中，设定一个文本情感分类的训练任务，从而得以使获取的情感向量富含情感种类的抽象信息；（2）对于情感强度的建模，需要利用传统的基于情感词典的匹配方法，例如：构建程度词、否定词、问号、感叹号等统计情感特征，来刻画文本的情感强度；（3）在语义编码和情感编码的表示融合时，借鉴门机制（Gate Mechanism）的思想，使模型能够自适应地学到对这两种编码的利用倾向。

4.1 概述

4.1.1 问题定义

在虚假信息检测领域，由于单条微博往往只含有几个简短的句子，所蕴含的信息有限，在实际研究中，往往需要利用与此微博相关的其他数据信息（例如：此微博下的评论微博、转发微博，或此微博的传播结构等）[15, 16]。因此，在本研究中我们把一个虚假信息事件定义为：一条虚假的新闻微博及其用户评论的集合体，本文主要针对于事件级别，而非单条微博级别的虚假信息检测。

在模型的设计上，本文也将对新闻微博文本、新闻微博的用户评论这二者采用分离的模块（新闻文本模态、社交评论模态）分别处理，并在模型的预测层之前进行融合。

4.1.2 模型概览

在本章中，首先在4.2节对情感嵌入表示的预训练过程做出了详尽地说明，具体包括预训练的模型、任务，以及语料。

之后，将详细阐述本研究设计的一个端对端（end-to-end）的基于情感的虚假信息检测模型（Emotion-based Fake News Detection Framework，EFN）。

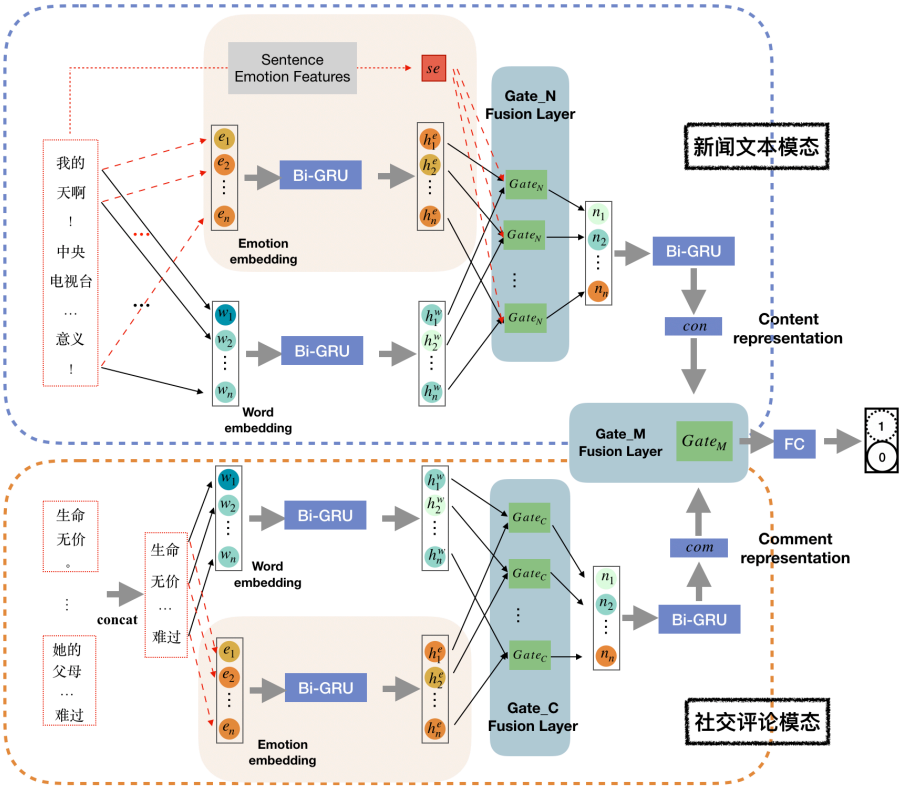

图 4.1　EFN 模型图

模型的整体架构包括三部分（如图4.1所示）：

（1）新闻文本模态（Content Module）。此模态主要蕴含了新闻发布者的有关信息，包括新闻文本内容的语义信息与情感信息，将在4.3节中详细说明；

（2）社交评论模态（Comment Module）。此模态捕捉的是用户对新闻事件的评论，同样包括评论的语义信息与情感信息，将在4.4节中详细说明；

（3）虚假信息检测模块（Fake News Prediction Component）。此处融合了新闻文本模态和社交评论模态的深层抽象表示，并最终进行虚假信息的检测预测，将在4.5节中详细说明。

在整个模型中，还包括三种基于门单元的融合层：新闻文本模态中的 *Gate_N Fusion Layer*，社交评论模态中的 *Gate_C Fusion Layer*，以及虚假信息检测模块中

的 *Gate_M Fusion Layer*。关于三种门单元的详细结构，将在4.3.4节、4.4节和4.5节中分别介绍。

4.2 预训练的情感嵌入表示

正如1.3.1节中所提到的，单纯的情感词典匹配技术具有匹配度低、不符合互联网行文风格等劣势。因此在本研究中的情感信号建模中，我们使用了情感分析领域中较为先进的情感嵌入表示（Emotion Embedding）方法，该技术的应用背景在2.1中已做出了一定的说明。与自然语言处理领域广泛使用的词向量（Word2Vec）技术相似，情感嵌入表示也是通过大量语料的预训练，从而得到每个词语的情感向量表示。此节将分别介绍我们在获取情感向量时所采用的预训练的任务与模型、预训练的语料，以及预训练得到的情感嵌入表示的可靠性分析。

4.2.1 预训练模型

此处我们借鉴了情感分析领域最先进的情感嵌入表示模型 [29]，其通过训练一个深度情感分类模型，从而获得每个词语的情感嵌入向量。在本文的预训练过程中，我们设计了五分类的情感文本分类任务，共包括愤怒（anger），悲伤（sadness），质疑（doubt），开心（happiness），以及无情绪（none，即不具备情感信号）这五种情感分类。

如图4.2所示，整个情感分类器模型包括嵌入层（Embedding Layer）、双向 GRU 层（Bi-GRU Layer）等部分。在预训练的过程中，模型的输入是经过词典索引表示的 one-hot 向量（与 Word2Vec[37] 的预训练输入相同），经过多轮的训练，模型的嵌入层将学习到每个词典索引所对应的嵌入向量，即为本研究所要获取的情感嵌入表示。在参数设置上，此处我们将词典大小设置为 30000，嵌入层的向量维度设置为 16，即：通过预训练，获得了词典中 30000 个词的 16 维向量表示。

4.2.2 预训练语料

对于预训练的情感文本语料，本文参考了 [38, 39] 的语料获取方式，我们首先获取了大规模的含有表情（Emoticon）的短文本微博，并将每条微博的表情作为其情感种类的标签（例如：含有表情“*[爱你]*”的文本，被标注为具有“开心”的情绪）。

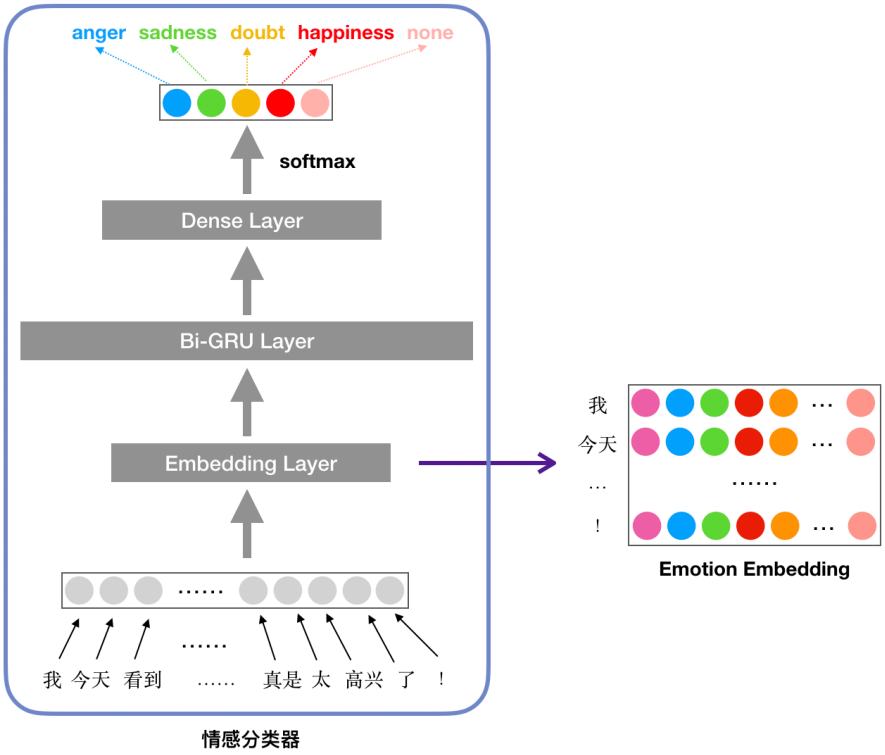

图 4.2　预训练模型图

在表4.1中展示了本研究中所使用的预训练语料，表中列举了我们对于表情标签的标注，以及每种情感文本的样本数量。

	选取表情	文本数量
愤怒	*[怒] [微笑] [拜拜] [摊手] [汗] [抓狂] [怒骂] [衰] [哼] [鄙视]* 等	10,155
悲伤	*[蜡烛] [悲伤] [可怜] [晕] [生病] [伤心] [委屈] [失望] [泪流满面]* 等	10,155
质疑	*[吃惊] [疑问] [傻眼] [震惊]* 等	10,155
开心	*[心] [爱你] [哈哈] [doge] [笑 cry] [偷笑] [嘻嘻] [喵喵] [鼓掌] [害羞]* 等	10,155
无	*[思考] [污] [话筒] [白眼] [懒得理你]* 等	10,155

表 4.1　预训练语料

4.2.3　预训练效果评估

通过深度情感分类器模型的预训练，我们得到了每个词语的情感嵌入表示，在进一步利用情感嵌入向量进行情感信号的建模之前，本文首先对预训练过程中分类器模型的性能、情感嵌入向量的可靠性进行一定的评估。

Label	Prec	Recall	F1
愤怒	0.41	0.38	0.40
悲伤	0.53	0.39	0.45
质疑	0.42	0.58	0.49
开心	0.44	0.39	0.41
无	0.29	0.31	0.30
Avg	0.42	0.41	0.41

表 4.2　预训练中情感分类器的性能

表4.2中展示了经预训练后的情感分类器性能，可以看到，对于此五分类问题，分类器平均能够对每个分类达到 42% 的分类准确率。由 [36] 所显现的，在情感分析领域，类似这样的细粒度情感分类任务的难度往往较高，因此在本研究中我们认为这样的分类器已经达到了细粒度情感分类问题的平均水平，预训练的效果可以满足后续情感信号的建模需求。

为了进一步验证预训练的效果，本文还对在后续模型中直接应用的情感嵌入向量，做出了一定的可视化分析。和3.4节中所使用的方法相同，本文利用 [36] 中计算语料中每个词语的情感权重的方法，针对愤怒、悲伤、质疑、开心、无这五种情感，各自选取情感权重最高的 100 个词。之后，对这 500 个情感词所对应的情感向量进行 t-SNE 可视化分析[40]，如图4.3所示。为了更清楚地展现情感嵌入向量的效果，此处我们还对比了这 500 个情感词的语义向量（Word2Vec）分布，如图4.4所示。

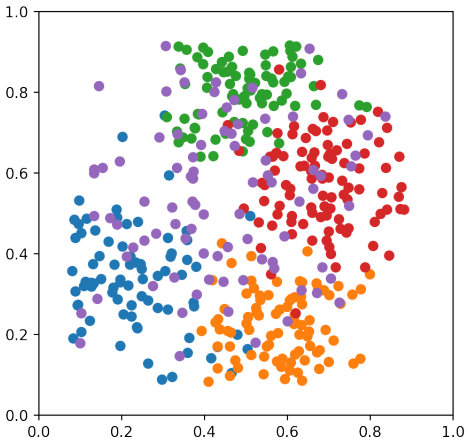

图 4.3　Emotion Embedding

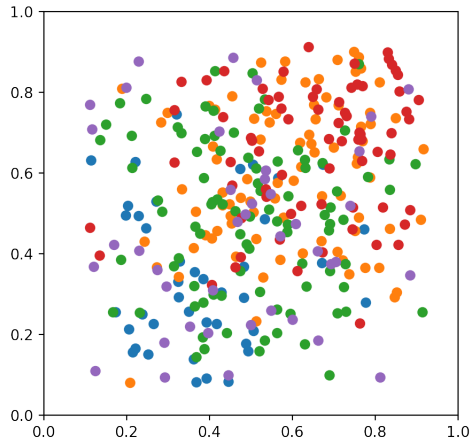

图 4.4　Word2Vec

能够看到，对于同一情感种类的词语，其经过 t-SNE 降维处理的 2 维情感向量，在空间上具有很高的相似度，而其语义向量的分布则较为杂乱。由此可见，本研究中的预训练过程提取到了较为可靠的情感嵌入表示，这为接下来虚假信息检测模型的设计奠定了基础。

4.3 新闻文本模态

新闻文本模态（Content Module）涵盖着虚假信息与真实信息的原始文本，即新闻事件发布的原微博内容。上一章中的数据分析实验，已经显示了虚假新闻与真实新闻在内容上的情感分布差异，因此本研究所设计的模型旨在更好地捕捉文本中的情感信号，并与语义信号进行有效融合。

从模型结构上看，新闻文本模态总共包括三部分：（1）语义编码，其通过双向 GRU 模型对每个词语的语义词向量进行了编码；（2）词粒度的情感嵌入编码，与语义编码相似，依然由双向 GRU 模型对词语的情感嵌入向量进行编码；（3）句子级别的传统情感特征编码，其主要是在句子级别对微博文本进行情感特征的提取，作为情感信号建模的进一步补充。

4.3.1 语义编码

本研究中，采用了基于循环神经网络（RNN）的语义编码（Word Encoder）来获取文本内容的语义表示。尽管从理论上讲，RNN 能够通过巧妙的参数设置来捕捉到句子的长期依赖，但实际应用中，这样完美的参数常常难以调节。基于门的循环单元（Gated Recurrent Unit，GRU）模型的出现解决了这一问题，其能够维持更长时间步内的记忆[34]。为了更好地利用文本内容的上下文信息，此处使用了双向的 GRU 模型，其能够同时利用句子词语间的前向与反向关系。

对于每一个词 t_i，向量 t_i^w 由预训练的词向量模型（Word2Vec）进行初始化[37]。假设每个句子 s 分词后的结果为 $\{t_0, t_1, \cdots, t_M\}$，在双向 GRU 模型中，前向 GRU 模型 $\overrightarrow{f}$ 采用 t_0 到 t_M 的顺序读取句子中词语的输入，后向 GRU 模型 $\overleftarrow{f}$ 则采用 t_M 到 t_0 的顺序逆向读取句子中词语：

$$\begin{aligned} \overrightarrow{h_i^w} &= \overrightarrow{GRU}(t_i^w), i \in [0, M], \\ \overleftarrow{h_i^w} &= \overleftarrow{GRU}(t_i^w), i \in [0, M]. \end{aligned} \tag{4.1}$$

对于每个给定的词语 t_i，记前向 GRU 模型所输出的隐向量表示为 $\overrightarrow{h_i^w}$，后向

GRU 模型输出的隐向量表示为 $\overleftarrow{h_i^w}$，则 h_i^w 即为词语 t_i 通过语义编码后得到的表示，其中 $h_i^w = [\overrightarrow{h_i^w}, \overleftarrow{h_i^w}]$。

4.3.2 词粒度的情感嵌入编码

经过4.2节的预训练，我们已经得到了每个词的情感嵌入向量。和语义编码器相似，本文仍采用双向 GRU 模型来获取文本模态在情感空间中的隐式表示。对于每个词 w_i，向量 t_i^e 代表经预训练得到的其情感向量。假设每个句子 s 分词后的结果为 $\{t_0, t_1, \cdots, t_M\}$，在双向 GRU 模型中，前向 GRU 模型 $\overrightarrow{f}$ 采用 t_0 到 t_M 的顺序读取句子中词语的输入，后向 GRU 模型 $\overleftarrow{f}$ 则采用 t_M 到 t_0 的顺序逆向读取句子中词语：

$$\begin{aligned} \overrightarrow{h_i^e} &= \overrightarrow{GRU}(t_i^e), i \in [0, M], \\ \overleftarrow{h_i^e} &= \overleftarrow{GRU}(t_i^e), i \in [0, M]. \end{aligned} \tag{4.2}$$

对于每个给定的词语 t_i，记前向 GRU 模型所输出的隐向量表示为 $\overrightarrow{h_i^e}$，后向 GRU 模型输出的隐向量表示为 $\overleftarrow{h_i^e}$，则 h_i^e 即为词语 t_i 通过情感编码后得到的表示，其中 $h_i^e = [\overrightarrow{h_i^e}, \overleftarrow{h_i^e}]$。

4.3.3 句子级别的传统情感特征编码

特征描述	**特征数量**
文本中的表情个数	1
文本中所含五类表情（愤怒、悲伤、质疑、开心、无）的分别数量	5
文本中的特定标点（！、？、…、。。。、，，，）个数	5
文本中正、负向情感词个数	2
文本中否定词个数	1
文本中程度词个数	1
文本的情感值	1
文本中人称代词（第一人称、第二人称、第三人称）数量	3
共计	19

表 4.3 句子级别的传统情感特征

4.3.2节的情感嵌入方法能够获得每个词语的情感向量表示，从而在词粒度完成了情感信号的有效编码。然而，对于一个句子而言，其全局性的情感统计特征也蕴含着丰富的情感信号。例如，程度副词（如“很”、“非常”）的个数就是一种

典型的情感统计特征，在模型的训练过程中，我们希望蕴含更多程度副词的样本，其情感信号的作用能够被有效加强。因此，利用传统方法抽取的句子级别的情感特征也被应用在了此模态中。

对于一条新闻文本 q，我们根据 [10] 的设定提取了情感特征，并在其基础上添加了具有情感含义的表情的统计特征，最终获得了新闻 q 所具有的 19 个情感特征 se，其中包含正/负向情感词的个数、句子的情感值等，如表4.3所示。

4.3.4 新闻文本模态表示

在获得了语义编码 h_t^w，词粒度的情感编码 h_t^e，以及句子级别的情感特征编码 se 之后，本文应用门机制（Gate mechanism）将这三种模态的特征做融合，如图4.1中的 *Gate_N Fusion Layer* 所示。

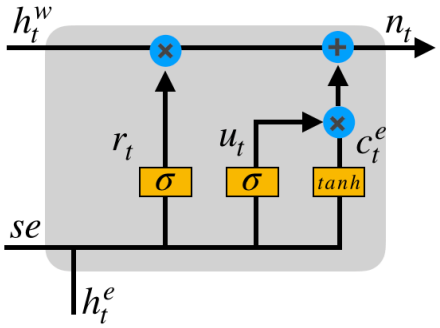

图 4.5　门单元 Gate_N

Gate_N 为本模态中使用的门单元，其结构如图4.5所示。Gate_N 门单元的设计借鉴了 GRU 中重制门（Reset Gate）、更新门（Update Gate）的思想，从而使门单元的输出更能强化每个词语的融合表示。在 Gate_N 中，两种情感信号（词语的情感嵌入、句子的情感特征）共同决定了通过两个 sigmoid 层值的输出 r_t 与 u_t（即：重制门的输出与更新门的输出），这样的门单元设计使得模型可以自适应地学习到每个单词的语义表示、情感表示分别的重要性。除此之外，tanh 层的设计也把输入的情感信号，映射到和语义信号相同维度的特征空间中。与 Gate_N 相关的输入、

输出处理如以下公式所示：

$$
\begin{aligned}
r_t &= \sigma(W_r.[se, h_t^e] + b_r) \\
u_t &= \sigma(W_u.[se, h_t^e] + b_u) \\
c_t^e &= tanh(W_c.[se, h_t^e] + b_c) \\
n_t &= r_t * h_t^w + u_t * c_t^e
\end{aligned}
\tag{4.3}
$$

最终，所有词语的新型表示依据原有的时序关系，被依次输送到双向的 GRU 层。经过双向 GRU 层处理，最后一个隐层单元的输出 con 代表了新闻模态的全部信息，被称作新闻模态表示（Content Module Representation）。

4.4 社交评论模态

社交评论模态（Comment Module）挖掘了新闻事件中参与用户所发表评论中的语义信息与情感信息。此模态的架构设计与新闻文本模态大体相似，仅有以下两处不同：（1）[15, 16] 等研究中均显示，由于单个用户的评论常常字数较少，不具备显著信息，因此在实际应用中常常把多个用户的评论合并起来。在此模态中，本研究也将一个新闻事件的所有评论拼接起来，形成用户群体对该事件的评论文本；（2）由于评论文本内容由多条评论组成，因此在本模态中并未抽取句子级别的情感特征。

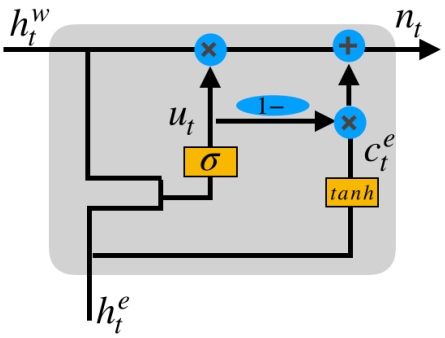

图 4.6　门单元 Gate_C

Gate_C 是社交评论模态中所使用的门单元，其结构如图4.6所示。不同于 Gate_N 单元，Gate_C 只拥有两个输入（只有语义编码的输入和词粒度情感嵌入编码的输入，缺少 Gate_N 中的句子情感特征输入）。因此我们仅使用一个更新门，来对这两个输入进行融合，并随后通过 sigmoid 层获得每个词语的新型表示 u_t。与 Gate_C

相关的输入、输出处理如以下公式所示：

$$
\begin{aligned}
u_t &= \sigma(W_u.[h_t^w, h_t^e] + b_u) \\
c_t^e &= tanh(W_c.h_t^e + b_c) \\
n_t &= u_t * h_t^w + (1 - u_t) * c_t^e
\end{aligned}
\tag{4.4}
$$

和新闻文本模态相似，所有词语的新型表示依据原有的时序关系，被依次输送到双向的 GRU 层。经过双向 GRU 层处理，最后一个隐层单元的输出 com 代表了评论模态的全部信息，被称作社交评论模态表示（Comment Module Representation）。

4.5 整体框架

在分别获得新闻文本模态（ 4.3.4节）和社交评论模态（ 4.4节）的深层抽象表示（con 和 com）后，本文采用了门单元 Gate_M 进行两种模态的融合（如图4.7），并输出融合后的向量表示 n：

$$
\begin{aligned}
r &= \sigma(W_u.[con, com] + b_u) \\
n &= r * con + (1 - r) * com
\end{aligned}
\tag{4.5}
$$

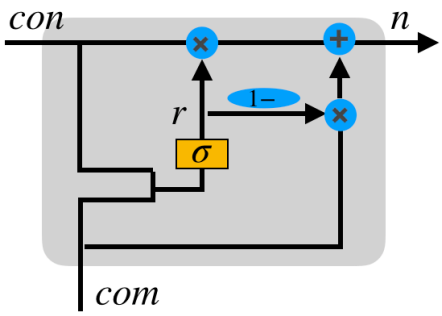

图 4.7　门单元 Gate_M

在模型的最后，本文先使用全连接层把融合向量 n 映射到二维空间内（共有虚假新闻、真实新闻两类），并通过 softmax 激活函数获得在每个类别内的概率分布：

$$
p = softmax(W_c n + b_c) \tag{4.6}
$$

模型使用二维的交叉熵作为训练的损失函数 L，假设共有 m 个样本，则所有样本造成的损失 S^m 为：

$$
L(S^m) = -[l^m p^m + (1 - l^m) \log(1 - p^m)] \tag{4.7}
$$

其中，p^m 代表第 m 个样本被判定为虚假新闻的概率值，l^m 为第 m 个样本的真实标签，其值为 1（样本为虚假新闻）或 0（样本为真实新闻）。

5 实验与评估

在本章中，展示了所提出的基于情感的虚假信息检测模型（Emotion-based Fake News Detection Framework，EFN）在真实环境的微博数据集下的性能体现。首先，在5.1节中介绍了整体实验所使用的微博数据集；其次，在5.2节中列举了实验中所要进行比较的基准模型，其中包括多种虚假信息领域内最先进的深度模型，结果表明：本研究所提出的 EFN 模型，能够比领域内最先进的模型拥有 4% 的 F1 值提高（性能对比结果将在5.3节中得到详细说明）；最后，还着重对本研究中所应用的两种关键性技术（情感嵌入表示、门融合机制）的性能做出多角度的评估（5.4节、5.5节），评估结果表明了这两种技术的有效性。

5.1 数据集

本研究基于新浪微博平台①构建了数据集，整个数据集包括 7880 个虚假新闻微博、7907 个真实新闻微博，以及对应新闻事件的评论（共约 16 万条）。

在进一步介绍本文的数据集之前，此处我们先对目前领域内数据集情况做出一定的说明。目前在整个虚假信息检测领域，较为权威的规模较大的中文数据集有 Ma Jing [15] 与 Jin Zhiwei [32] 等人构建的微博数据集。Ma Jing 的数据集主要针对于用户对虚假信息微博的转发行为，其数据集共有 2312 个虚假新闻微博、2334 个真实新闻微博（共计 4646 个新闻事件），数据规模相较本文较小[15]；Jin Zhiwei 的数据集来源与本文相同，其中的虚假新闻事件均来源于微博社区管理中心的官方辟谣系统②，其爬取了自 2012 年 5 月至 2016 年 12 月的虚假微博数据[32]，而本研究中爬取了 2012 年 5 月至 2018 年 10 月的虚假微博，因此可以看作是对其数据集的扩充。

正如上文所叙述，本文的虚假新闻微博的来源是微博社区管理中心的官方辟谣系统；对于真实新闻，本研究利用了 NewsVerify 系统③，该系统能够实时抓取微博中产生的争议性新闻，我们选取了其中经判定为真实新闻的争议事件[41]。为了更真实地反映新闻事件发布早期在社交平台中所引起的讨论，对于每个新闻事

①https://www.weibo.com
②http://service.account.weibo.com
③https://www.newsverify.com

件，都抓取距离其发布时间 24 小时内的用户评论（因此，可能有些新闻事件没有用户评论数据）。数据集的统计信息如表5.1所示。在训练过程中，本研究采用了 4:1 的比例划分训练集与测试集。

	虚假新闻	真实新闻	共计
# 新闻文本	7,880	7,907	15,787
# 用户评论	109,154	47,037	1,56,191

表 5.1　数据集

5.2　领域内基准模型

在实验中，EFN 对比了虚假信息检测领域中多种较为先进的模型，如下所示：

- DTC [10]，此模型也是虚假信息检测领域的首个代表性模型，其使用了 J48 决策树模型，模型的输入为人工提取的 68 个特征（包括 21 个文本内容特征、7 个用户特征、35 个事件特征、5 个传播特征），其中也蕴含了4.3.3节所使用的句子级别的文本情感特征。
- ML-GRU [15]，该模型是领域内性能优异的深度学习模型。作者把社交平台内对于一个新闻事件的讨论（转发与评论）按照时序关系划分时间步，并输入到多层的 GRU 模型中。
- HSA-BLSTM [16]，该模型是近两年用深度学习方法检测虚假信息的最先进模型之一。作者通过多层次的注意力模型，捕捉一个新闻事件发展中的级联架构，并提取了用户与新闻的社交特征（其中蕴含一些基本的情感特征）作为注意力机制。
- Basic-GRU，此模型分别用两个双向 GRU 模型捕捉新闻模态、评论模态的语义信号，在融合两个模态时采用简单拼接的方式。Basic-GRU 模型相当于 EFN 除去了情感信号，并不采用门机制的融合策略，因此可看作是 EFN 的自比较模型。

所有的实验，都采用了领域内的通用评价指标：分类准确率（Accuracy）、查准率（Precision）、查全率（Recall）以及 F1 值。

5.3 性能比较

表5.2显示了 EFN 和所有基准模型的实验结果。较为突出地，本文所提出的 EFN 模型达到了 87.2% 的分类准确率以及 87.4% 的 F1 值，显著高于所有其他模型的性能。这样优异的实验结果也表明，对于词语的情感嵌入表示方法，以及基于门单元的融合机制的利用，能够有效提高虚假信息检测模型的性能。

Methods	Acc	Prec	Recall	F1
DTC	0.756	0.754	0.758	0.756
ML-GRU	0.799	0.810	0.790	0.800
Basic-GRU	0.835	0.830	0.850	0.840
HSA-BLSTM	0.843	0.860	0.810	0.834
EFN	**0.872**	**0.860**	**0.890**	**0.874**

表 5.2 EFN 与领域内基准模型的比较

除此之外，实验结果还表明现有的基于深度学习的模型（ML-GRU、HSA-BLSTM 以及 Basic-GRU）都优于传统的基于手工特征提取的模型（DTC）。这也在某种程度上证明了基于 RNN 的网络模型，在处理时序化的文本结构方面具有天然的优势，能够挖掘到文本内容的深层抽象表示。在深度学习模型间的纵向比较中，我们还能够发现 Basic-GRU 模型比 ML-GRU 表现更好，这极有可能是因为 Basic-GRU 模型对于新闻模态和评论模态作了显式的划分，而 ML-GRU 却将这二者简单拼接起来，这样的结果也说明新闻事件的文本内容和用户评论内容之间的差异较大。

从整体上看，本文所提出的基于情感的虚假信息检测模型（EFN）在真实环境的数据集中达到了优异性能。较领域内最原始的模型 DTC 提高了约 12% 的精度（从 75.6% 提升至 87.2%），较领域内最先进的模型 HSA-BLSTM，也拥有 4% 的 F1 值提高。

5.4 情感嵌入表示的性能评估

由上一章中模型的设计可以看出，EFN 模型最独特的两个模块有：（1）应用了词语的情感嵌入表示方法（Emotion Embedding），从而加强了情感信号对于虚假信息检测的重要性；（2）应用了基于门单元的融合机制（Gate mechanism），从而

在多模态融合时充分利用了各模态的信息。因此，本研究用进一步的实验验证了这两个模块的重要性。在本节中先介绍情感嵌入表示方法的性能评估，在5.5节中介绍门融合机制的评估。

为了全面分析情感嵌入表示方法对性能带来的影响，此处将分别评估该方法在（1）新闻文本模态（Content Module）、（2）社交评论模态（Comment Module）、以及（3）二者的模态融合（Content + Comment）后上所带来的性能改变。

我们用 WE 和 EE 分别代表：单纯使用词向量嵌入表示（Word Embedding）和单纯使用情感嵌入表示（Emotion Embedding），并用 WEE 代表同时使用这两种嵌入表示。除此之外，为了对比词粒度的情感嵌入编码和句子级别的传统情感特征编码的性能，我们用 SE 表示单纯使用句子级别的统计情感特征，并用 WSE 代表同时使用词向量嵌入、以及句子级别的统计情感特征。实验的结果如表5.3所示。

Module	Methods	Acc	F1
Content	WE	0.790	0.801
	EE	0.700	0.719
	SE	0.623	0.618
	WSE	0.793	0.790
	WEE	**0.813**	**0.810**
Comment	WE	0.667	0.550
	EE	0.619	0.553
	WEE	**0.669**	**0.560**
Content + Comment	WE + WE	0.835	0.840
	WEE + WE	0.860	0.859
	WEE + WEE	**0.866**	**0.868**

表 5.3　情感嵌入表示的性能评估

从表5.3的实验结果来看：（1）与单纯使用词嵌入的语义信号相比，融合使用语义信号和情感信号对所有模态都有性能的提升；（2）对于本文中所使用的数据集，情感信号的重要性在新闻文本模态中体现的更为明显（这样的结果可能是因为数据集中评论数据的稀疏性，即：有相当部分的新闻事件没有用户评论，因此情感信号的有效性在社交评论模态中难以体现）；（3）对于新闻文本模态而言，本文提出的情感嵌入表示方法，比传统的句子级别的情感特征提取具有更好的效果；（4）从模态之间的对比来看，单独使用新闻文本模态最高能够达到 81.3% 的准确率，而单独使用社交评论模态仅能达到 66.9% 的准确率，在两个模态融合之后，最终的

准确率能够比单独使用新闻文本模态提高超过 5%。

5.5 门融合机制的性能评估

Module	Methods	Acc	F1
Content	WEE(c)	0.813	0.810
	WEE(att)	0.799	0.793
	WEE(gn)	**0.851**	**0.854**
Comment	WEE(c)	0.669	0.560
	WEE(gc)	**0.671**	**0.563**
Content + Comment	(WEE(gn)+WEE(gc))(c)	0.866	0.868
	(WEE(gn)+WEE(gc))(gm)	**0.872**	**0.874**

表 5.4 门融合机制的性能评估

接下来，本文将进一步探究基于门单元的融合机制对于整个检测模型的性能影响。本文提出的融合策略所涉及到的门单元有 Gate_N，Gate_C 与 Gate_M，此处分别记为 gn，gc 与 gm。我们对比了领域内应用较为广泛的拼接融合策略（Concatenation，记为 c）；以及 [32] 中所提到的基于注意力机制（Attention）的融合策略（记为 att），其将一种模态的深层抽象表示作为注意力向量，从而动态引导另一种模态的学习。

与上一节相似，为了全面分析门融合机制对性能带来的影响，此处将分别在（1）新闻文本模态、（2）社交评论模态、以及（3）整体的两种模态上评估上述几种融合策略的优劣。实验结果如表5.4所示。

由此可以看到：（1）对新闻文本模态而言，在融合语义嵌入与情感嵌入的表示时，直接的拼接（c）比 [32] 提出的基于注意力机制的融合（att）有更好的性能，而本文提出的门单元 Gate_N（gn）相较拼接策略仍有超过 4% 的 F1 值提高；（2）对于评论模态而言，门单元 Gate_C（gc）依然比拼接策略（c）拥有更好的性能；（3）在融合抽象程度更高的新闻模态与评论模态时，门单元 Gate_M（gm）也取得了一定的性能提升。

为了更好地理解基于门单元的融合机制，增强模型的可解释性，此处本文以一些样本为例来展现门融合策略的效果。以门单元 Gate_C 为例，由公式4.4可知，门单元中的 u_t 向量可看作每个词语的语义向量 h_t^w 和情感向量 c_t^e 的权重，即：它决定了对于每个词语而言，其语义信息和情感信息分别有多少的重要性。

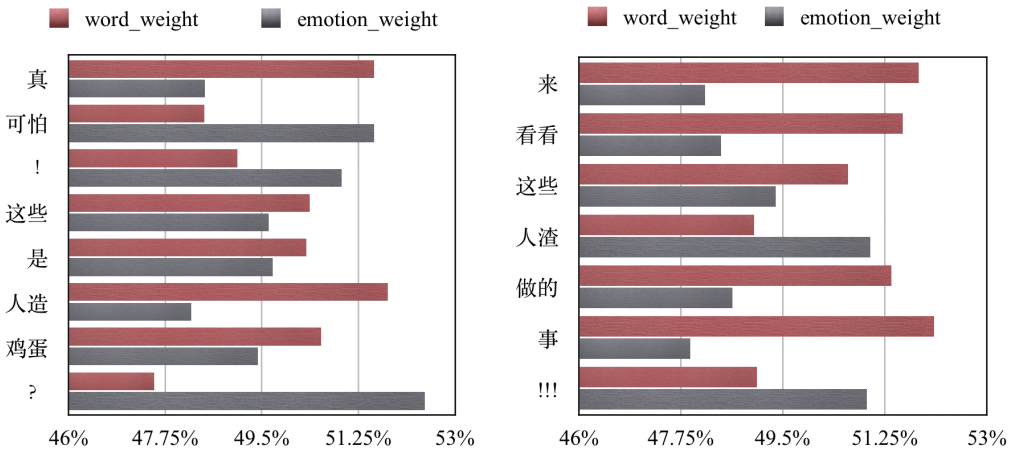

图 5.1　门单元 Gate_C 的评估

图5.1展示的是一条对于虚假新闻事件的用户评论，我们能够看到：通过模型的训练，“可怕”、“人渣”、“！”、“？”等词语学习到了更高的情感权重，而“这些”、“鸡蛋”、“事”等词则具有更高的语义权重。因此，门融合机制能够更好地利用情感词的信息，从而最终提高了虚假信息检测模型的性能。

6 总结与展望

6.1 本文的工作总结

在本篇论文中，我们首先进行了大量的对虚假信息的情感信号挖掘工作，从数据挖掘的研究角度确认了情感信号对虚假信息检测的重要性，为整个社交媒体虚假信息检测领域提供了一些基础的理论性结论。其次，本文还提出了一种基于情感的虚假信息检测模型（Emotion-based Fake News Detection Framework，EFN），其能够同时捕捉新闻事件发布者、参与讨论的用户这两者的情感信号，并分别使用了新闻文本模态与社交评论模态，来利用新闻文本与用户评论的语义信息和情感信息。

EFN 模型在技术上主要有两个创新之处：第一点是情感嵌入方法的应用。通过预训练得到每个词语的情感嵌入向量，这相比于传统的基于词典的情感值计算方法拥有匹配度更高、更能够利用上下文语境信息等优势；第二点则是基于门单元的融合策略。通过三种门单元的设计，模型能够自适应地学习到对于不同模态的信息利用。

通过在真实的微博数据集上的实验，我们确认了 EFN 模型相较领域内的其他基准模型，甚至最先进的深度模型，均有优异的性能体现。与此同时，还通过大量评估性的对比实验以及案例分析，证明了 EFN 模型中的情感嵌入表示方法、门融合机制方法，均能为虚假信息的检测带来很大的性能提升。

6.2 研究缺陷与未来展望

由于时间与个人精力所限，本研究中也存在一些潜在的缺陷，例如：本文主要针对中文社交平台微博的虚假信息检测，对于领域内较为重要的英文平台 Tweet、Facebook 等没有作出相关的对比实验，因此 EFN 模型对于英文数据集的性能体现还是未知数；其次，本研究中所使用的数据集中，由于只抓取了每个新闻事件发布的前 24 小时的评论，因此有大量的事件不存在用户评论数据，这对于 EFN 模型中两种模态的设计（新闻文本模态、社交评论模态）具有一定的影响。

除此之外，从模型的设计上看，由于 EFN 模型中所使用的情感嵌入表示（Emotion Embedding）需要经由预训练过程，从而在某种程度上提高了整体研究的难度，也增加了模型训练所需要的时间。在工业界的实际应用中，对于此类需要预训练

流程的模型，在工程应用中也存在着极其重要的技巧。因此，在理论研究之外，也希望日后的学习生活中能够提高对工业界中类似应用的关注。

在未来的研究工作中，除了对上述缺陷的改进，在本文的基础上仍有一些十分重要的研究点可以探索。本文中已经通过大量的数据分析工作，确认了情感信号对于社交媒体虚假信息检测问题的重要性，在此结论的基础上，如何有效利用情感信息则成为了一个新的话题，而 EFN 模型的提出仅仅只是这个话题的冰山一角。例如，虚假信息的早期检测一直是行业内亟需解决的难题，那么我们是否能够探究新闻事件在早期传播中的情感信号的时序变化，从而使虚假信息在早期的发展阶段就得到检测与预防。此处我们也期待大量类似的研究点能够得到探索。

参考文献

[1] 刘知远, 张乐, 涂存超, et al. 中文社交媒体谣言统计语义分析 [J]. 中国科学: 信息科学, 2015, 45(12) : 1536.

[2] ALLCOTT H, GENTZKOW M. Social media and fake news in the 2016 election[J]. Journal of Economic Perspectives, 2017, 31(2) : 211 – 36.

[3] 姜金贵, 闫思琦. 基于主题和情绪相互作用的微博舆情演化研究——以“红黄蓝虐童事件”为例 [J]. 情报杂志, 2018, 37(12) : 118 – 123.

[4] VOSOUGHI S, ROY D, ARAL S. The spread of true and false news online[J]. Science, 2018, 359(6380) : 1146 – 1151.

[5] GRINBERG N, JOSEPH K, FRIEDLAND L, et al. Fake news on twitter during the 2016 US Presidential election[J]. Science, 2019, 363(6425) : 374 – 378.

[6] PIERRI F, CERI S. False News On Social Media: A Data-Driven Survey[J]. CoRR, 2019, abs/1902.07539.

[7] SHAO C, CIAMPAGLIA G L, VAROL O, et al. The spread of low-credibility content by social bots[J]. Nature Communications, 2018, 9(1) : 4787.

[8] SHU K, SLIVA A, WANG S, et al. Fake News Detection on Social Media: A Data Mining Perspective[J]. SIGKDD Explorations, 2017, 19(1) : 22 – 36.

[9] SHAO C, HUI P, WANG L, et al. Anatomy of an online misinformation network[J]. CoRR, 2018, abs/1801.06122.

[10] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on twitter[C] // Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011. 2011 : 675 – 684.

[11] YANG F, LIU Y, YU X, et al. Automatic Detection of Rumor on Sina Weibo[C] // Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics. New York, NY, USA : ACM, 2012 : 13:1 – 13:7.

[12] GUPTA M, ZHAO P, HAN J. Evaluating Event Credibility on Twitter[C] // Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012.. 2012 : 153 – 164.

[13] JIN Z, CAO J, JIANG Y, et al. News Credibility Evaluation on Microblog with a Hierarchical Propagation Model[C] // 2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014. 2014 : 230 – 239.

[14] JIN Z, CAO J, ZHANG Y, et al. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs[C] // Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA. 2016 : 2972 – 2978.

[15] MA J, GAO W, MITRA P, et al. Detecting Rumors from Microblogs with Recurrent Neural Networks[C] // Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016. 2016 : 3818 – 3824.

[16] GUO H, CAO J, ZHANG Y, et al. Rumor Detection with Hierarchical Social Attention Network[C] // Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018. 2018 : 943 – 951.

[17] STIEGLITZ S, DANG-XUAN L. Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior[J]. Journal of Management Information Systems, 2013, 29(4) : 217 – 248.

[18] FERRARA E, YANG Z. Quantifying the effect of sentiment on information diffusion in social media[J]. PeerJ Computer Science, 2015, 1 : e26.

[19] ZHAO Z, RESNICK P, MEI Q. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts[C] // Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015. 2015 : 1395 – 1405.

[20] LIU B, ZHANG L. A Survey of Opinion Mining and Sentiment Analysis[G] // Mining Text Data. 2012 : 415 – 463.

[21] LIU B. Sentiment Analysis and Subjectivity[G] // Handbook of Natural Language Processing, Second Edition.. 2010 : 627 – 666.

[22] PANG B, LEE L. Opinion Mining and Sentiment Analysis[J]. Foundations and Trends in Information Retrieval, 2007, 2(1-2) : 1 – 135.

[23] PARROTT W G. Emotions in social psychology: Essential readings[M]. [S.l.] : Psychology Press, 2001.

[24] KU L-W, LO Y-S, CHEN H-H. Using polarity scores of words for sentence-level opinion extraction[C] // Proceedings of NTCIR-6 workshop meeting. 2007 : 316 – 322.

[25] KAMPS J, MARX M, MOKKEN R J, et al. Using WordNet to Measure Semantic Orientations of Adjectives[C] // Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal. 2004.

[26] WIEBE J, WILSON T, CARDIE C. Annotating Expressions of Opinions and Emotions in Language[J]. Language Resources and Evaluation, 2005, 39(2-3) : 165 – 210.

[27] BENGIO Y, DUCHARME R, VINCENT P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3 : 1137 – 1155.

[28] TANG D, WEI F, QIN B, et al. Sentiment Embeddings with Applications to Sentiment Analysis[J]. IEEE Trans. Knowl. Data Eng., 2016, 28(2) : 496 – 509.

[29] AGRAWAL A, AN A, PAPAGELIS M. Learning Emotion-enriched Word Representations[C] // Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. 2018 : 950 – 961.

[30] 刘建伟, 丁熙浩, 罗雄麟. 多模态深度学习综述 [J]. 计算机应用研究, 2019, 1(18).

[31] SILBERER C, FERRARI V, LAPATA M. Visually Grounded Meaning Representations[J]. IEEE Trans. Pattern Anal. Mach. Intell., 2017, 39(11) : 2284 – 2297.

[32] JIN Z, CAO J, GUO H, et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[C] // Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017. 2017 : 795 – 816.

[33] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8) : 1735 – 1780.

[34] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C] // 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015.

[35] WANG S, ZHANG J, ZONG C. Learning Multimodal Word Representation via Dynamic Fusion Methods[C] // 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. 2018 : 5973 – 5980.

[36] LI C, WU H, JIN Q. Emotion Classification of Chinese Microblog Text via Fusion of BoW and eVector Feature Representations[C] // Natural Language Processing and Chinese Computing - Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings. 2014 : 217 – 228.

[37] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality[C] // Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. 2013 : 3111 – 3119.

[38] HU X, TANG J, GAO H, et al. Unsupervised sentiment analysis with emotional

signals[C] // 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013. 2013 : 607 – 618.

[39] 何炎祥, 孙松涛, 牛菲菲, et al. 用于微博情感分析的一种情感语义增强的深度学习模型 [J]. 计算机学报, 2017, 40(4) : 773 – 790.

[40] MAATEN L V D, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(Nov) : 2579 – 2605.

[41] ZHOU X, CAO J, JIN Z, et al. Real-Time News Cer tification System on Sina Weibo[C] // Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume. 2015 : 983 – 988.

致谢

东湖之滨，珞珈山上，如今坐在“全国最美的大学校园”里，过往的时光还历历在目。感谢武大在过去的日子里给我的包容与陪伴，她聆听着我平日里的欢声笑语，也注视着我许多深夜里的郁郁寡欢。感谢自己在武大留下的点滴回忆，教学楼的自习室还留存着考试周复习的身影，珞珈山的环山路也惦念着夕阳下跑步的足迹。

在这里，我首先要感谢贾向阳老师对我毕业设计的悉心指导。贾老师治学严谨，看待问题一针见血，在我论文的写作过程中提出了大量宝贵的意见。更难得的是，其为人宽大谦和，对待学生认真负责，在同学之中拥有很好的人气与口碑。

其次，我要感谢中科院计算所-前瞻实验室-跨媒体课题组的全体老师和同学。作为我毕业设计的校外指导老师，曹娟导师耐心指导了我的整个工作，并言传身教，令我学习到科研人员所应具备的浓烈求知欲以及不断求索的品质与精神。也感谢郭川、亓鹏等师兄师姐对我论文的各项指导与帮助，与他们的沟通与交流，能让我不断修正实验方案与思路，并对研究的优劣之处拥有更清晰的认识。

除此之外，还有数不清的老师与同学在过去四年里给我的陪伴。感谢我的辅导员黄轶雯老师，她对我的学习生活给予了极大的支持与关怀；感谢刘峰老师，能够成为他的助教，令我受益匪浅；感谢陈艳姣老师，我在陈老师实验室里的日子异常充实，她也是我得以在本科生涯就迈入科研的领路人；感谢雷云聪、岳翔等学长学姐，他们在我大学的迷茫时期，给了我太多的信心与勇气；感谢我的室友、软工 5 班的同学、以及文艺部的小伙伴们，和他们一起的日日夜夜，就像照亮平淡生活的点点星光。

最后，我要感谢我的家人，感谢他们这么多年对我的养育和教诲。感谢我的父母对我永恒的理解、支持与信任，正是有他们的后盾我才能昂首阔步，不断奋进。

毕业在即，分离将至，此日一别或数载，不知何时返珈园。

自强弘毅，求是拓新，母校训诫伴心间，永是珞珈一少年。