情报理论与实践
Information Studies:Theory & Application
ISSN 1000-7490,CN 11-1762/G3

《情报理论与实践》网络首发论文

题目：基于 BiLSTM-CRF 的政府微博舆论观点抽取与焦点呈现
作者：胡吉明，郑翔，程齐凯，张岩
网络首发日期：2020-09-18
引用格式：胡吉明，郑翔，程齐凯，张岩．基于 BiLSTM-CRF 的政府微博舆论观点抽取与焦点呈现．情报理论与实践.
https://kns.cnki.net/kcms/detail/11.1762.g3.20200917.1734.008.html

●胡吉明[1,2]，郑　翔[1,2]，程齐凯[1,2]，张　岩[3]
（1.武汉大学信息管理学院，湖北 武汉 430072；2.武汉大学信息检索与知识挖掘研究所，湖北 武汉 430072；3.武汉大学测绘遥感信息工程国家重点实验室，湖北 武汉 430079）

基于 BiLSTM-CRF 的政府微博舆论观点抽取与焦点呈现*

摘　要：**[目的/意义]**准确把握公众微博评论中所反映的公众观点并总结舆论焦点，有助于及时获取和引导社会舆情态势，对政府公信力、快速响应能力及执行力提升具有支撑作用。**[方法/过程]**文章针对当前政府微博评论社会功能发挥的现实要求和其文本特征挖掘的技术需求，从基于深度学习的文本智能语义理解和挖掘出发，提出了适用的细粒度四元组标注策略，构建了政府微博评论观点抽取与焦点呈现的深度学习模型 POF-BiLSTM-CRF，即通过细粒度标注策略确定、Word2vec 训练词向量、BiLSTM 评论特征学习进行标签及其概率输出、CRF 学习上下文实现微博评论标注优化，以及观点聚类和主题词提取后最终呈现舆论焦点。**[结果/结论]**针对“中国警方在线”微博评论的实验表明，本文所提研究框架和模型能够有效进行舆论观点的智能化提取，为快速把握公众观点及为政府决策提供了参考。
关键词：政府微博评论；舆论观点抽取；深度学习模型；BiLSTM-CRF 模型；POF-BiLSTM-CRF 模型

Public Opinion Extraction and Focus Presentation in Government Microblog Based on BiLSTM-CRF

Abstract:**[Purpose/significance]**Accurately understanding the public opinion reflected in comments of government microblog and summarizing the focus of public opinion will help to obtain and guide the situation of public opinion timely. It will also improve the government’s credibility, responsiveness and executive capability.**[Method/process]**Based on the current requirements of social function exertion and text feature mining of government microblog comments, this paper proposes an applicable fine-grained quaternion annotation strategy and the POF-BiLSTM-CRF model to extract and present government microblog comments opinion, from the perspective of text intelligent semantic understanding and mining based on deep learning. In this paper, fine-grained annotation strategy is used to extract comment connotations;Word2vec is used to train word vector; BiLSTM model is used to learn text features of Public microblog comments and output possible labels and probabilities; CRF model is used to optimize annotation by learn the context; clustering algorithm and topic words extraction algorithm are used to present the focus of public opinion. **[Result/conclusion]**The experiment of comments of “China police online” shows that the research framework and model proposed in this paper can effectively solve the problem of public opinion extraction intelligently and rapidly, which provides a reference for the rapid grasp of public opinion and government

* 本文为国家自然科学基金面上项目“基于深度学习的政务新媒体互动内容摘要自动生成与情感分析模型研究”的成果，项目编号：71874125。

decision-making.

Keywords: public microblog comments; opinion extraction; deep learning model; BiLSTM-CRF model; POF-BiLSTM-CRFmodel

目前，官方政府微博已成为传播政务信息、了解公众观点及把握舆论走向的重要渠道[1]，在了解社情民意、提供政务信息、打造阳光政府等方面起到了积极作用[2]。准确把握政府微博评论中所反映的公众观点并总结舆论焦点，有助于政府部门及时获取和引导社会舆情态势，提升政府公信力、快速响应能力及执行力。

但微博评论的海量化对政府部门及时有效识别和归纳公众舆论提出了新的挑战，特别是加大了迅速准确把握评论主旨要义的难度；因此从政府微博评论中准确抽取公众观点以支撑舆论焦点的实时把握，成为当前普遍关注的热点问题。同时，政府微博评论中包含了大量的非正式文法、口语化表达和缩写、不规范写法和噪音、表情符号等，导致其结构松散、数据稀疏且特征空间高维等特性[3]，为浏览和抽取评论内涵带来了更多困难。

综上，政府对公众舆论精准把握的需求与微博评论复杂的内容与结构特征之间的矛盾，要求重新审视公众观点抽取与舆论焦点把握的理论架构和实现策略。基于此，本文立足于文本内容智能语义理解和挖掘的通用框架，探索基于深度学习的政府微博评论分析框架与模型计算实现策略，以期达到快速、准确及有效地抽取与挖掘公众观点和呈现舆论焦点的目的。

1 政府微博舆情研究的技术演进

政府微博作为政府面向公众的重要媒体，受到了研究者、政府机构等的广泛关注，对其研究特别是内容研究的热度日益凸显。目前有关政府微博的研究大多集中在其信息传播效果、整体影响力、运营模式或在突发事件中舆情应对等方面，如从公众微观行为的视角，探究政府微博传播效果的影响机制[4]，进而研究政府微博影响力提升策略[5]和运营优化思路[6]；从政府角色出发，多维度评估突发事件中政府微博的作用[7]。针对政府微博评论内容的研究较少，且多基于关联规则[8]实现了公众对热点事件关注的词语挖掘，并未进行大规模数据样本基础上的政府微博评论内容焦点分析。

政府微博舆情研究的推进既需要充分考虑政府微博评论内容的独有特征，同时也需要大规模文本挖掘自动学习与分析的技术支撑。深度学习及相关技术在文本内容挖掘方面的优势日渐凸显。深度学习神经网络模型的集成运用更是成为当前文本内容分析的一大主流，如 BiLSTM 和条件随机场（CRF）模型的结合[9]则能够降低对词嵌入的依赖，较纯粹的 LSTM 和 BiLSTM 更具优势，在医学[10]、食品安全[11]、涉恐信息[12]等领域的应用证明，其对命名实体识别[13]、否定焦点识别[14]、网购商品评论观点抽取[15]等序列标注任务及后续文本深度挖掘任务的效果较好。

因此，本文将借助深度神经网络模型在自然语言处理中智能学习与训练的优势，构建面向政府微博评论观点挖掘的创新框架，即创新细粒度标注策略，集成深度学习模型 BiLSTM、CRF 以及聚类算法和主题词提取算法，提出面向政府舆情把控的公众观点（Public Opinion Focus，POF）抽取与焦点呈现框架（POF-BiLSTM-CRF），实现对大规模政府微博评论观点的自动化抽取与焦点呈现。

2 基于深度学习的政府微博舆论焦点挖掘模型

在细粒度标注策略确定与实现的基础上，本文集成深度学习模型 BiLSTM、CRF、K-means 聚类算法和 TF-IDF 算法，构建了政府微博评论观点抽取和舆论焦点呈现框架（POF-BiLSTM-CRF），其模型框架见图 1。

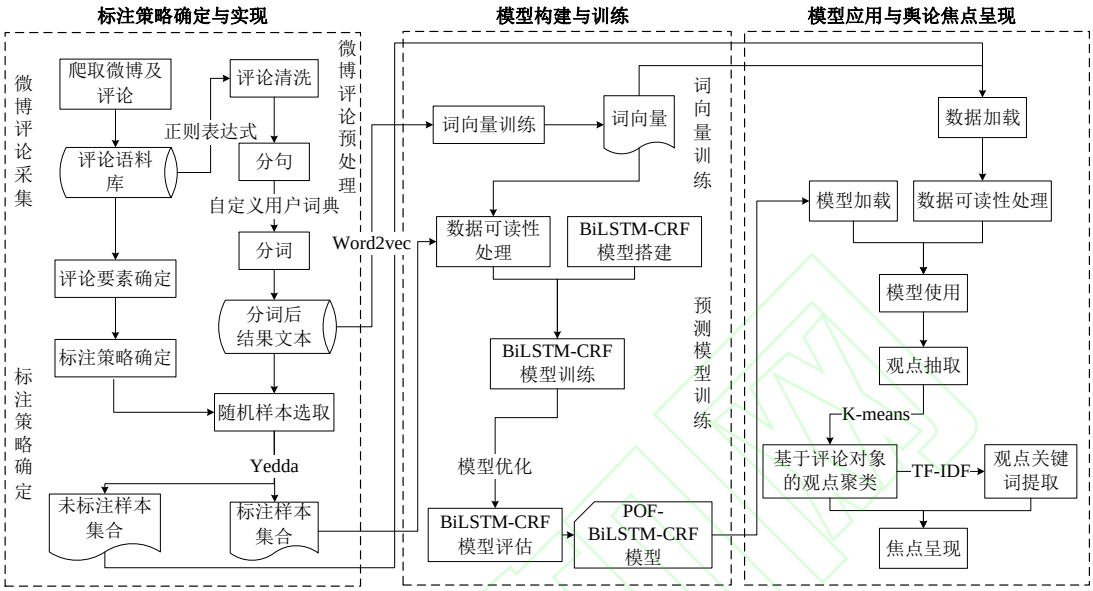

图 1 面向政府微博的舆论焦点挖掘框架 POF-BiLSTM-CRF

2.1 标注策略确定与实现

观点抽取的首要工作是进行内容标注，本文在微博评论采集和预处理的基础上，结合政府微博评论内容的特点，确定适用的标注策略。

1）基于 scrapy 框架的微博评论采集。构建基于 scrapy 框架的微博爬虫，自动抓取指定政府微博的评论数据，并搭建爬虫系统的 cookie 池与代理池，保证大规模互联网数据采集的完整性，在 MySQL 数据表中实现政府微博评论的格式化存储。

2）基于规则与词典的微博评论预处理。本文通过正则表达式过滤用语不规范的评论数据，包括：①删除评论中的表情及颜文字。微博表情与发言者真实意图并不完全对应，如“微笑”常表示讽刺、“哭泣”有时也表示感动。②删除评论中的@信息。@用于提醒他人注意评论或微博内容，无法表达评论者观点。③删除过短评论句。过短的中文文本无法准确表述语句观点[16]，因此删除长度小于 5 个字符的评论句。④删除仅有标点而无文字的评论。同时，本文采用人工判读方式扩充统计词典，依此经过多轮分词后最大程度保证分词的准确率。

3）基于要素分析的微博评论标注策略确定。不同于传统的商品评论，政府微博评论的内容要素构成较为复杂，除包含评论对象、程度及表达评论者态度的观点外，还常常包含描述评论对象已发生或可能发生的动作状态。例如，评论“警方积极作为，依然迅速，加快行动”中，除包含评论对象“警方”、程度“依然、快”及观点“积极作为、迅速”之外，还包含评论者希望评论对象“警方”实现的动作“行动”。因此，本文参考评论观点抽取的通用策略，即<评论对象，程度，观点>三元组[17]，基于对微博评论的内容要素分析，从评论

要素内涵区分和标签数量平衡角度出发，创新性地摸索出一套契合的细粒度四元组标签标注模式，即按照<评论对象（Object）、程度（Degree）、观点（Opinion）、动作（Action）>四元组分类标注，并在具体实现上对评论对象细化，以达到面向对象的观点抽取目的。

最后，本文采用 random 函数[18]随机划分评分文本，使用 BIO 标注方式[19]，结合简单易用且一致性较好的 YEDDA[20]，人工标注随机抽取的微博评论以训练模型，从而在测试模型准确率的基础上实现大规模语料的自动标注。

2.2 观点抽取模型构建

本文以 BiLSTM 和 CRF 为基础模型，结合词向量处理、文本特征学习等要求，集成模型以实现大规模政府微博评论的自动标注与观点挖掘。

1）基于 Word2vec 降维的词向量表示。词的向量化表示是自然语言处理的基础，本文首先将语料库中的词汇全部转化为词向量，以便模型输入。微博评论文本的数据稀疏，存在大量出现频次较低的生僻词。而广泛应用的 Word2Vec[21]能够将高维词向量嵌入到低维空间，并引入词的上下文信息。特别是其中的 Skip-gram 模型能够在已知当前词的前提下预测其上下文，利用周围词的预测结果不断调整中心词词向量，以此得到全部词向量。当数据量较少、生僻词出现较多时，Skip-gram 模型的多次训练与调整会使词向量更加准确[22]。因此，本文选择 Skip-gram 模型进行词向量表示，同时采用层次 Softmax 训练方法[23]简化参数及降低计算复杂度，既能充分利用上下文信息，又能保证整体预测的高效准确。通过 Word2vec 词向量表示后，评论文本中的词被映射到 N 维语义空间中，为后续模型输入及文本特征学习提供基础。

2）基于 BiLSTM 评论特征学习的标签及其概率输出。微博评论文本常出现大跨度的依赖关系，如“受害人受到伤害也不是一两天了，应该报警”中，评论对象是“受害人”，观点词是“报警”，二者位置跨度较大，为准确识别与抽取观点带来困难。但这种复杂的上下文关系对理解微博评论内涵至关重要。为解决上述长期依赖问题[24]、捕获上下文信息，本文使用 BiLSTM 模型生成更为全面的语义向量。其中，正向和逆向 LSTM 网络分别依次接受政府微博评论的正向和逆向输入，并依次计算正向和逆向隐藏状态，得到各输入词正向和逆向特征，拼接后得到双向表达向量，通过全连接神经网络处理后，获取各词所对应的标签及概率，从而提高标签预测的准确性。

然而 BiLSTM 模型忽略了输出标签序列间的依赖关系，导致标签预测结果顺序错乱。以 BIO 标注评论对象为例，若仅采用 BiLSTM 在学习文本特征后标注“受害人芝芝”，标注结果可能为“受害人 B-Victim 芝芝 B-Victim”，这明显与 BIO 标注方式不符。为提升文本特征学习的准确度和效率，针对上述标签预测优化的需求，本研究在 BiLSTM 模型处理的基础上，引入 CRF 模型优化标签预测的输出结果。

3）基于 CRF 上下文学习的微博评论标注优化。CRF 模型是由 J. Lafferty 等[25]在 2001 年提出的条件概率分布模型 $P(Y|X)$，表示给定一组输入随机变量 X 的条件下，另一组输出随机变量 Y 的马尔可夫随机场。不同于 BiLSTM 模型将每个标注词的最优标签拼接后输出，CRF 模型通过对标签序列依赖关系与顺序的学习，能够有效处理输出标签及概率，以计算标签间最优联合概率的方式对整个标签序列进行优化，从而使输出标签序列更加正确地符合输出规则，有效解决上述预测标签顺序不合理的问题。因此，本文引入 CRF 模型，充分考虑输出

标签序列的依赖关系与顺序，以此做出更加准确的判断，最终得到全局最优的标注序列，将关注点放在微博评论的句子级别上[9]。

综上，本文对词向量、BiLSTM 模型和 CRF 模型的集成能够有效利用上下文输入特征和句子级标签信息，并能通过学习标签间的约束条件，保留输出标签间的顺序关系，提升政府微博评论文本标签预测的准确性（见图 2）。

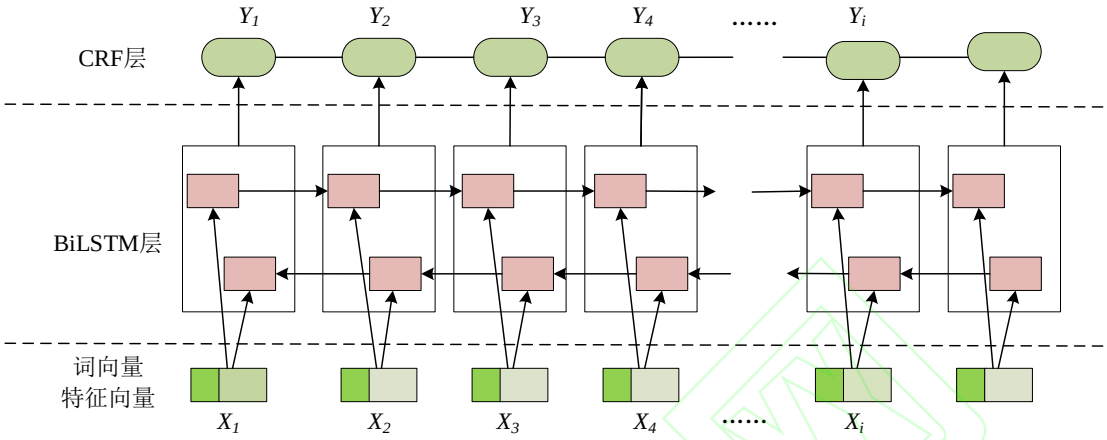

图 2　BiLSTM-CRF 模型结构

最后，本文将微博评论的标注文本按比例分为训练集、验证集与测试集，分别用于模型训练、参数设置和结果测试，以确保模型应用的准确率和适用性。模型训练完成后，本文将其用于采集的全部政府微博评论中，实现大规模评论文本的自动标注，继而判断预测文本是否属于评论者观点，即判断标签是否属于评论对象、程度、观点和动作之一。

2.3 基于观点聚类的舆论焦点呈现

在上述观点抽取模型构建的基础上，本文保留属于评论者观点的文本，依据不同评论对象分别提取相关观点，并呈现舆论焦点。

1）基于 K-means 的评论观点聚类。为实现面向评论对象的政府微博舆论焦点呈现，本文借助 K-means 聚类算法[26]对评论不同对象的观点聚类。

K-means 聚类算法为常用的聚类算法[27]，具有相对可伸缩和高效率等优点[28]。该算法采用距离作为相似性的评价指标，即两观点距离越近、相似度越大。本文使用 K-means 算法评论同一对象的观点聚类，随机选取 K 个政府微博评论观点作为初始聚类中心，计算每个观点与各聚类中心之间的距离，从而把每个观点分配给距离它最近的聚类中心。聚类中心及其所分配的政府微博评论观点自动形成一个聚簇。每分配一个政府微博评论观点，K-means 算法将根据当前类中现有的所有对象重新计算聚类中心，该过程将不断重复直至所有观点被合理分配为止。

2）基于 TF-IDF 主题词提取的焦点呈现。为减少主观因素、更好呈现舆论焦点，本文利用 TF-IDF 提取不同聚簇的主题词，辅助人工判读总结评论观点，呈现公众对不同评论对象的舆论焦点。

TF-IDF[29]是较为常用和基础的主题词提取算法。该算法认为当给定词在某类政府微博评论观点中出现次数越多，而在所有政府微博评论观点中出现次数越少时，越能够代表该类评论观点的核心思想。其中，TF 衡量了给定词在该类评论观点中的重要程度；IDF 表达了给定词的类别区分能力。

因此，本文综合选择 K-means 实现基于评论对象的政府微博评论观点聚类，

TF-IDF 提取具有较强代表性的政府微博评论观点主题词，准确高效地为焦点呈现提供参考，也减少人工判读呈现焦点中的主观因素。

3 公安微博评论观点提取与舆论焦点呈现

为验证本文所提标注策略及所构建模型的有效性，本文选择公安部新闻中心与公安部治安管理局官方微博账号“中国警方在线”[30]为实验样本来源。将其微博“‘起侮辱性绰号’也属欺凌！严惩涉校违法犯罪！”[31]作为实证分析对象，进行观点抽取与舆论焦点呈现实验。

3.1 公安微博评论标注

本文共计抓取其评论数据 11987 条，清洗后得到可用数据 11360 条，结合自定义词典和 Jieba 库[32]对评论文本分词。依据所选评论领域，本文将评论对象细化为警方（Police）、校方（School）、施暴者（Perpetrator）、受害者（Victim）、暴力（Violence）4 个标签。其中，警方标签包括警方等政府部门，校方标签包括学校及校内教师等，施暴者指代对他人施以校园暴力或欺凌的人，受害者指代校园暴力事件或校园欺凌事件中的受害方，暴力则指代校园暴力事件或校园欺凌事件本身。结合细粒度四元组标注策略及细化后的评论对象，随机选取 5000 条分词后的评论文本进行人工标注。

3.2 公安微博评论观点抽取模型训练

本文使用采集到的全部公安微博评论训练词向量，并将上述标注好的公安微博评论用于观点抽取模型训练。

3.2.1 公安微博评论文本的词向量集构建　本文使用 Word2Vec 工具包训练词向量。其中，向量维度设置为 300 维，上下文窗口设为 5，最小训练词频设为 3，二次采样为 1e-5[33]。如表 1 所示，通过选定词语相似词的对比后发现，相似词与选定词语义相关度较高，词向量训练结果较好，可进行下一步实验。

表 1 词向量模型处理效果

选定词	相似词（相似度）
山西公安	晋中公安（0.8715246915817261）、太原公安（0.8598653078079224）、山西警方（0.8419739007949829）、晋中（0.8242400884628296）
施暴者	加害者（0.7612622380256653）、施暴人（0.7426093816757202）、罪犯（0.706917941570282）、恶人（0.6959208250045776）
太原师范学院	太原师范（0.7069997787475586）、山西师范大学（0.6703776121139526）、太原校园（0.6641935706138611）
感动	激动（0.9082546234130859）、好想哭（0.8390811681747437）、太好（0.8351080417633057）
保护	维护（0.7142535448074341）、帮助（0.6918110251426697）、捍卫（0.6716572046279907）

3.2.2 公安微博评论观点抽取模型实现　本研究使用 TensorFlow 框架[34]和 Python 中的 Keras 深度学习库[35]实现 POF-BiLSTM-CRF 模型。按照 6:2:2[36]的比例，将人工标注的 5000 条评论数据，随机分为 3000 条训练集、1000 条验证集和 1000 条测试集。

1）实验参数设置。为防止模型过拟合，使用 Dropout 率为 0.5 的正则化方法[37]。根据文本实际长度，将训练集的 batch_size 定为 32，测试集的 batch_size

定为 64。经测试，迭代次数定为 50，BiLSTM 中前后方向隐藏状态的维度定为 128[38]。同时选择优化器 Adam[39]最小化模型损失[40]，提升模型准确率。

2）模型评价。本文采用准确率（Precision，P）、召回率（Recall，R）和 F 值作为评价指标。准确率衡量了模型预测标签中正确的比例。召回率衡量了所有人工标注标签中被正确识别的比例。F 值综合考虑了准确率与召回率，是两者的调和平均值。评价指标越高，说明实验结果越好。这三种指标的计算公式如下所示：

$$P = \frac{t_{TP}}{t_{TP} + t_{FP}} \tag{1}$$

$$R = \frac{t_{TP}}{t_{TP} + t_{FN}} \tag{2}$$

$$F = \frac{2PR}{P + R} \tag{3}$$

式中，t_{TP} 为正确识别的标签数量；t_{FP} 为错误识别为该标签的数量；t_{FN} 为人工标注标签没有被系统正确识别的数量。

将预测结果与人工标注结果对比后发现，所有标签的准确率、召回率和 F 值均在 80%以上，部分标签达到了 90%以上（见表 2），说明模型的预测效果较好。

表 2 各标签预测效果

标签	准确率（%）	召回率（%）	*F* 值
警方	87.99	86.03	86.99
校方	93.23	81.78	86.79
受害者	91.67	91.12	91.39
施暴者	90.91	83.33	86.96
暴力	92.98	80.30	86.18
程度	88.31	80.42	84.18
观点	82.42	80.31	81.35
动作	81.95	84.07	83.00

3）常用模型对比。本文将所提模型与 CRF 模型和 LSTM-CRF 模型[41]在相同的实验环境和数据集下进行对比。对比结果（见表 3）表明，本文所构建的 POF-BiLSTM-CRF 模型具有更高的准确率、召回率及 *F* 值，对政府微博评论的观点抽取更加有效，具有较好的预测能力及优越性。

表 3 常用序列标注方法对比

模型	准确率（%）	召回率（%）	*F* 值
CRF	79.16	68.35	73.36
LSTM-CRF	84.15	80.72	82.41
POF-BiLSTM-CRF	86.46	82.37	84.36

3.3 公安微博评论舆论焦点呈现

本文对全部微博评论进行观点抽取模型计算后，部分结果见表 4。

表 4 政府微博评论部分预测结果

预测文本	我	真的	想说	一句	晋中
预测标签	O	O	O	O	B-Police
预测文本	公安	不作为	啊	，	芝芝
预测标签	I-Police	B-Opinion	O	O	B-Victim
预测文本	太	可怜	了	吧	！
预测标签	B-Degree	B-Opinion	O	O	O

如表 4 所示，评论“我真的想说一句晋中公安不作为，芝芝太可怜了吧！”抽取到的评论对象为“晋中公安”和“芝芝”，观点词分别为“不作为”和“可怜”，评论对象“芝芝”对应的程度词为“太”，抽取到的观点为“晋中公安不作为”和“芝芝太可怜”。

同时，本文面向不同评论对象分别抽取对应观点，将初始聚类数目 K 设置为 3，使用 K-means 分别对不同评论对象的观点聚类。继而通过 TF-IDF 分别提取出不同评论对象、不同聚簇的前 10 个主题词，结合含有主题词的观点，人工判读聚类结果，归纳出面向不同对象的多个舆论焦点（见表 5）。例如，有关“警方”这一评论对象的观点数共为 3520 条，其舆论焦点可总结为：“希望警方用法律还受害者公道”“警方终于发声，希望持续关注”和“山西公安关闭微博评论”。

表 5 基于评论对象的舆论焦点展示

评论对象	观点数	部分舆论主题词	舆论焦点	舆论焦点占比（%）
警方	3520	警方、学校、太原、希望、法律、山西、受害者、政府、公信力、公道	希望警方用法律还受害者公道	66.59
		大佬、终于、发声、山西、警方、太原、希望、关注、学校、管管	警方终于发声，希望持续关注	15.88
		公安、山西、评论、晋中、警方、关闭、微博、大佬、学校、公安部	山西公安关闭微博评论	17.50
校方	4503	学校、校园、老师、太原、学生、处理、受害者、真的、希望、同学	希望校方处理该事件	78.93
		没有、施暴者、受害者、严惩、真的、希望、学校、解决、暴力、校园	希望学校严惩施暴者，解决校园暴力	15.50
		威胁、学生、老师、应该、社会、高中、势力、报警、事情、学校	学校应报警，防止其他势力威胁受害学生	5.57
受害者	1754	希望、学校、校园、欺凌、可怕、保护、女孩、真的、救救	受害者应被好好保护	29.87

		受害者、学校、希望、交代、惩罚、跟进、继续、关注、不能、不了了之	希望持续关注受害者	20.98
		受害者、学校、施暴者、校园、公道、交代、保护、控制、调查、关注	希望调查后还受害者公道	49.14
施暴者	724	施暴者、起来、暴力、老师、裸露、法律制裁、室友、视频、学校、一定	施暴者太过嚣张	12.30
		施暴者、学校、受害者、希望、校园、惩罚、真的、暴力、得到、法律	施暴者涉嫌违法，应受到法律制裁	71.96
		严惩、施暴者、受害者、希望、交代、校园、一定、暴力、跟进、欺凌	施暴者应被公正处理	15.88
暴力	1592	校园、欺凌、霸凌、学校、事件、暴力、希望、关注、太原、法律	遭受校园欺凌等应向法律求助	61.12
		暴力、校园、解决、希望、严惩、真的、学校、抵制、受害者、网络	校园欺凌相关事件应被坚决抵制	28.33
		欺凌、绰号、侮辱性、孩子、严惩、可怕、希望、不要、伤害、欺负	起侮辱性绰号也属校园欺凌	10.55

4 结论与展望

本文进行了大规模政府微博评论中公众观点抽取和舆论焦点呈现研究，研究结果表明：所提出的细粒度政府微博评论标注策略与所构建的POF-BiLSTM-CRF 模型具有有效性与普适性，领域适应性较好，应用优势明显；能够自动学习政府微博评论文本特征，并准确标注评论对象、程度、观点和动作，基于不同评论对象准确迅速地呈现舆论焦点，对快速把握公众观点及决策提供了参考。

同时，本文研究仍存在情感指标缺乏和焦点呈现效率提升等问题。后续研究将考虑纳入评论情感因素，揭示观点的情感倾向；以及引入多种文本特征优化焦点呈现过程，进一步减少人工主观因素作用，更为全面高效地呈现政府微博舆论焦点。□

参考文献

[1] 中华人民共和国国家互联网信息办公室. 第 44 次《中国互联网络发展状况统计报告》[EB/OL]. [2020-03-13].http://www.cac.gov.cn/2019-08/30/c_1124938750.htm.

[2] 曾润喜,王君泽,杜洪涛.新媒体时代网络评论观点信息发现机制研究[J].图书情报工作,2015,59(14):111-116,148.

[3] 王连喜. 微博短文本预处理及学习研究综述[J]. 图书情报工作, 2013,

57(11):125-131.
[4] 冯小东,马捷,蒋国银.社会信任、理性行为与政务微博传播:基于文本挖掘的实证研究[J].情报学报,2019,38(9):954-965.
[5] 冯子昂. 政务微博影响力提升策略研究——以“共青团中央”官方微博为例[D].济南：山东师范大学,2019.
[6] 孟川瑾,卢靖.基于新公共服务的政务微博运行机制——“@问政银川”案例研究[J].电子政务,2016(4):45-53.
[7] 安璐,陶延芳.突发事件情境下政务微博的角色评估[J].情报工程,2019,5(4):19-32.
[8] 汪祖柱,阮振秋.基于关联规则的政务微博公众评论观点挖掘[J].情报科学,2017,35(8):19-22.
[9] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.
[10] 王序文,李姣,吴英杰,李军莲.基于 BiLSTM-CRF 的中文生物医学开放式概念关系抽取[J].中华医学图书情报杂志,2018,27(11):33-39.
[11] 徐飞,叶文豪,宋英华.基于BiLSTM-CRF模型的食品安全事件词性自动标注研究[J].情报学报,2018,37(12):1204-1211.
[12] 黄炜,黄建桥,李岳峰.基于 BiLSTM-CRF 的涉恐信息实体识别模型研究[J].情报杂志,2019,38(12):149-156.
[13] 张俊飞,毕志升,王静,吴小玲.基于BLSTM-CRF中文领域命名实体识别框架设计[J].计算技术与自动化,2019,38(3):117-121.
[14] 陈世梅,伍星,唐凡.基于 BiLSTM-CRF 模型的汉语否定信息识别[J].中文信息学报,2018,32(11):55-61.
[15] 张诗林.基于 Bi-LSTM 和 CRF 的中文网购评论中商品属性抽取[J].计算机与现代化,2019(2):93-97.
[16] 苏圣疃. 微博热点话题发现系统的设计与实现[D].上海：复旦大学,2014.
[17] 睢国钦,那日萨,彭振.基于深度学习和 CRFs 的产品评论观点抽取方法[J].情报杂志,2019,38(5):177-185.
[18] Python random() 函数[EB/OL].[2020-03-25]. https://www.runoob.com/python/func-number-random.html.
[19] 余云秀. 基于分层标注的地理领域嵌套命名实体识别研究[D].南京：东南大学,2018.
[20] YANG J, ZHANG Y, LI L, LI X. YEDDA: a lightweight collaborative text span annotation tool [EB/OL]. [2020-03-13].https://arxiv.org/pdf/1711.03759.pdf.
[21] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.
[22] MIKOLOV T. word2vec-toolkit[EB/OL].[2020-06-5]. https://groups.google.com/forum/#!searchin/word2vec-toolkit/c-bow/word2vec-toolkit/NLvYXU99cAM/E5ld8LcDxlAJ.
[23] RONG X. Word2vec parameter learning explained[J]. Computer Science, 2014.
[24] Understanding LSTM Networks[EB/OL].[2020-06-26].http://colah.github.io/posts/2015-08-Understanding-LSTMs/.
[25] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proc. 18th International Conf. on Machine Learning，2001:282-289.

[26]LLOYD S. Least squares quantization in PCM[J]. Information Theory, IEEE Transactions on, 1982, 28(2): 129-137.
[27]李芳. K-Means 算法的 k 值自适应优化方法研究[D].合肥：安徽大学,2015.
[28]赵松. 数据挖掘中基于遗传算法的 K-means 聚类算法的研究及应用[D].杭州：浙江工业大学,2014.
[29]SALTON G，BUCKLEY C.Term-weighting approaches in automatic text retrieval[J].Information Processing& Management, 1988, 24 (5) :513-523.
[30]中国警方在线[EB/OL].[2020-03-25]. https://www.weibo.com/dshcsh?topnav=1&wvr=6&topsug=1.
[31]中国警方在线.“起侮辱性绰号”也属欺凌！严惩涉校违法犯罪！[EB/OL].[2020-03-25]. https://weibo.com/2328516855/HwytOnQor?filter=hot&root_comment_id=0&type=comment#_rnd1585145055647.
[32]SUN J. Jieba[EB/OL].[2020-03-25]. https://pypi.org/project/jieba/.
[33]Gensim. models.word2vec – Word2vec embeddings[EB/OL].[2020-03-25]. https://radimrehurek.com/gensim/models/word2vec.html.
[34]Tensorflow[EB/OL].[2020-03-25]. https://tensorflow.google.cn/.
[35]Keras: 基于 Python 的深度学习库[EB/OL].[2020-03-25]. https://keras.io/zh/.
[36]吴恩达.模型选择和训练、验证、测试集[EB/OL].[2020-03-25]. https://study.163.com/course/courseLearn.htm?courseId=1004570029#/learn/video?lessonId=1051790987&courseId=1004570029.
[37]张应成,杨洋,蒋瑞,全兵,张利君,任晓雷.基于 BiLSTM-CRF 的商情实体识别模型[J].计算机工程,2019,45(5):308-314.
[38]武惠,吕立,于碧辉.基于迁移学习和 BiLSTM-CRF 的中文命名实体识别[J].小型微型计算机系统,2019,40(6):1142-1147.
[39]KINGMA D, BA J. Adam: a method for stochastic optimization[J]. International Conference on Learning Representations, 2014.
[40]张慧. 深度学习中优化算法的研究与改进[D].北京：北京邮电大学,2018.
[41]张聪品,方滔,刘昱良.基于 LSTM-CRF 命名实体识别技术的研究与应用[J].计算机技术与发展,2019, 29(2): 106-108,142.

作者简介：**胡吉明**（通讯作者），男，1985 年生，副教授，硕士生导师。研究方向：政务信息资源管理与服务。**郑翔**，女，1996 年生，硕士生。研究方向：政务信息服务。**程齐凯**，男，1989 年生，讲师。研究方向：自然语言处理与信息检索研究。**张岩**，男，1997 年生，博士生。研究方向：社会地理计算。
作者贡献声明：**胡吉明**，提出研究思路，设计研究方案，修改论文。**郑翔**，进行实验，撰写论文。**程齐凯**，讨论与分析实验结果。**张岩**，采集数据。
录用日期：2020-08-24