基于概率主题建模和深度学习的公众舆情分析*

劳鑫[1]，马宝君[1]，张楠[2]，万岩[1]
(1. 北京邮电大学 经济管理学院，北京 100876；2. 清华大学 公共管理学院，北京 100084)

摘　要：随着互联网的迅猛发展，公众逐渐开始通过社交媒体来发表自己对于社会事件的看法。在 Web 2.0 时代公众广泛参与的理念下，对上述信息的有效分析并将结果以合理形式展现对于舆情监测有着重要意义。鉴于此，本文提出了一种新颖的基于社交平台的舆情分析方法，采用概率主题模型 LDA 抽取公众对于某一事件的不同主题观点，利用深度学习的 word2vec 模型计算出每条文本的情感强度，分别对事件整体以及事件下各潜在主题进行情感强度时序分析，跟踪事件发展过程中公众的情感强度变化，并通过案例分析验证了该方法的合理性与有效性。

关键词：公众舆情分析；概率主题模型；深度学习；情感强度；时序分析

1　引言

随着互联网的不断普及与发展，公众的参与性正在逐步提高，他们不仅仅只是从网络上获取信息，更多地开始参与信息的发布、表达自己的看法与见解。特别是随着各类社交平台的广泛应用，越来越多影响力较大的社会热点事件都是通过社交媒体爆料而引发公众关注，社交平台开始作为一项重要的互联网舆情源，网络舆情的“微内容”程度化日益加剧。在这种背景下，挖掘社交平台上公众对于某一热点事件的意见与看法，并根据事件发展对公众情感强度的变化趋势进行分析与预测，对于舆情监测预警具有非常重要的意义。

由于社交平台上相关数据量的巨大规模，人们往往无法在短时间内详细了解公众对于某一社会热点事件的不同观点。并且随着时间的推移，公众对于该事件也会产生新的认识与看法。因此，随着信息技术的不断发展，基于舆情分析的需要，能够简单、全面、快捷、有效地了解公众对于某一社会事件的讨论，获取公众讨论内容的潜在主题，就显得愈发重要。此外，随着对应事件的不断发展，公众对其的情感强度也会随之改变。特别是在当事人对该事件表态时，公众情感强度往往会发声较大的变化。这不仅需要运用相关模型对与事件相关的各类讨论进行情感值计算，还需要根据时间段划分对其进行时序分析，并结合社会学与政治学的有关知识进行进一步讨论。由于社会热点事件常常较为复杂，并且公众对于该类事件一般存在着不同的观点，无论是对于社会热点事件的发展走势，还是对于社会舆情的分析与管理，甚至是对于有关人员解决该事件的方法，对公众情感强度变化进行分析和预测都是极其重要的。

以往的相关研究更多地是关注于热点事件的发现，而对于热点事件的情感强度追踪关注较少。本文从概率主题模型以及深度学习的视角出发，结合时间序列分析方法，提出了一种新颖的基于话题情感强度的社会事件舆情分析方法，实现了基于社交平台的社会热点事件追踪。通过 LDA 模型，发现公众对于某个热点事件的不同观点与看法，并将其分布到各个对应的主题当中；通过 word2vec 模型，计算对应语料中所有词的词向量，通过计算余弦距离的方式获取情感词及其对应的情感强度；通过时间序列分析方法，将各时间段内各项数据的情感强度进行累加，跟踪事件发展过程中公众的情感强度变化情况。本文选取了 2015 年 5 月发生于四川成都的“女司机变道被暴打”事件作为研究案例，并抓取了新浪微博上关于该事件的相关评论进行分析，展示了初步分析的结果以及该研究方法的有效性。

2　相关工作

由于本文所提出和运用的研究方法本质是运用 LDA 模型与 word2vec 模型，结合情感词典，对社会热点事件进行公众情感强度的时序分析，因此该部分的内容将围绕概率主题模型、神经网络语言模型以及基于情感词典的情感分析三部分来展开论述。

*基金项目：国家自然科学基金(71402007, 71473143, 71471019)；中央高校基本科研业务费专项资金(2014RC0601)
作者简介：劳鑫，男（汉族），硕士研究生；马宝君，男（汉族），讲师；张楠，男（汉族），副教授；万岩，女（汉族），教授。
通讯联系人：马宝君，讲师，E-mail: mabaojun@bupt.edu.cn

2.1 概率主题模型

概率主题模型是主题建模方法中最为常见的一种，其本质是运用统计学的方法和理论，从大量的文本信息中发现并提取主题信息，在信息检索领域有着广泛的应用[1]。

作为最为初始的文本表示模型，TF-IDF 模型[1]以及空间向量模型[3]可以通过对文档中所包含信息的间接处理，对文档内容及相互之间的相似度进行初略描述与建模。这两种模型所运用的数学原理较为简单，无法有效区别含义相似的词语或文档。为了解决这个问题，有关学者又引入了奇异值分解（Singular Value Decomposition）的计算方法，提出了潜在语义分析（Latent Semantic Analysis, LSA）模型，将高维的文档词语空间转化映射到了低维的文档词语空间[4]。相比于该模型中运用启发式的物理距离表示主题，随后提出的概率潜在语义分析（probabilistic Latent Semantic Analysis ,pLSA）模型在概率框架下生成文档的主题集合[5]。由于不需要进行复杂度较高的奇异值分解，pLSA 方法被广泛应用于许多信息检索领域的大规模数据建模与分析当中。

相比于前面所说的模型，潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）对 pLSA 模型与方法进行了进一步改善，运用了三层贝叶斯框架，假定文档到主题以及主题到词均服从多项式分布，是一种更为有效的概率主题模型，其建模效果也要明显优于其他模型，可以用来识别大规模文档集或语料库中潜藏的主题信息。因此，本文的研究方法即采用 LDA 模型来提取和分析文本中的主题信息。

2.2 神经网络语言模型

统计语言模型的目标是学习语言中单词序列的联合概率函数，其最大的难点在于维度灾难。神经网络语言模型的出现则有效地解决了这个问题。最早的神经网络语言模型是由 Bengio 系统化提出的 NNLM（Neural Network Language Model）[6]，其基本思想是通过对模型的不断训练与优化，使得语料库中的每个单词获取一个分布式表示，不仅去除了维度灾难，还可以让模型能够了解语义相近的句子的数量。神经网络语言模型包含两大核心部分：一个是分布式表征[7]，即所谓的词向量（Word Embedding）;另一个是运用神经网络建立语言模型[8]。

神经网络模型在近十多年取得了较大的发展，所提出的各类模型也都是以 NNLM 作为模板，例如比 NNLM 更为简单的 CBOW 模型[9-10]、Skip-gram 模型[9-10]等。此外，为了使模型的训练更为快速有效，又出现了相应的 Hierarchical Softmax 算法[11]、Negative Sampling 算法[12]等。2013 年 Google 公司所开放的 Word2vec 深度学习工具则集合了上述模型与算法，能够快速有效地将语料库中词语以词向量的形式表示，获取词与词之间的语义相关性，为神经网络语言模型在各个领域的广泛应用提供了更为有效的方法[13]。基于 word2vec 的有效性，及其能准确捕捉词与词之间的潜在语义相似性，本文的研究方法即基于 word2vec 工具来计算文本的情感强度。

2.3 基于情感词典的情感分析

早在上世纪 90 年代初，国外就有相关学者通过情感词典进行文本的情感分析。Riloff 与 Shepherd 是最早进行相关研究的学者，提出了基于语料库数据来构建语义词典的方法[14]。Hatzivassiloglou 和 McKeown 在考虑了大规模语料数据中形容词语义情感倾向的限制性影响的基础上，尝试对单词的情感倾向进行判断[15]。在此之后，越来越多的研究开始关注情感词或情感短语与特征词之间的关联。Turney 等人则使用 PMI 方法，通过计算语料中非情感词的情感倾向使得情感词典得到了扩展，并运用语义极性算法（Semantic Polarity Algorithm）分析文本情感，最终取得了 74%的准确率[16]。

我国在这方面的研究也取得了较大进展。朱嫣岚等人基于 HowNet 情感词典，分别提出了基于语义相似度和基于语义相关场的词汇语义倾向性计算方法，判别准确率可达 80%以上[17]。李钝、曹付元等人从语言学的角度出发，采用“情感倾向定义”权重优先的计算方法获取短语中各词的语义倾向度，同时分析短语中歌词组合方式的特点，提出“中心词”的概念来对各词的倾向性进行计算来识别短语的倾向性和倾向强度[18]。白雪等人则在运用 word2vec 将语料中的所有词转换为词向量后，使用基于 PMI 方法改进的 SO-SD 算法计算词与情感词之间的语义距离，判断词的情感倾向，构建微博情感词典[19]。随后，基于所构建的微博情感词典，结合微博中表示程度的副词、感叹词、否定词以及表情图标等，对微博的情感倾向进行分类。其结果表明情感强度越高的微博，分类效果越好。

总体来说，基于情感词典的情感分析方法准确度较高，并且能够进行细粒度情感的相关研究。然而局限于自然语言处理技术以及相关的数据提取方法，该类方法难以发现和获取数据当中的隐藏信息，其后续研究具有很大的发展空间。

3 研究方法与框架

为了能够有效研究社交平台上公众对于社会

热点事件的情感强度变化趋势，本研究结合了潜在狄利克雷分配（Latent Dirichlet Allocation，简称 LDA）以及 word2vec 深度学习工具，设计了社交平台上基于概率主题建模以及深度学习的情感强度分析框架。在该分析框架当中，LDA 模型被用来发现从社交平台上与某一热点事件相关的信息中发现和标注相应的主题信息，找出公众对于某一事件的不同观点与看法；word2vec 工具则可将所有词转化成向量的表现形式，基于情感词典计算得出潜在表达的情感表达词，从而获取每条信息的情感强度，并将其应用到主题分布当中，观察各个主题下情感强度随时间的变化趋势。基于此，我们不仅能够结合对于相关事件的报道准确预测公众舆情的情感强度走势，还可以有效地指导有关部门如何有效地引导舆情走势。

3.1 潜在狄利克雷分配模型（LDA model）

概率主题模型（Statistical Topic Models）是一类从文本文档中提取潜在语义信息的有效方法[1,4]，近年来得到非常广泛的应用，在文本分类、信息检索等相关领域取得了非常好的应用效果。概率主题模型的基本原理认为文档是若干主题的混合概率分布，而每个主题又是一个关于单词的混合概率分布，可以看作是文档的一种生成模型。在概率主题的各项方法当中，潜在狄利克雷分配模型（LDA model）是最为有效的模型之一[1]。

LDA 模型是一项无监督的生成统计模型，其目的是通过提出一种文档中词语生成的随机过程，找到文档的主题信息。该模型应用了贝叶斯统计理论中的标准方法，在确定了最优主题数的基础上，通过利用 MCMC（Markov chain Monte Carlo）中的 Gibbs 抽样进行推理，计算得到模型的相关参数，从而获取文本在主题集，以及词语在各个主题下的概率分布。概括地说，LDA 模型是在 PLSI（Probabilistic Latent Semantic Indexing）模型的基础上，用服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题混合比例，模拟文档的产生过程[20]。所有主题在任何文档生成之前就已经确定，因此在文档的生成过程中，只需通过随机抽样确定每篇文档的主题多项式分布，然后反复抽样，根据该文档主题的分布概率产生每篇文档中的词语。可以发现，每篇文档中的每个词语都是基于不同的、随机选择的主题所产生的。

总体来说，LDA 模型不需要对文档做任何人工标注，所有的主题信息都是通过对原始文档集合的分析所获取。LDA 模型不仅可以帮助我们了解和分析大规模文本的主要内容，还可以得知这些文本的主题分布情况，而完全依靠人工标注来完成这些工作是极其困难的。在本文的研究框架中，LDA 模型主要被用来对相关的文本数据进行主题建模，获取这些数据所对应的主题内容以及主题分布情况，作为下一步计算各个主题下情感强度的变化情况的基础。

3.2 word2vec

随着计算机应用领域的不断发展与扩大，以及非结构化数据的爆发式增长，自然语言处理逐渐开始受到人们的高度重视，各种处理自然语言的数学模型也纷纷应运而生。自然语言建模的方法经历了从基于规则的方法到基于统计方法的转变[21]。通过基于统计建模方法所获取的自然语言模型称为统计语言模型，较为常见的统计语言模型包括 n-gram 模型、神经网络语言模型等，这些模型普遍存在维数灾难、词语相似性难以表达、模型泛化能力不足等相关问题。

在不断深入对统计语言模型的研究、以求提高计算机的自然语言处理能力的背景之下，2013 年 Google 公司开发了一款基于深度学习的工具 word2vec[13]。该工具可以根据给定的语料库，通过不断优化训练模型的方式获取语料库中每个词语的向量表达形式，并通过计算词向量之间的余弦距离来表示词与词之间的语义距离，为自然语言处理领域的应用提供了更为有效的方法。

word2vec 包括了两种词向量训练模型，分别是 CBOW 和 Skip-gram。两者均包含输入层、投影层以及输出层，不同的是 CBOW 模型是通过上下文来预测当前的词，而 Skip-gram 模型则是通过当前的词来预测上下文。同样，word2vec 也包括两种词向量优化模型，分别是 Hierachy Softmax 模型以及 Negative Sampling 模型。其中 Hierachy Softmax 优化方法是以词语在语料库中的词频作为权值构造一棵二叉树，并将叶子节点对应到所有词语，而 Negative Sampling 优化方法则是通过采用相对简单的复采样来提高词向量的训练速度。通过对两种训练模型与两种优化方法的组合，一共可得到四种训练词向量的框架。

作为一款用于训练词向量的工具，word2vec 不仅能够将语料库中的词语快速高效地表达为向量，而且能够捕获词语之间的语义特征与相似性，从而可以供其他相关的应用研究使用，有效地推动了自然语言处理领域相关研究的发展。在本文的研究框架中，我们利用 word2vec 可以获取潜在表达情感的隐性情感词，并通过计算其与情感词典中显性情感词的余弦距离来定义情感强度，进而得到每条文本数据的情感强度，并最终运用于各个主题下情感强度随时间变化情况的相关分析中。

3.3 具体研究流程

本文所提出和应用的基于概率主题建模以及深度学习的情感强度分析方法包括四大模块（如图 1 所示）：根据所选定的热点事件，从社交平台上爬取与其相关的文本数据，并对其进行预处理；运用 LDA 模型对所采集的文本数据进行主题建模，并通过人工分析选取有意义或感兴趣的主题；基于显性情感词，运用 word2vec 获取文本数据中的隐性情感词，并计算每条文本的情感强度；结合前两项工作的成果，在所选取主题下进行情感强度时序分析。

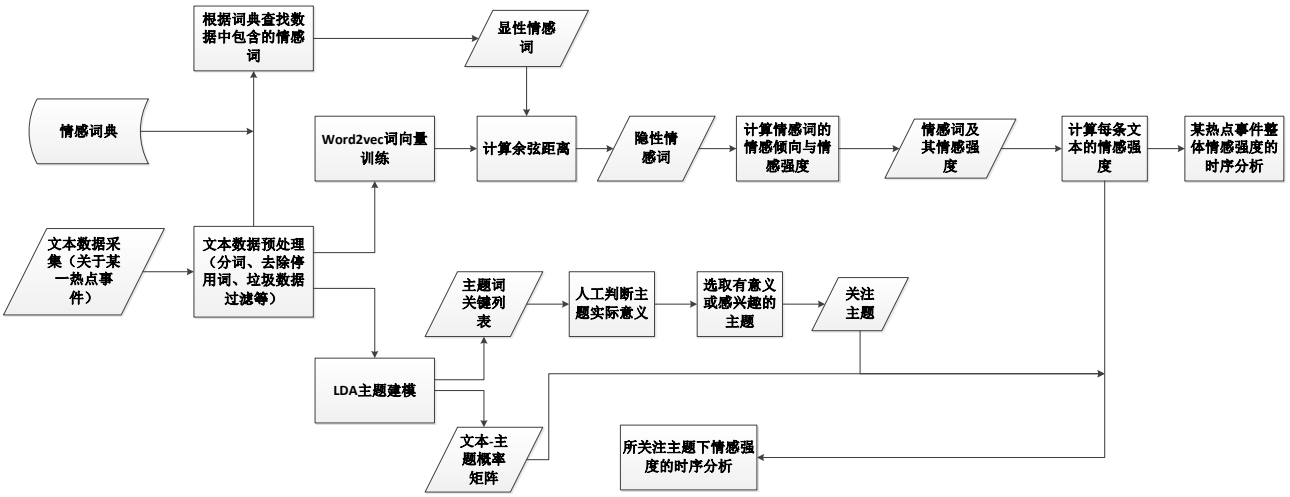

图 1　情感强度分析具体流程

1） 数据抓取与预处理

主要包括文本数据采集与文本数据预处理两项工作。

文本数据采集主要是用爬虫的方法，从所关注的社交网络平台（微博、论坛等）上抓取与热点事件有关的具有时序性的数据。

在抓取数据之后，为了能够有效地将数据导入到后续分析的各个模型当中，需要对所抓取的数据进行分词、垃圾信息过滤、去停用词等相关预处理工作。

2） 主题建模与筛选

在获取数据并对数据进行预处理之后，利用 LDA 模型进行概率主题建模。LDA 模型对所有的文本向量进行训练和推理，提取出蕴含在这些文本数据中的潜在主题信息[1]。通过概率主题建模，可以获取以下两组有用的信息：①主题关键词列表：表示与各个潜在主题最为相关的一些词语；②文档-主题概率矩阵：在该矩阵中每行对应一条文本，每列对应一个潜在主题。矩阵中的数值即表示某条文本属于某个对应主题的概率。

在得到以上两项信息之后，需要人工对主题关键词列表进行观察与判断，找出各个主题所表达的内容与实际含义，并筛选出若干有意义或者感兴趣的主题，以进行后续分析[22]。此外，为了验证所选取的主题内容是否与实际情况相符合，可以结合文档-主题概率矩阵，找出相对应主题下概率最高的若干条文本，并对这些文本的含义进行观察确认，最终确定该主题的内容信息。

3） 情感强度计算

情感强度的计算是与 LDA 主题建模同时进行的。相比于主题建模与筛选，情感强度计算的工作量更大，步骤更为繁琐，所产生的数据量也更多。

在通过数据预处理将文本数据表现为若干词语组成的向量之后，结合情感词典，找出所有文本数据中所包含的情感词，即显性情感词。同时，将文本数据导入到 word2vec 模型当中，获取文本中所有词的词向量形式。遍历除去显性情感词以外的其他所有词，通过计算余弦距离的方式，找出与该词最为相关的若干个词，并观察这些词中是否包含显性情感词，如果包含则认为该词可能在表达上具有一定的情感，并作为隐性情感词保留，反之则认为该词是中性词。随后，对于所有隐性情感词，运用文献[19]中所提出的 *SO-SD* 方法判断隐性情感词的情感倾向，并将所求得的值作为对应隐性情感词的情感强度。*SO-SD* 的计算公式如下：

$$SO-SD(word) = \sum_{pword \in Pwords} SD(word, pword) - \sum_{nword \in Nwords} SD(word, nword) \quad (1)$$

其中 *pword* 表示与 *word* 最为相关的若干个词中

所包含的某个正向显性情感词；*Pwords* 表示与 *word* 最为相关的若干个词中所包含的全部正向显性情感词；*nword* 表示与 *word* 最为相关的若干个词中所包含的某个负向显性情感词；*Nwords* 表示与 *word* 最为相关的若干个词中所包含的全部负向显性情感词。

此外，其中的：

$$SD(word1, word2) = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^{2}} \sqrt{\sum_{k=1}^{n} x_{1k}^{2}}} \quad (2)$$

式中 n 为词向量的维度，x_{1k} 为第一个词向量中第 k 维度的值，x_{2k} 为第二个词向量中第 k 维度的值。

对于计算得到的 *SO-SD* 值，我们采用 p 和 q 作为判断阈值：

$$SO-SD(word)\begin{cases} > p \text{ 该词为正向隐性情感词} \\ \in [q, p] \text{ 该词为中性词} \\ < q \text{ 该词为负向隐性情感词} \end{cases} \quad (3)$$

而对于所有的显性情感词，将其所有的正向词情感强度用+1 表示，负向词情感强度用-1 表示。随后，根据所获取的所有情感词以及各情感词所对应的情感强度，便可计算得到每条文本的情感强度。

4）情感强度时序分析

情感强度时序分析包括两种：对热点事件整体的情感强度进行时序分析，以及对热点事件对应主题下的情感强度进行时序分析。对于前者，只需在完成每条文本的情感强度计算后，分别累计各个时间段的情感强度，便可实现对于事件整体的情感强度时序分析；而对于后者，针对所关注的主题，则需要将各条文本的情感强度与其属于该主题的概率相乘，得到文本在该主题下的情感强度，再累计各个时间段的情感强度，便可对该主题下的情感强度进行时序分析。

3.4 研究方法的效率讨论

由于本研究方法的两大核心模块是 word2vec 与 LDA 模型，因此需要对这两个模型的时间复杂度等进行相关讨论，从而对本研究方法的效率进行详细讨论。

从优化与效率提升的角度上来说，相比其他神经网络语言模型，Word2vec 的最大特点是运用了层次 softmax 函数。所谓的层次 softmax 函数即是运用霍夫曼编码构造二叉树，并对每个词赋予了一条唯一路径，即这个词所对应的编码。如果没有运用这个二叉树，而是直接从隐层直接计算每一个输出的概率——即传统的 softmax，就需要对语料中的 V 个词语都算一遍，这个过程时间复杂度是 O(|V|)的。而使用了霍夫曼二叉树，其时间复杂度就降到了 $O(\log_2(|V|))$，速度明显加快。在情感强度计算的其他步骤中，其时间复杂度都要大于 $O(\log_2(|V|))$。其中隐性情感词计算的时间复杂度最高，达到了 $O(|V|^2)$。

LDA 模型的复杂度则取决于语料中词的数量、所设置的主题数量以及所运用的优化迭代方法。在单次迭代中，LDA 的时间复杂度为 O(K*|V|)，其中 K 为主题数量，|V|为词语的数量。而相比于[1]中所用到的 EM 算法，吉布斯采样的速度更快，时间复杂度更低，其所产生的样本也更加服从真实样本的分布，因此在本文中所运用的 LDA 模型通过吉布斯采样来进行迭代优化。

在本研究方法中，情感强度计算与 LDA 主题建模是同时进行的，因此该方法的运行效率取决于时间复杂度更高的模型。无论是 LDA 模型还是 word2vec 模型，其时间复杂度都与语料中词语的数量相关。因此可以说，语料中词语的数量在很大程度上决定了该方法的效率，同时语料中文本数量、主题建模中所设定的主题数等也对其有一定的影响。

4 案例分析

4.1 案例背景

为了展示该研究方法的有效性，我们选取了近期讨论较为广泛的“成都女司机被打事件”作为案例进行分析。2015 年 5 月 3 日下午，成都市三环路娇子立交桥附近发生一起打人事件，卢女士在驾车前往三圣乡途中，因行驶变道原因在娇子立交被张某驾车逼停，随后遭到殴打致伤。该事件的视频在网络上发布后，网友纷纷对男司机张某下手之凶狠表示震惊。然而随着其行车记录仪的视频公布之后，公众舆论转而谴责女司机的开车方式。特别是随着女司机的父母纷纷站出替其女儿辩解，该事件的关注热度不断升级，部分网友甚至对女司机卢某进行“人肉搜索”，其各类相关信息陆续被曝光。除了对事件双方的指责与批判之外，也有人开始对此事件中所出现的“个人隐私”、“行车安全”等社会问题进行反思。一时间，关于该事件的各种讨论与意见充斥着网络。

之所以选择“成都女司机被打事件”作为案例来验证本文所提出研究方法的有效性与合理性，原因有以下三点：①社会舆论存在转变：从一开始一边倒地认为男司机下手凶狠，到行车记录仪视频公布之后指责女司机的开车方式，再到后来女司机的父母亲纷纷接受采访替自己女儿辩解，引发网友对女司机的进一步批判与谴责，公众对于该事件的态度与情绪一直随着事件的发展发生着较为明显的变化；②数据量较为充足：在该事

件发生后，各大门户网站纷纷创建了对该事件的专题报道，每条相关的新闻都有上千条评论。以新浪微博为代表的各大社交平台上也充满了众多网友对其的观点与看法；③观点差异较为明显：由于该事件后续发展的时间长、变化大，不同时期网友对于该事件的观点与看法也存在着较大差异，适合于运用主题建模的方法来找出公众讨论的潜在主题，并对其进行进一步探讨与分析。

4.2 数据采集与整理

本研究的案例数据来自于新浪微博上与该事件相关的所有兴趣主页中的微博数据，这些数据都会带有一个与之相关的 hashtag（例如：#成都女司机变道遭殴打#）。运用基于 AJAX 的定址网络爬虫对相关的 url 地址进行 HTML 解析，我们抓取了相应兴趣主页上的全部微博数据，并选取了发布日期在 5 月 17 日之前的微博，共得微博数据 11899 条。

在获取了全部有效数据的基础上，首先需要对数据进行预处理。我们发现，有一些微博的内容是完全或基本相同的，因此需要使用文本匹配的方法将内容完全重复的微博删除，只保留其中的一条。此外，微博中含有大量 hashtag、表情图标等相关内容，在后续的研究并没有太大的意义，因此还需要将微博中所含的这些内容删除。最后得到实际有效的微博 6989 条。

接下来，将处理完后的微博内容作为每一条微博的微博信息，对每一条微博信息进行分词、去除停用词，将其表示称为由若干词语组成的向量，为使用 LDA 模型进行主题建模以及使用 word2vec 模型计算词向量做好数据准备。在此过程中，由于微博信息中的内容基本为中文，因此使用了 ICTCLAS 工具包进行分词处理；去除停用词则采用了哈工大的停用词词表。此外，在后续的 word2vec 情感强度计算过程中，将词向量的维度设置为 50，所使用的情感词典为台湾大学所发布的情感词典 NTUSD。包括数据抓取在内的绝大部分工作都是通过 Java 编程实现。

4.3 结果与分析

1） 情感强度时序分析

对于 6989 条微博数据，我们使用 LDA 模型，分别以 5—15 作为潜在主题数进行了概率主题建模。通过人工阅读不同主题数目下的主题关键词列表，可以发现主题数目为 11 的情况下主题词列表的语义信息表现最好。我们选取了其中最具有代表性的 5 个潜在主题的长度为 10 的关键词列表，如表 1 所示。

表 1　案例中 5 个潜在主题的关键词列表

Topic0（从法律角度批判男司机）	Topic1（女司机母亲辩解）	Topic2（女司机被人肉）	Topic6（女司机病情加重）	Topic8（男司机下手凶狠）
人	女司机	人肉	称	女司机
法律	视频	转发	女司机	下手
社会	慈善	中国	舆论	脸
暴力	母亲	开车	女儿	凶狠
危险	机构	删除	加重	变道
交通	女儿	挑衅	病情	男司机
文明	搞	道德	母亲	显示
强行	别车	四川	别车	视频
马路	辩解	女人	呕吐	脑震荡
网络	采访	开房	发高烧	骨折
……	……	……	……	……

通过 LDA 主题建模，我们对公众针对该事件的各类观点与看法有了初步的了解与认识。在随后的研究与分析当中，将结合主题建模所得结果以及后续的情感强度计算，对该主题的情感强度变化进行相应的时序分析。

2） 情感词典构建

根据 3.3 节中所介绍的方法， 我们共从微博数据中共找出 612 个正向显性情感词和 1315 个负向显性情感词。同时，利用 word2vec 与余弦距离计算公式，选取与词语最为相关的 20 个词，将词性判断阈值分别设置为 0.05 与-0.05，运用 *SO-SD* 方法计算得到了 1343 个隐性正向情感词以及 5533 个隐性负向情感词。限于篇幅，表 2 中分别只展示了较为明显的 10 个正向隐性情感词和负向隐性情感词，以及这些词所对应的情感强度。

结合所获得的显性情感词与隐性情感词，我们便构建了与该热点事件相关的微博情感词典。它是计算每条微博情感强度的基础。

表 2　从微博数据中发现的部分隐性情感词

正向隐性情感词		负向隐性情感词	
合情理	0.328784	心如蛇蝎	-1.553315
耐心	1.362254	违法乱纪	-0.933532
坚守	1.215394	错误百出	-2.803829
正名	1.177136	不规范	-2.271391
理所当然	1.147314	坎坷	-2.188523
霸气	0.595432	轻狂	-2.182535
温良恭俭	0.662376	酒驾	-1.943230
有理有据	0.309414	大打出手	-1.874091
以理服人	0.621868	病入膏肓	-1.872049
规范化	0.588630	吃闭门羹	-1.226578

3） 情感强度时序分析

利用所构建的情感词典以及各情感词所对应的情感强度，便可计算得到每条微博的情感强度，

并根据微博的发布时间将其划分到不同的时间段中，结合每条微博属于各个主题的概率值，对相关主题的情感强度进行时序分析。

仍然以前文中所提及的 5 个代表性主题作为主题情感强度的分析案例，以每 12 小时作为一个时间段，通过绘制时间序列图的方式，我们便可大致了解情感强度随着时间变化的情况（见图 2）。

其中，横轴上各个点所对应的时间段，以及该事件后续发展的几个关键时间点分别如表 3 和表 4 所示。结合图 2 以及表 3 与表 4 的内容，下面我们将展开相关分析与讨论。

由于公众针对该事件的负面评价较多，各条微博的情感强度大多是为负值。因此相比于情感值的正负，我们更为关注情感值随时间的变化情况。结合图 2 中各主题的情感强度变化情况，可以发现其与事件的后续发展紧密相关。

对于谴责男司机下手凶狠的主题（topic8），从事件曝光开始就具有很高的情感强度。而随着男司机道歉以及记录仪视频的公布，该主题下的情感强度迅速下降。这说明在事件曝光之后，绝大多数网友均认为男司机下手过于凶狠，对其行为纷纷进行谴责，因此情感强度较高。随后，在男司机对此进行道歉，特别是记录仪中视频表明女司机的开车习惯确实较为恶劣，矛头纷纷指向女司机一方，对于男司机的批判也就大大减少。而在此之后，网友开始对女司机进行“人肉搜索”，女司机表示要追究法律责任，与此相关的主题（topic2）情感强度在时间点 5 立即上升，网友对其的指责与批判愈发强烈。随后，女司机的母亲开始为其女儿进行辩解。她先是称女司机变道是为了赶着去做慈善，随后又在当天晚上称因为舆论的压力导致女司机的病情加重。然而在第二天，医院方面却否认女司机的病情有所加重。这一系列事情的发生也使得有关主题的情感发生了较为明显的变化：在女司机的母亲为其辩解之后，可以看到与之相关的主题（topic1）情感强度立即上升，表明网友并不相信其母亲的说法；而在医院否认女司机病情加重之后，对应的主题（topic6）情感强度也有所上升，网友纷纷讽刺与谴责女司机及其母亲对公众的欺骗。而由于女司机道歉的时间太晚，此时公众的关注热度已经很低，公众的情感强度并未因此而明显增强。

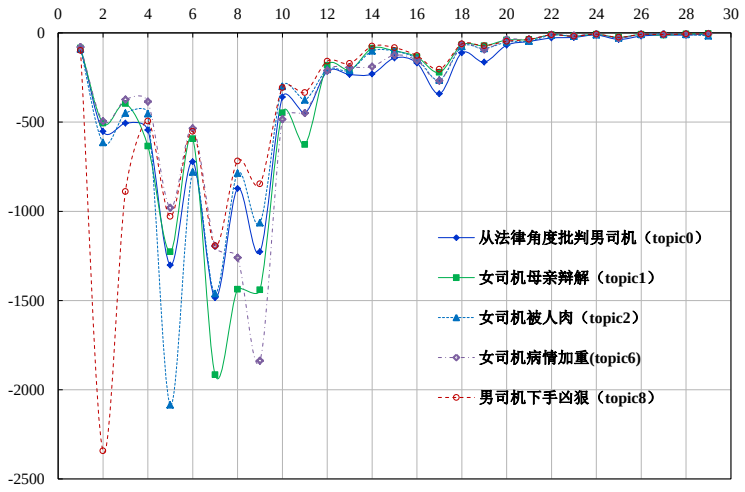

图 2　各主题情感强度随时间的变化图示

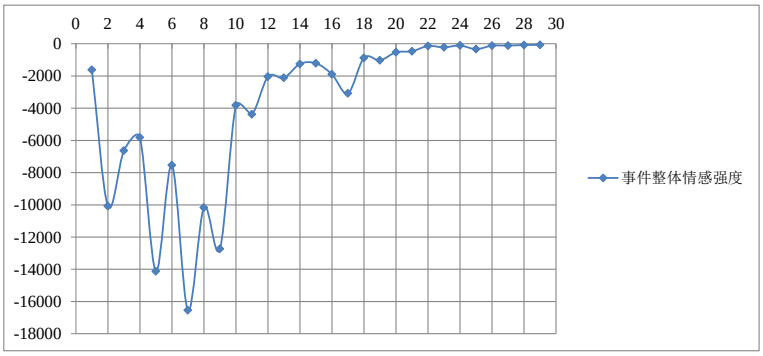

图 3　事件整体情感强度随时间的变化图示

表 3　各时间点所对应的时间段

时间点	对应时间段	时间点	对应时间段
1	5/3 12:00-24:00	16	5/11 0:00-12:00
2	5/4 0:00-12:00	17	5/11 12:00-24:00
3	5/4 12:00-24:00	18	5/12 0:00-12:00
4	5/5 0:00-12:00	19	5/12 12:00-24:00
5	5/5 12:00-24:00	20	5/13 0:00-12:00
6	5/6 0:00-12:00	21	5/13 12:00-24:00
7	5/6 12:00-24:00	22	5/14 0:00-12:00
8	5/7 0:00-12:00	23	5/14 12:00-24:00
9	5/7 12:00-24:00	24	5/15 0:00-12:00
10	5/8 0:00-12:00	25	5/15 12:00-24:00
11	5/8 12:00-24:00	26	5/16 0:00-12:00
12	5/9 0:00-12:00	27	5/16 12:00-24:00
13	5/9 12:00-24:00	28	5/17 0:00-12:00
14	5/10 0:00-12:00	29	5/17 12:00-24:00
15	5/10 12:00-24:00		

表 4　“成都女司机被打事件”的关键时间点

时间	发生事件
5 月 3 日 18:22	事件曝光
5 月 4 日 18:46	男司机对此道歉
5 月 4 日 20:22	记录仪视频曝光，舆情反转
5 月 5 日 13:01	女司机被人肉，要求追究网友的法律责任
5 月 5 日 15:41	女司机父亲表示不接受道歉
5 月 6 日 17:38	女司机母亲称其变道是为了去做慈善
5 月 6 日 22:20	女司机母亲称其病情恶化
5 月 7 日 11:49	医院方面否认女司机病情加重
5 月 11 日 10:42	女司机对此道歉

可见事件整体情感强度的时序变化与主题情感强度的时序变化相对应，其变化情况与事件的后续发展紧密相关，符合实际情况，可认为该研究方法是合理有效的。

总体来说，通过分析可以发现，本文所提出的基于概率主题模型与深度学习模型的情感强度分析方法能够较为准确地获取某一社会热点事件下公众的不同观点与看法，同时对各观点以及事件整体的情感强度随时间的变化情况进行合理有效的分析。该方法不仅有效地弥补了传统舆情分析方法无法获取情感变化趋势的不足，而且还可以根据公众的不同观点与看法，针对该事件下各个主题进行情感分析，能够更为精确地计算与分析公众舆情及其情感，对于舆情的监控与管理具有重要意义与价值。此外，该方法对于事件当事人还具有一定程度上的指导意义，可以帮助其通过合理的应对方法避免公众舆论的攻击。

5　结语

本研究以对社交平台上公众对于社会热点事件的观点与看法进行分析为目标，基于概率主题建模的 LDA 模型以及基于深度学习的 word2vec 模型，提出了一种新颖的从大规模舆论数据中提取具有代表性或有意义的主题进行情感强度时序分析的研究框架，并以“成都女司机被打事件”作为案例，结合其实际相关数据进行分析，通过案例分析的形式充分体现了本文所提出研究方法的合理性与有效性。

基于所存在的一些问题，未来的研究可以针对以下几个方面展开：①采用相关方法更为合理地标注和设置情感词的情感强度，使得情感强度的计算更为精确；②考虑结合热点事件相关微博下的评论与转发数据，使得微博数据的表达更为全面与准确；③考虑到微博的短文本特性，后续研究可以考虑运用或设计更适合于短文本主题建模的相关模型，从而使得研究结果更为严谨客观。

参 考 文 献

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] Jones K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, 28(1):11-21.

[3] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM, 1975, 18(11):613-620.

[4] Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[5] Hofmann T. Probabilistic latent semantic analysis[C]. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc., 1999: 289-296.

[6] Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3(6):1137--1155.

[7] Hinton G E. Learning Distributed Representations of Concepts[C]// In Proceedings of CogSci. 1986.

[8] Xu W, Rudnicky A I. Can artificial neural networks learn language models?[J]. In International Conference on Statistical Language Processing, 2000.

[9] Mikolov, Kopeck J, Burget L, et al. Neural network based language models for highly inflective languages[C]// Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Computer Society, 2009:4725 - 4728.

[10] Hinton G E, Mcclelland J L, Rumelhart D E. Distributed representations[J]. Parallel Distributed Processing Eds Rumelhart Et Al, 1986:77 - 109.

[11] Morin F, Bengio Y. Hierarchical probabilistic neural network language model[J]. Aistats, 2005.

[12] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Eprint Arxiv, 2013.

[13] word2vector: https://code.google.com/p/word2vec/.

[14] Rilo E, Shepherd J. A corpusbased approach for building semantic lexicons[C]. //In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-97), 1997:117--124.

[15] Hatzivassiloglou V, Mckeown K. Predicting the semantic orientation of adjectives. ACL[J]. Proceedings of the Acl, 1997:174--181.

[16] Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. Acm Transactions on Information Systems, 2003, 21(4):315--346.

[17] 朱嫣岚, 闵锦, 周雅倩,等. 基于 HowNet 的词汇语义倾向计算[C]// 全国第八届计算语言学联合学术会议（JSCL-2005）论文集. 2005:14-20.

[18] 李钝, 曹付元, 曹元大,等. 基于短语模式的文本情感分类研究[J]. 计算机科学, 2008,35(4):132-134.

[19] Bai X, Chen F, Zhan S. A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec[C]// Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, 2014:358 - 363.

[20] 姚全珠, 宋志理, 彭程. 基于 LDA 模型的文本分类研究[J]. 计算机工程与应用, 2011,47(13):150-153.

[21] 周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济, 2015, (2):145-148.

[22] 马宝君, 张楠, 孙涛. 智慧城市背景下公众反馈大数据分析：概率主题建模的视角[J]. 电子政务, 2013, 12: 9-15.

Public Opinion Analysis Based on Probabilistic Topic Modeling and Deep Learning

LAO Xin[1], MA Baojun[1], ZHANG Nan[2], WAN Yan[1]

(1. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. School of Public Policy and Management, Tsinghua University, Beijing 100084, China)

Abstract: With the rapid development of Internet, people begin to express their views and opinions about social events through social media. In the Web2.0 era, with the concept of broad public participation, effective analysis of the above information and displaying it in a reasonable style is rather essential for monitoring public opinions. This paper proposes a novel method about public opinion analysis based on social platforms, which first utilizes probabilistic topic model (i.e., LDA) to extract public's different perspectives on certain event, and then uses word2vec model to calculate the emotional intensity for each text. Then, time-series analysis is carried on the emotional intensity of the overall social event as well as certain topics selected manually, to track public emotional changes during whole incident. We finally show the rationality and effectiveness of the proposed method by a case study.

Key words: Public Opinion Analysis; Probabilistic Topic Modeling; Deep Learning; Emotional Intensity; Time-series Analysis