基于社交媒体数据的2023年土耳其地震舆情分析——主题演变、情感变化及公众反应研究

廖瑾瑜

2024年7月4日

摘要

本研究基于 2023 年 2 月土耳其地震期间的推特数据，运用文本挖掘和自然语言处理技术，深入分析了社交媒体上的舆情演变、主题变化及公众情感反应。研究采用高频词分析、潜在语义分析(LSA)、潜在狄利克雷分配(LDA)等方法进行主题建模，并结合时间序列情感分析探讨公众情感变化。

研究发现，公众关注呈现明显的阶段性演变，从初期的事件信息关注，到救援需求和行动，再到情感表达，最后扩展至国际影响。热点词分析表明公众更关注具体、及时的事件信息和救援动态。情感分析揭示了公众情感遵循“急剧上升-迅速下降-趋于稳定”的模式，反映了面对突发灾害时的心理适应过程。

本研究不仅深化了对公众在地震灾害中反应模式的理解，也为灾害管理、舆情监测和社交媒体应用等领域提供了有价值的实证依据和理论洞见。

关键词：主题演变 情感分析 社交媒体 自然灾害 舆情监测

目录

表格与插图清单

基于社交媒体数据的2023年土耳其地震舆情分析——主题演变、情感变化及公众反应研究

一、绪论

（一）研究背景

2023年2月，土耳其发生7.8级地震，这次地震不仅造成了严重的物质损失和人员伤亡，也引发了全球范围内的广泛关注。在当今数字化信息时代，社交媒体已成为公众获取信息和表达意见的重要平台。地震灾害发生后，社交媒体上的信息传播速度极快，内容涉及灾情报告、救援行动、公众情感表达等多个方面。通过分析社交媒体数据，可以实时反映出公众对灾害的反应和态度，以及救援工作的进展和效果。

研究自然灾害期间的社交媒体舆情，不仅能够帮助政府和相关机构了解公众的需求和情感变化，还能为灾后恢复和重建提供有价值的参考。同时，这类研究也有助于揭示社交媒体在灾害信息传播中的作用和特点，为提升应急管理和舆情引导能力提供理论依据。

（二）研究意义

本研究通过对2023年土耳其地震期间推特数据的分析，旨在揭示自然灾害背景下公众舆情的演变规律及其情感变化特征，并促进信息传播效率。作为信息传播的重要渠道，社交媒体的特点和作用在本研究中得到了深入探讨。研究结果将有助于理解在灾害情境下信息传播的模式和公众的互动行为，为提高信息传播的效率和准确性提供参考依据。

此外，本研究通过数据挖掘和分析技术，展示了如何利用大数据进行舆情监测和分析，为政府和组织在灾害应对中的数据驱动决策提供了示范。通过科学的数据分析，可以提高决策的准确性和针对性，增强整体应对灾害的能力。

（三）创新性

本次研究中，根据词频进行了主题选定，将困惑度由盲选383.9584的降到156.9179。

本次研究中，加入了情感时间序列的模型，并进行可视化。在统计建模大赛中一直想用但没有用出来的模型，此次论文中得到了使用。

二、文献综述

本文结合课上的数据预处理、特征工程及各模型相关课件内容，进行舆情分析。研究主要借鉴了课上“文本分析与主题模型”的模型和代码，使用了主题模型和情感分析，结合研究的具体问题，另外加入了一点时间序列情感分析的内容。

三、研究方法

（一）数据描述

本文选择了 2023 年 2 月 6 日至 2023 年 2 月 10 日期间关于土耳其地震的推特信息，旨在分析社交媒体在自然灾害背景下的舆情演变。尽管原计划使用 tas 上的数据集，但是由于 RStudio 环境中反复出现数据不存在的报错，最终选择通过网络获取相关数据。

具体而言，从 kaggle 上下载了五天内的推特数据并手动合并。合并后的数据集包括了 tas 数据集中的字段，两个数据应该区别不大。

（二）数据预处理

本文采用了一系列文本预处理技术，以确保数据质量并为后续分析奠定基础。预处理步骤主要包括文本清理、标准化、停用词移除、词干提取和文档-词项矩阵的构建。

首先，本文使用 quanteda 包中的 tokens()函数对原始文本进行分词处理。在此过程中，本文同时移除了标点符号、特殊符号、数字和 URL，以消除可能对分析造成干扰的非语义元素。随后，通过 tokens_tolower()函数将所有词项转换为小写，以统一文本格式。

为了进一步提高分析效率和准确性，本文采用 tokens_remove()函数移除了英语停用词。这些高频但低信息量的词语（如"the"、"is"、"and"等）往往不能为文本主题分析提供有意义的信息。接着应用了 tokens_wordstem()函数进行词干提取，将词语还原为其基本形式，以减少词形变化带来的复杂性。

为确保数据的纯净度，本文使用正则表达式通过 tokens_replace()函数移除了所有非字母字符，仅保留纯文本信息。这一步骤有助于降低数据噪声，提高后续分析的准确性。

最后，构建了文档-词项矩阵（Document-Term Matrix，DTM）。为了降低矩阵的稀疏性并聚焦于最具代表性的词语，仅保留了在语料库中出现频率不少于 10 次的词

项。这一做法不仅有助于减少计算复杂度，还能够提高模型的稳定性和解释力。

通过这一系列预处理步骤，本文成功地将原始文本数据转化为结构化的数值表示，为后续的主题建模和文本分析奠定了坚实的基础。这种预处理方法不仅提高了数据质量，还有效降低了模型的困惑度，为本文挖掘文本中潜在的语义结构和主题提供了可靠的数据支持。

（三）词频分析

为了深入理解文本数据的核心内容和主要话题，本文首先进行了词频分析。使用 quanteda 包中的 textstat_frequency()函数，本文计算了语料库中每个词项的出现频率。通过提取并展示前 20 个高频词，本文得以快速概览数据集中最常出现的关键词，从而初步把握文本的主要主题和焦点。这种方法不仅为后续的主题建模提供了基础，还帮助本文构建了一个更为精确的词典，显著降低了模型的困惑度（之前自己设定的词典模型困惑度为 383.9584，而经过高频词修正后的词典模型困惑度为 156.9179，显著降低）。

（四）主题模型

为了揭示文本数据中潜在的主题结构，本文采用了两种主题建模技术：潜在语义分析（Latent Semantic Analysis， LSA）和潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）。

1.潜在语义分析（LSA）

本文首先将文档-词项矩阵（DTM）转换为 TF-IDF（词频-逆文档频率）加权矩阵，以平衡词频与其在整个语料库中的重要性。随后，本文使用 textmodel_lsa()函数执行 LSA，提取潜在主题。通过分析词汇-主题矩阵（lsa.mod$features）和文档-主题矩阵（转置后的 lsa.mod$docs），本文能够洞察每个词汇在不同主题中的重要性，以及每个主题在各文档中的分布情况。

2.潜在狄利克雷分配（LDA）

本文使用 topicmodels 包中的 LDA()函数实现 LDA 模型，设定主题数 K=4，采用 Gibbs 抽样方法。模型的困惑度（perplexity）被用作评估指标，以衡量模型对未见文本的泛化能力。本文提取了每个主题的前 20 个特征词，并计算了每个词在主题中的后验概率（beta）。为了直观展示结果，本文为每个主题生成了词云图，使用 wordcloud 包可视化主题中最具代表性的词汇及其重要性。

值得注意的是，本文选定的词典显著降低了困惑度，在词典选择上，根据高频词，选择了四个主题，并且将主题中的词汇选定为较高频的词汇。主题最终确定为：

灾害救援（disaster_rescue）：包括词汇“earthquake”、“help”、“rescue”、

“victim”、“relief”、“survivor”、“support”、“injur”。这些词汇直接与地震事件及其后的救援行动相关，能够帮助识别讨论中关于救援和受灾情况的内容。

社会（society）：包括词汇“people”、“community”、“public”、“resident”、“citizen”。这些词汇涉及公众和社区的反应，能够反映社会层面的讨论和情感变化。

媒体（media）：包括词汇“news”、“report”、“media”、“broadcast”、“journalist”。这些词汇与新闻报道和媒体传播相关，能够帮助分析媒体在地震事件中的作用和影响。

组织（organization）：包括词汇“respond”、“reaction”、“organize”、“crisis”、“effort”、“coordinate”。这些词汇与组织和协调工作相关，能够帮助了解各类组织在应对地震事件中的表现和行动。

（五）情感分析

为了探索文本中蕴含的情感倾向及其随时间的变化，本文采用了多层次的情感分析方法。

1.基于词频的情感倾向分析

本文首先基于点赞数（likeCount）将评论分为正面（≥10 个赞）和负面两类。使用 quanteda 包的 textstat_keyness()函数，本文比较了这两类评论中关键词的分布差异，从而识别出与正面和负面情感相关的特征词。这种方法能够快速揭示不同情感倾向文本的词汇使用特点。

尽管这种方法可以识别出正面和负面评论中的特征词汇，但在结果中，相比于纯粹的情感倾向，这种分析可能更能反映出热点词汇的分布。因此在研究结果部分，更加偏向将其判断为关注度较高的词汇。

2.时间序列情感分析

使用 syuzhet 包，本文对每条推文进行了情感得分计算。首先，本文使用 get_sentiment()函数计算了整体情感得分，并绘制了日均情感得分的时间序列图，以观察情感随时间的总体变化趋势。

随后，本文采用 NRC 词典（get_nrc_sentiment()函数）进行了更细致的情感分类，包括愤怒、期待、厌恶、恐惧、喜悦、悲伤、惊讶和信任等八种基本情感，以及积极和消极两种情感倾向。本文特别关注了愤怒和恐惧情绪，绘制了这两种情绪随时间变化的趋势图。通过使用 ggplot2 包中的 geom_smooth()函数（采用 LOESS 平滑方法），本文能够更清晰地观察情感变化的整体趋势，同时保留了原始数据的波动信息。

这种多层次的情感分析方法能够全面把握文本中的情感动态，不仅反映了整体情感氛围，还能捕捉特定情绪（如愤怒和恐惧）的变化模式，为理解公众对事件的情感反应提供了丰富的洞察。

四、研究结论及分析

（一）高频词分析

高频词汇如“earthquake”、“help”、“turkey”和“people”表明，公众的讨论主要集中在地震事件本身、救援需求、地理位置和人群等方面。这些词汇的高频出现反映了事件的严重性和公众的高度关注。

“earthquake”：作为高频词汇之一，出现了 3739 次，直接指向了此次地震事件，表明这是讨论的核心话题。

“help”：出现了 3506 次，显示了公众对救援和支持的强烈需求，反映了社会的同情和援助意愿。

“turkey”：出现 3487 次，强调了事件的地理位置，指向了土耳其这一受灾国家。

“people”：出现 2528 次，表明讨论中频繁提到人群，涉及受灾人员及其反应。

（二）主题演变分析

从结果上来看，虽然 LDA 模型的困惑度显著降低了，但是主题似乎与开始设定的不完全一致，可能是因为 LDA 模型的无监督性质。不过困惑度的显著降低也说明词典的使用可能通过提供更有意义的特征选择来改善模型的整体性能，因此选用了更加合适的词典依旧是值得骄傲的一部分。

本文通过潜在语义分析（LSA）和潜在狄利克雷分布（LDA）模型提取了文本中的潜在主题。每个主题的特征词和词云图展示了各主题的核心词汇及其权重分布。以下是对四个主要主题的详细分析：

1.主题一：地震及其影响

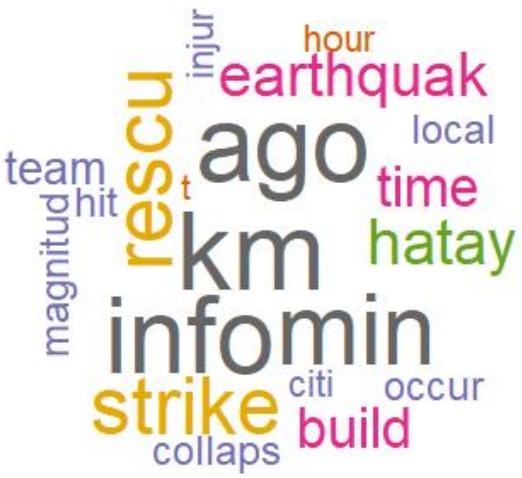

图 1 主题一的词云图

主题一的词云图（见图 1）显示，核心词汇包括“earthquake”（地震）、“km”（千米）、“ago”（之前）、“magnitude”（震级）等。这些词汇集中反映了地震的时间、地点和强度等信息。在地震发生的初期阶段，社交媒体上的讨论重点在于传递事件的基本信息。公众主要关注地震发生的具体细节，如震源深度、震中位置和震级等。这表明，地震事件的基本事实是初期讨论的核心内容。

2.主题二：救援行动与呼救

图 2 主题二的词云图

主题二的词云图（见图 2）显示，核心词汇包括“help”（帮助）、“people”（人们）、“need”（需要）、“support”（支持）等。这些词汇反映了地震发生后，公众对救援行动和救助需求的高度关注。社交媒体成为呼吁救援和提供帮助的重要平台，充斥着寻求援助和组织救援的信息。这表明，在灾后初期，救援行动是讨论的主要焦点，公众展现了强烈的社会同情心和援助意愿。

3. 主题三：感谢与祈祷

图 3 主题三的词云图

主题三的词云图（见图 3）展示了如“thank”（感谢）、“amin”（阿门）、“grate”

（感激）等核心词汇。这些词汇反映了在救援行动取得一定进展后，公众对救援人员和上天的感激之情以及祈祷祝福。社交媒体上的讨论转向表达感谢和祈求祝福的内容，显示了灾后公众情绪的积极变化和社会凝聚力的增强。这一阶段的讨论集中在对救援人员的感谢和对受灾人员的祝福上。

4. 主题四：国际援助与影响

图 4 主题四的词云图

主题四的词云图（见图 4）中，核心词汇包括“amp”、“syria”（叙利亚）、“affect”（影响）、“aid”（援助）、“pray”（祈祷）等。这些词汇显示了地震不仅对土耳其造成了影响，也波及了周边国家，如叙利亚。讨论中包含了对国际援助的需求和对受灾地区的祈祷。这表明，地震的广泛影响引起了国际社会的关注，讨论的焦点扩展到了国际援助和跨国合作。

通过对四个主题的词云图和特征词汇的分析，可以看出在 2023 年土耳其地震期间，社交媒体上的讨论主题随着事件的发展呈现出明显的演变趋势：

初期阶段：讨论集中在地震事件的具体细节，包括时间、地点和强度等。

救援阶段：公众高度关注救援行动和救助需求，社交媒体成为呼吁援助和组织救援的重要平台。

情感表达阶段：随着救援行动的进展，讨论逐渐转向对救援人员的感谢和对受灾人员的祈祷祝福。

国际关注阶段：地震的影响扩展到国际社会，讨论中出现了对国际援助的需求和对其他受灾国家的关注。

这些主题的演变反映了地震事件的不同阶段和公众情感的动态变化。通过深入分析社交媒体数据中的潜在主题，可以更全面地理解公众在地震事件中的反应和情感变化，为未来的灾害应急管理和社会舆情监测提供有价值的参考。

（三）热点词分析

likeCount（点赞数）大于等于 10，则该行被标记为 'positive'（正面）；否则，标

记为 'negative'（负面）。在本研究中，结果中的“正面关键词”更加倾向于热点词汇，“负面关键词”更加倾向于不受人们关注或人们不认同的词汇。本文利用点赞数作为区分标准，将评论分为“热点”（点赞数≥10）和“冷门”（点赞数<10）两类。

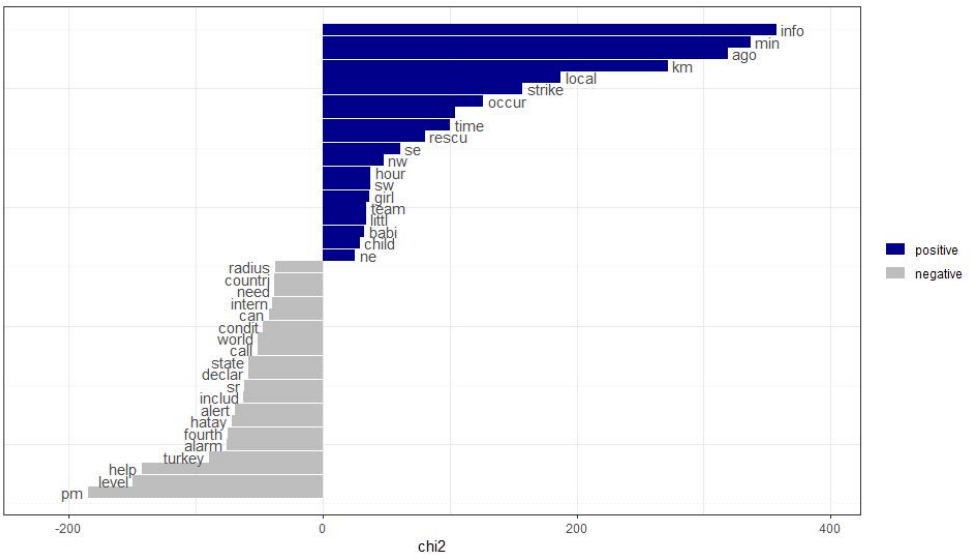

图 5 热点词汇分布图

在“热点”评论中，高频词汇主要集中在提供具体信息和时间点的词语上，如“info”（信息）、“min”（分钟）、“ago”（之前）、“local”（本地）、“km”（公里）等。这些词汇的高频出现，表明观众对具体事件的详细信息有较高的关注度。尤其是“info”和“min”这类词汇，表明在地震事件中，观众非常关注事件发生的具体时间和相关信息更新。

此外，像“strike”（袭击）、“occur”（发生）、“rescu”（救援）等词汇的出现频率较高，显示出观众对地震灾害过程及其救援行动的高度关注。这些关键词强调了观众对事件动态和后续救援情况的即时关注。

相对而言，“冷门”评论中出现的高频词汇包括“radius”（半径）、“country”（国家）、“intern”（国际）、“condit”（条件）等。这些词汇更多地涉及宏观背景和条件性描述，反映出观众对地震事件的宏观背景和影响较少关注。这可能与这些评论提供的信息较为笼统、缺乏具体细节有关。

（四）情感分析

1. 总的情感分数时间序列

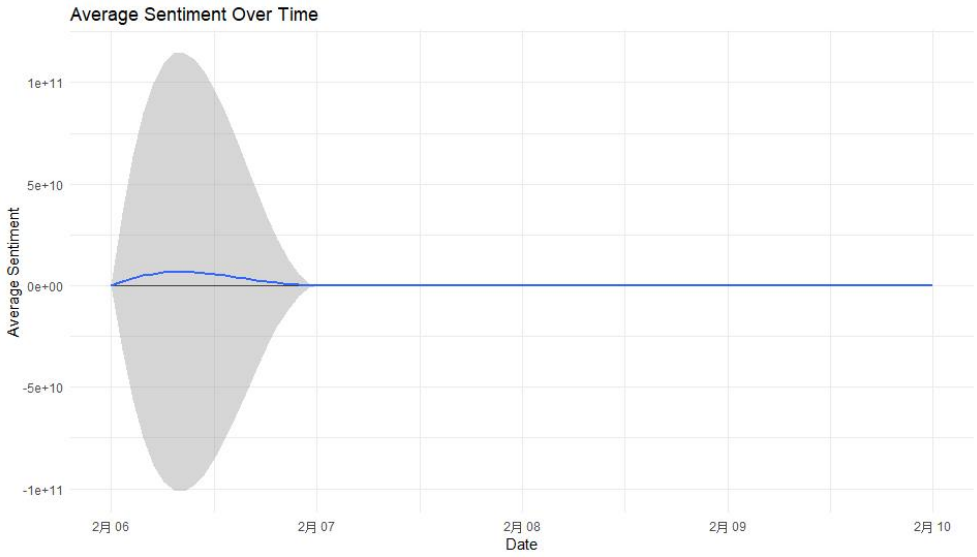

图 6 总的情感分数时间序列图

在图 6 中，我们观察到总的情感分数在时间序列上的变化几乎保持平稳，波动范围较小。蓝色线代表每日平均情感分数，而灰色阴影区域则表示置信区间。公众的情感分数在事件发生后的第一天内迅速达到顶峰，并在随后的几天内迅速消退。这种波动性较大的情感变化可能反映了地震事件初期公众情感的剧烈波动。

2. 愤怒的情感分数时间序列

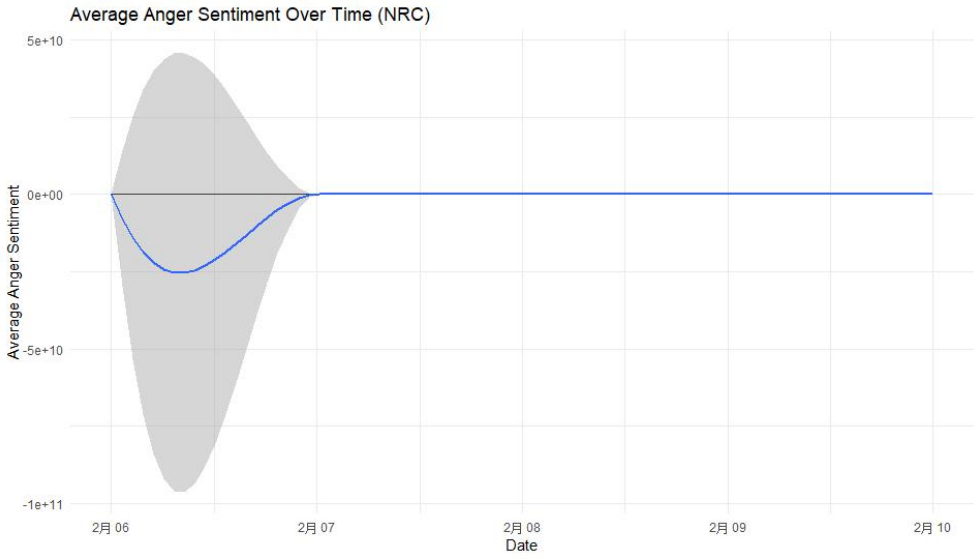

图 7 愤怒的情感分数时间序列图

图 7 展示了每日平均愤怒情感分数的时间序列变化。从图中可以看出，蓝色线在前几天有一个显著的下降趋势，之后趋于平稳。有关负面情感词汇在地震初期达到顶峰，随后迅速下降并保持在较低水平。这种情感波动性强且差异性大的现象反映了公众在事件发生初期的强烈反应，愤怒情感在短时间内迅速消退并趋于稳定。这种变化表明公众对地震事件的初期反应激烈，但愤怒情感在事件发生后的第一天内就消失了，

并在随后的时间内保持平稳状态。

3. 恐惧的情感分数时间序列

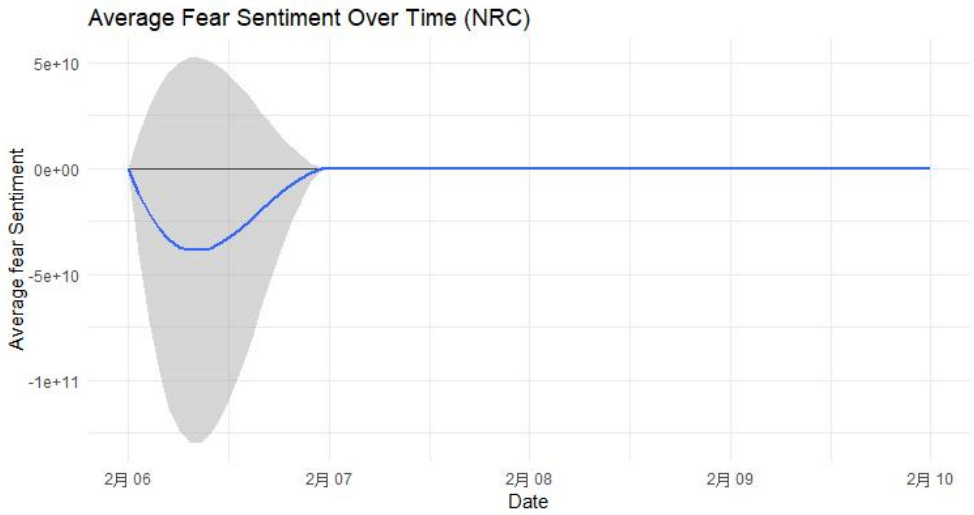

图 8 恐惧的情感分数时间序列图

图 8 显示了每日平均恐惧情感分数的时间序列变化。蓝色线在前几天呈现出明显的下降趋势，随后趋于平稳。有关恐惧的负面情感词汇在地震初期达到顶峰，之后迅速下降并趋于稳定。这与愤怒情感的变化趋势相似，显示出公众在事件初期的恐惧情感在短时间内迅速减弱，并在之后保持平稳。恐惧情感的波动性较大，差异性明显，但这种情感在事件发生后的第一天内就消失了，表明公众对地震的初期反应虽然强烈，但这种负面情感的持续时间较短，迅速趋于稳定。

这部分分析为我们提供了关于地震事件对公众情感影响的洞见，有助于进一步理解公众在面对自然灾害时的情感反应及其变化规律。

（五）主要结论总结

本研究通过对 2023 年土耳其地震期间社交媒体数据的多维度分析，揭示了公众反应的动态特征及其潜在机制。研究结果表明，公众关注重点呈现明显的阶段性演变: 从初期的事件信息关注，到救援需求和行动，再到情感表达，最后扩展至国际影响。这一演变过程反映了灾害事件中公众认知和情感的动态调整过程。

在信息传播方面，研究发现公众更倾向于关注和传播具体、及时的事件信息和救援动态，而非宏观背景描述。这一发现为优化灾害信息传播策略提供了重要启示。情感分析结果揭示了公众情感反应遵循"急剧上升-迅速下降-趋于稳定"的模式，这一模式在整体情感、愤怒情感和恐惧情感中均有体现，反映了公众面对突发灾害时的心理适应过程。

研究同时凸显了社交媒体在灾害应对中的多重作用，包括信息传播平台、救援协调工具、情感支持渠道和国际关注窗口。这些发现强调了社交媒体在现代灾害管理中的重要性，为优化其在灾害应对中的应用提供了实证依据。

本研究采用的多维度分析方法，包括高频词分析、主题模型、热点词分析和情感分析，展现了其在处理大规模社交媒体数据和揭示复杂社会现象方面的有效性。通过结合定量和定性分析，本研究不仅深化了对公众在地震灾害中反应模式的理解，也为灾害管理、舆情监测和社交媒体应用等领域提供了有价值的实证依据和理论洞见。未来研究可进一步探索不同类型灾害事件中的公众反应差异，以及社交媒体数据分析在预警系统和应急决策支持中的应用潜力。

致谢

感谢黄一凡教授提供的相关参考资料。

2024 年 7 月 4 日