中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

题目：性别与经济舆情在时间序列上的主题演化研究
——“胖猫”事件在百度贴吧

学年学期：2023-2024-2

课程名称：舆论学

课程编号：2011030016

课程序号：01

任课教师：韩运荣

姓　　名：黄淼森

学　　号：2021201013045

评分区域（由阅卷老师填写）：

结课成绩：

总评成绩：

提交时间：2024 年 6 月 30 日

目录

性别与经济舆情在时间序列上的主题演化研究
——“胖猫”事件在百度贴吧

摘　要　【目的】了解在男性用户为主的社交平台下，用户对性别与经济议题社会舆情事件的主题和态度，为社交媒体影响下的网络舆情引导和干预提供实证检验和数据参考。【方法】本研究以“胖猫”事件在百度贴吧的舆论表现为对象，采用机器学习方法对百度贴吧文本数据进行向量化，并使用深度学习方法对进行聚类，分析平台中主要话题的时间序列网络，从而进一步揭示舆情形成和演化的规律。【发现】百度贴吧中对于“胖猫”事件展开讨论时，涉及主题在时间序列上没有显著变化，经济问题讨论度最高，性别问题次之，事件本身讨论度最小。在时间序列上，三个主题的讨论在事件初期和通报发出后达到高峰，随后逐渐回落并保持在较低水平，但在某些时间点上仍然有明显的波动。其中性别问题波峰的出现早于其他主题。【局限】分析数据体量较小，不能完全反映真实情况，缺乏对不同社交平台数据的实验验证和分析比较。【结论】“胖猫”事件折射出社交媒体影响下的权力结构、资本流动、性别关系、媒介生态以及公共治理等多重复杂的社会现象。百度贴吧舆论虽因男性用户占主导而面临性别偏见放大的挑战，但也孕育了探讨社会、政治、经济议题的土壤。其舆论特征反映出人们对于社会不公的敏感度以及寻求改变现状的意愿。

关键词　主题演化　网络舆情　时间序列分析　百度贴吧

一、 研究背景

（一）社会背景

近年来，随着互联网的普及和社交媒体的迅猛发展，网络空间成为了公众表达意见、分享信息以及进行社会互动的重要平台。在这一背景下，网络舆情逐渐成为反映社会问题、引导公众情绪的重要媒介。然而，在于全球化进程中，中国互联网空间中社会观念的多元化和区隔化现象日益显著，特别是在文化主体性相关的观念之争中，互联网新媒体的影响尤为突出(陈云松, 2022)[1]。在舆情事件发生过程，社交媒体中社会观念的多元化表达成为深刻理解公众的核心关注与诉求，成为重要的研究内容。

涉及情感纠纷、经济纠纷等个人事件常常因其复杂性和争议性而引发广泛关注，而性别议题的讨论热度在涉及情感和经济纠纷的事件中更为显著。因为这些事件不仅反映了个体之间的矛盾，也折射出社会层面的深层次问题，如信任危机、情感操控、经济压力等。在舆情事件中，公众对社会中性别不平等和性别歧视的问题较为敏感，一旦事件激发了公众的情感共鸣，就容易导致网络上的激烈争论，甚至出现性别对立情绪的升温。

此外，在当代社会，经济压力和财务纠纷常常成为个人和家庭矛盾的根源。公众对相关舆情事件的高度关注反映了社会中普遍存在的对财务安全和经济正义的关注。“胖猫”事件正是在这种社会背景下发生的，其引发的广泛关注和激烈讨论说明了公众对男女情感、经济诈骗、网络暴力等问题的高度敏感。

（二）舆情演化

“胖猫”事件始于一名游戏昵称为“胖猫”的湖南男子，在与重庆女子谭竹的网恋关系中遭遇情感和经济纠纷后，于 2024 年 4 月 11 日在重庆跳江自杀，遗体于 4 月 23 日被打捞确认离世。事件中，男生疑似遭受 PUA（情感操控），并在生命的最后时刻给女友转账 6.6 万元，标注为“自愿赠予”。舆论爆发于“胖猫”姐姐指控谭竹利用感情骗取胖猫 50 多万元，而谭竹则否认骗

[1] 陈云松. 观念的“割席”——当代中国互联网空间的群内区隔[J]. 社会学研究, 2022, 37(4): 117-135+228.

钱，称分手是因双方频繁争吵及胖猫的冷暴力，且已协议退还部分款项。

“胖猫”事件在网络空间引发了极大的关注，自 5 月 1 日起至 6 月 1 日，事件持续发酵，网络相关信息总量高达 3621414 条[1]，热度显著超越同期其他重大社会事件，大量民众以送外卖和鲜花的形式前往大桥进行悼念，显示了公众对此事件的高度关注。事件之所以能迅速聚集如此高的热度，是因为它触及了公众对于情感诈骗、性别对立、网络暴力等敏感社会议题的深切关切，同时，事件本身的悲剧性和复杂性也激发了公众的情感共鸣和探究真相的欲望。事件在网络的发酵，形成了几个关键阶段：

1. 潜伏期——事件发生之后

2024 年 4 月 11 日凌晨“胖猫”跳江，4 月 23 日其遗体被打捞上岸。虽然“胖猫”的游戏代练身份，使得该事件获得相关人群的关注和讨论。但是由于细节缺失，该事件的冲突性和反常性尚未展现，网络上对于该事件的报道和互动较少。

2. 爆发期——双方对峙冲突

2024 年 5 月 2 日，胖猫姐姐通过网络平台曝光事件，揭示了情感和经济纠纷，引发初步关注。随后，“胖猫”女友发声，对相关质疑进行反驳，但很快删除，相关图片被网友保存了下来，舆论场上的紧张情绪开始升温。在双方的回应中，女方否认金钱动机并指责男方家庭引导网络暴力，而男方姐姐则坚称会为弟弟讨回公道，双方向公众的直接回应导致舆论两极分化严重。虽然“胖猫”事件本身是个体事件，但其背后映射的情感剥削和经济纠纷问题，在现代社会中并非孤立存在。“胖猫”事件中疑似存在的情感操控和金钱争议引发了部分网友的强烈不满。

3. 蔓延期——次生舆情涌现

随着事件的扩散，公众的同情心被激发。2024 年 5 月 3 日，大量网友通过外卖平台送餐、送花悼念，但这也导致了相关品牌因外卖“空包”问题陷入舆论风波。同时，网络上出现了模仿账号和不当玩梗现象，进一步复杂化了舆情环境。此外，性别对立情绪被点燃，出现了大量嘲讽逝者的行为和诋毁其女友的言论。2024 年 5 月 4 日，部分账号因煽动性别对立被处理。2024 年 5 月 9 日，据泸州网警报道，有网友在短视频平台冒充“胖猫”女友，发布虚假的资

[1] 基于舆情研究平台“舆情通”监测，2024 年 5 月 1 日至 2024 年 6 月 1 日期间数据，详见附录 3。

金往来信息，扰乱公共秩序，被采取行政处罚。

4. 缓解期——舆情持续发酵

2024 年 5 月 4 日至 2024 年 5 月 17 日，尽管事件长时间占据社交媒体热搜，但真相仍未明了，部分网民开始呼吁理性讨论，对持续的“小作文”和聊天记录表示疲惫。一方面，网络出现了不少事件梳理与推测分析，但由于官方提供的线索不足、双方各执其词，这些讨论难免偏颇。另一方面，也有声音提醒关注其他重要社会事件，防止舆情过度泛化。

5. 第二爆发期——官方公告与舆论反转

2024 年 5 月 19 日晚间，重庆市公安局南岸区分局发布警情通报，而微博等平台也加强了对违规内容的清理。该阶段，对真相的判断与等待成为了核心议题，虽然警方通报专业详实，还原了真相、解答了疑问，但是其过分的滞后也引发了部分网友的不满。此外，法律追偿、网络悼念的商业道德、以及网络暴力等问题也成为广泛讨论的焦点。

根据舆情研究平台“舆情通”检索，2024 年 5 月 1 日至 2024 年 6 月 1 日期间，舆情整体呈现双峰型（如图 1），对应 5 月 4 日事件在双方回应、网友悼念的推动下走向高潮，以及 5 月 19 日警方通告引发的“舆论反转”。

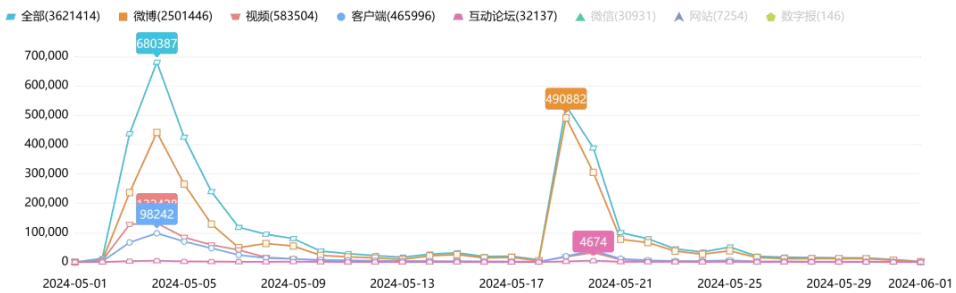


图 1　2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”信息走势图

由图 2 可知，该时间段内敏感信息共有 3369671 条（占 93.05%），非敏感信息共有 229790 条（占 6.34%），中性信息共 21953 条（占 0.61%）。经分析，敏感信息在 2024-05-04 达到舆论最高峰值，信息量为 618091 条，值得重点关注。

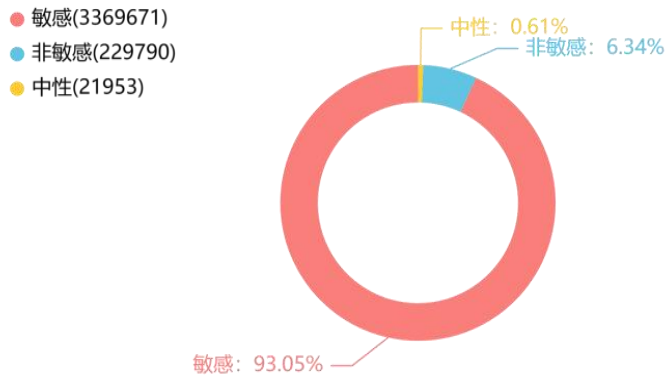


图 2　2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”敏感占比图

从图 3 可看出，该时间段内微博信息中“中性”情绪的发文最多，共 1179830 篇（占 47.17%）。但值得注意的是，愤怒、悲伤为主的负面情绪相比积极情绪显著更高，证明该舆情事件给社会带来巨大的负面情绪影响。

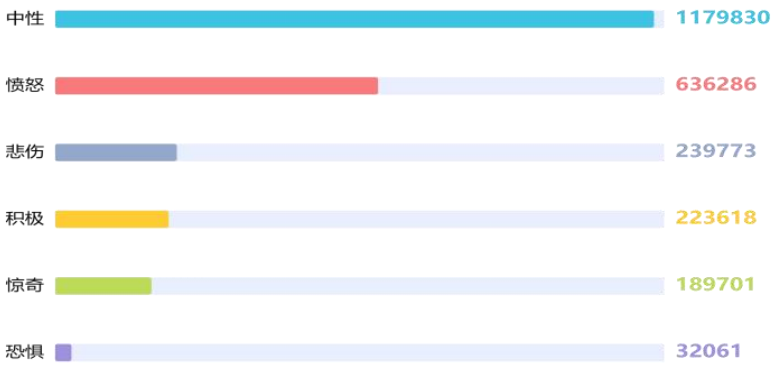


图 3　2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”微博情绪地图

（三）议题分析

整个舆情发展过程中，事件经历了从个体情感悲剧到社会广泛关注的转变，伴随着复杂的情感表达、激烈的网络骂战、性别对立的升级，以及次生舆情的不断涌现，形成了一个高度复杂且充满挑战的舆论场。

根据舆情研究平台“舆情通”检索，2024 年 5 月 1 日至 2024 年 6 月 1 日期间，微博热门话题 Top100 中有 30 条属“胖猫”事件相关，总热度达到 6411694。根据整理和人工识别，该时期对该事件的讨论议题主要分布在以下方面：双方本人情况、双方经济往来、双方法律纠纷、冒充散布谣言、煽动性别对立、操纵引导舆论、游戏代练情况、外卖悼念情况。

二、 相关研究

（一）研究对象

百度贴吧成立于 2003 年，是百度公司推出的一款基于关键词搜索的互动平台。随着互联网的发展，百度贴吧曾一度成为中国最大的网络社区之一，用户可以围绕共同兴趣爱好和话题进行讨论。然而，随着新兴社交媒体平台的崛起，如微信、微博、抖音等，百度贴吧的用户活跃度和影响力逐渐下降，逐步退出主流社交平台的行列。

尽管如此，百度贴吧在热点话题上仍然展现出其独特的舆论生态。在平台类型上，百度贴吧作为一个高度社区化的社交媒体，其用户在平台内部也存在多元化与区隔化的特点。用户在各自的社区中组成社群共同体，表现出了强烈的社区参与动机。这也可能会导致社区用户的表达更加情绪化而非理性,从而在社区内部形成一种情绪的传染和共振（胡杨涓等,2019）[1]。理解社群共同体这一概念在网络社会的变化及发展,对于新时代合理引导和规范青少年社会价值观具有一定的启示意义（赵艳娇，2019）[2]。在用户结构上，百度贴吧以男性用户居多，且集中于经济发达地区及一线、超一线、二线城市。用户对游戏、数码、院校等话题表现出较高的关注度。[3]在传播表现上，在百度贴吧的传播模式中，每个网民都可以自由创建贴吧或发布信息，传统“把关人”的作用被削弱。这种模式是基于用户关键词搜索的主题讨论社区，具有高度的自主性和互动性。这些特点使得百度贴吧在某些特定的舆论事件中，能够展现出不同于其他社交平台的讨论特点和舆论走势。

在“胖猫”事件中，百度贴吧的舆情表现相较于其他社交平台和网络总体舆情，展现出“双主峰+多侧峰”的特点。舆情研究平台“舆情通”数据显示（如图 4），2024 年 5 月 1 日至 2024 年 6 月 1 日期间，百度贴吧监测到总信息 18518 条，与网络总体舆情的双峰型不同，百度贴吧的舆情呈现出多峰特点，

---

[1] 胡杨涓, 胡千红. 虚拟社区中的用户特征与情绪表达——对“知乎”社区五类新闻议题讨论的实证分析[J]. 青年记者, 2019(33): 22-24.

[2] 赵艳娇. 网络空间的社群共同体——基于百度贴吧粉丝群的考察[J]. 北方民族大学学报(哲学社会科学版), 2019(5): 82-87.

[3] 走心发布丨贴吧垂类生态数据报告_兴趣[EB/OL]. [2024-06-28]. https://www.sohu.com/a/www.sohu.com/a/447474731_165158.

这反映了用户在讨论主峰之外还存在多次衍生讨论。

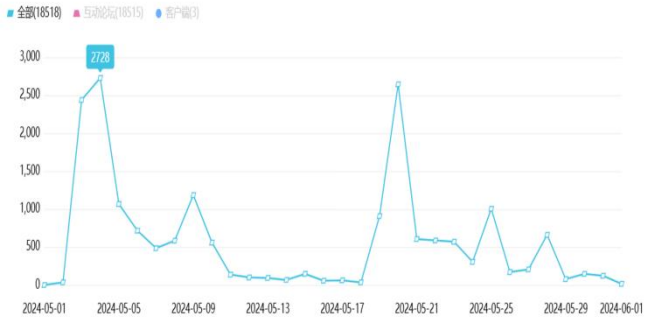


图 4　2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”信息走势图-百度贴吧

此外，百度贴吧在该阶段的互动声量走势独特。由图 4 和图 5 可知，在舆论总量呈现双主峰、第一主峰总量更高的同时，第一主峰的互动声量（回帖数）极低。这表现出虽然当时有较多人希望将这一社会热点事件引入百度贴吧展开讨论，但是平台的一般用户对此并不感兴趣。而当新闻通报发布之后，新旧的贴子同时获得了激烈的互动讨论。

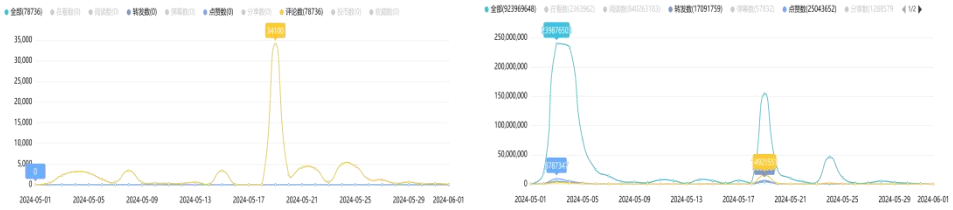


图 5　2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”互动声量图

（左：百度贴吧，右：全部）

（二）网络舆情研究

随着互联网和信息科技的发展，人们已经习惯于从社交媒体获取信息。然而，社交媒体中信息的复杂性和多样性使得我们更难找到想要的信息。使用话题发现方法分析新闻事件中的话题，不仅能帮助人们更好地理解事件的发生和演化，还能分析公众关注的问题，理解舆论的关注点。因此，话题发现的分析方法吸引了众多学者的关注，并在网络舆情的相关研究中得到了广泛应用。[1]

[1] Yan X, Guo J, Lan Y, et al. A probabilistic model for bursty topic discovery in microb logs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2015, 29(1).

目前，关于性别议题网络舆论的话题研究主要集中在舆论话题发现和传播演变。其中，网络舆论的话题发现主要是针对社会网络中意外事件话题的识别，或者基于原始模型改进和优化预测模型的准确性。研究中通常采用聚类分析和话题模型。前者主要基于同类文档相似度大或不同类文档相似度小的聚类假设，然后将文本信息转化为数字信息，并通过机器学习方法进行处理。例如，一些研究利用社交媒体中的大规模文本数据，通过细分构建基于推文的事件监测系统，并考虑信息内容的频率分布和相似性，将紧急事件推文片段检测为事件片段，然后将事件片段聚类为事件，以实现相关事件的识别。[1]也有研究利用机器学习聚类算法来识别社交媒体中的话题趋势，并提供语义分析，以综合准确描述每个话题[2]。后者是基于文本挖掘中广泛使用的话题发现工具，从大规模文本数据中自动寻找潜在隐藏的话题并进行建模。最广泛使用的模型是 LDA（潜在狄利克雷分布）模型及其改进模型[3]。

总结来说，网络舆情的话题研究相对成熟，已经取得了良好的成果，但很少有研究将特定平台，尤其是用户局限、特点鲜明的社区化社交媒体作为研究对象。此外，传播学领域的相关探讨存在面向单一、逻辑断裂、缺乏纵深考量等问题。[4]因此，我们以“胖猫”事件在百度贴吧的舆情表现为研究对象，舆情话题的传播演化规律，并比较进一步探讨不同类型紧急事件的舆情规律。

（三）数学方法

本文主要构建了三个模型，即文本数据的向量化、网络舆情的主题识别和网络舆情的时间序列模型。研究主要使用 Word2Vec 词向量模型将文本数据转换为词向量，然后使用 K-means 聚类方法实现主题识别。在时间序列模型中，研究主要使用 ARIMA 模型对舆情进行时间-主题分析，并进一步比较不同类型的主题，以探讨性别主题与其他衍生主题在时间网络上的演变共性和差异。本节将简要介绍 Word2Vec 词向量模型、K-means 主题分类模型和 ARIMA 时间序

---

[1] Li C, Sun A, Datta A. Twevent: segment-based event detection from tweets[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 155-164.

[2] Mathioudakis M, Koudas N. Twittermonitor: trend detection over the twitter stream[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. 2010: 1155-1158.

[3] Stieglitz S, Mirbabaie M, Ross B, et al. Social media analytics–Challenges in topic discovery, data collection, and data preparation[J]. International journal of information management, 2018, 39: 156-168.

[4] 董晨宇,林琦桁.“黏连剂”与“节拍器”：网络舆论议题的主题演化——对李佳琦“怒怼网友”事件的多维社会网络分析[J].传媒观察,2024(06):24-34.

列模型的原理和优势。

1. Word2Vec 词向量模型

社交媒体为人们了解与突发事件相关的信息提供了便捷的渠道。然而，网络数据的规模庞大，大量的无用信息和噪声的存在使得提取有效信息变得更加困难。为了及时准确发现重要信息，主题识别方法已经在社交媒体的事件检测和信息提取中得到了广泛应用[1]。然而，在以短文本信息为主的社交媒体平台如 Twitter 和 Facebook 上，传统的以 LDA 为代表的主题识别模型并不适用，主题识别的有效性无法得到良好保证[2]。作为一种典型的主题模型，LDA 主要使用软聚类方法来聚类文档，并通过研究文档矩阵，利用文档中词语的共现关系实现主题聚类。但是，当主题模型应用于短文本文档时，经常会发生数据稀疏的问题。具体来说，为了进一步提高主题识别的准确性，研究人员提出了一系列基于深度学习的主题识别和文本挖掘方法，其中最具代表性的模型是 Word2Vec，它能够生成词向量，并由一个浅层双层神经网络结构表示[3]。与以 LDA 为代表的主题模型相比，Word2Vec 主要表现为具有神经网络结构的词嵌入模型。通过学习上下文-词语矩阵，将词语转换为词向量，这在短文本信息的主题识别中表现出了更好的结果。

Word2Vec 是一个简化的神经网络模型，通过处理大量文本数据，学习词与上下文之间的关系，并将词的语义映射到向量空间。Word2Vec 模型包括两种主要训练方法：CBOW 模型和 Skip-Gram 模型。区别在于，前者主要通过当前词的语义实现上下文词的预测，而后者主要通过上下文的语义来预测当前词。尽管在遍历所有文本后，Skip-Gram 模型在获取所有文本的词向量方面更为准确。考虑到在分析较小规模社交媒体数据集时的确保模型精度和预测准确度，本研究采用了 Skip-Gram 模型。

Skip-Gram 模型具体实现如下：

输入层：一个 one-hot 编码的向量，维度为词汇表大小。

隐藏层：一个包含若干神经元的全连接层，神经元数量通常为词向量的维

---

[1] An L, Zhou W, Ou M, et al. Measuring and profiling the topical influence and sentiment contagion of public event stakeholders[J]. International Journal of Information Management, 2021, 58: 102327.

[2] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. 2013: 1445-1456.

[3] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013: 746-751.

度。通过输入词的 one-hot 编码向量乘以输入层到隐藏层的权重矩阵将输入向量转换为隐藏层向量。

输出层：一个包含词汇表大小神经元的全连接层，输出每个词的概率。通过 softmax 函数将输出向量转换为概率分布，表示每个上下文词的概率。

2. K-means 聚类模型

K-means 是一种无监督学习算法，用于将数据集中的样本划分为多个簇（clusters）。其基本原理是通过迭代优化过程，将数据点分配到最近的簇中心，并更新簇中心以最小化簇内数据点的平方误差和。

K-means 的优势在于其简单性和高效性，适用于大规模数据集。然而，它对初始簇中心的选择敏感，可能陷入局部最优解。本研究会结合文献研究中的议题分析和预处理过程的 tf-idf 频数分析对 Word2Vec 词向量模型训练结果和 K-means 聚类结果进行评估，以减小误差。

K-means 的主要步骤包括：

1. 初始化：随机选择 K 个数据点作为初始簇中心。
2. 分配：将每个数据点分配到距离最近的簇中心。
3. 更新：计算每个簇的新中心，即簇内所有数据点的平均值。
4. 重复步骤 2 和 3，直到簇中心不再显著变化或达到预设的迭代次数。

3. ARIMA 时间序列模型

时间序列是由相同对象的观测值按照时间顺序排列形成的一系列序列。其目的是利用现有历史数据来预测未来的数据。为了实现观测值的预测，已经建立了基于时间序列数据的随机和动态模型。常见模型包括自回归（AR）、移动平均（MA）和向量自回归（VAR），以及基于 AR 和 MA 模型的自动回归移动平均（ARMA）和自回归积分移动平均（ARIMA）模型。

ARIMA（AutoRegressive Integrated Moving Average）模型是一种广泛用于时间序列分析和预测的统计模型。ARIMA 模型结合了自回归（AR）、差分（I）和移动平均（MA）三个部分，以捕捉时间序列中的趋势和周期性。ARIMA 模型的基本原理是通过数学方法逼近观测对象在时间上的数据变化，并基于历史数据的拟合实现对未来的预测。

ARIMA 模型的优势在于其能够处理非平稳时间序列，并通过差分和自回归项捕捉序列中的复杂动态。本研究主要使用 ARIMA 模型来实现对“胖猫”事

件衍生的舆论主题情绪演化的分析，然后探索百度贴吧中舆论主题的内部形成和传播规律。

ARIMA 模型的主要步骤包括：

1. 平稳性检验：检查时间序列是否平稳，若不平稳则进行差分处理，直到序列平稳。

2. 模型识别：通过自相关函数（ACF）和偏自相关函数（PACF）确定 AR 和 MA 的阶数。

3. 参数估计：使用最大似然估计等方法估计模型参数。

4. 模型检验：检查模型的残差是否为白噪声，以确保模型拟合良好。

5. 预测：使用拟合好的模型进行未来值的预测。

三、 研究设计

在研究中，首先选取“胖猫”事件作为主要研究对象，然后抓取了该事件不同时间下在百度贴吧的舆论数据集，并通过去除停用词、分词、过滤文本内容完成文本预处理。其次，利用 Word2Vec 和 K -means 聚类方法提取该事件下的讨论主题。最后，运用时间序列分析方法对舆论内容进行建模，探究“胖猫”事件舆情的主题演化机制，绘制演化图，并对不同主题进行对比，了解其演化规律的共性与差异。研究框架如图 6。

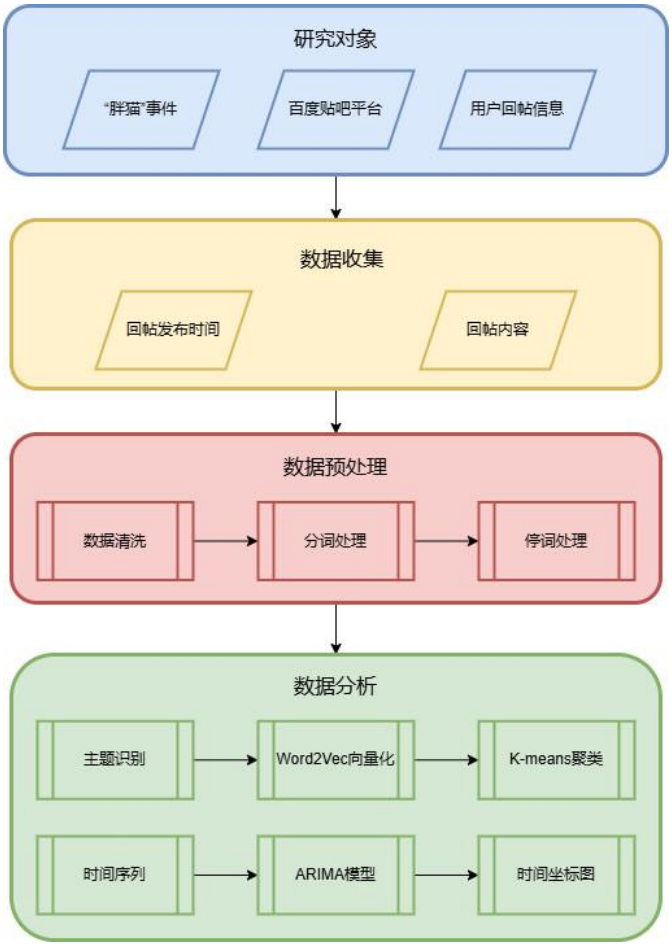


图 6　研究框架

（一）数据收集与预处理

研究选取“胖猫”事件作为案例材料，并以百度贴吧中与该事件有关帖子的回帖信息为数据来源，以对应研究背景中百度贴吧舆论在该事件中表现特殊的“互动声量”。

百度贴吧是曾是中国最具代表性和影响力的信息互动与交流平台之一。类

似于 Facebook 等由社群主导的社交媒体平台，用户在百度贴吧中发帖与回帖，对事件表达观点，并与他人互动，从而构建了一个信息交换与共享网络。通过限定时间和搜索关键词的方法，研究人员获得了“胖猫”事件相关的回帖和时间信息，为后续的研究提供了充足的数据支持。

在数据分析之前，为了进一步提高模型运行的效率和结果的准确性，需要采取一些数据预处理步骤。首先，对获取的数据进行初步浏览、清洗，并通过正则表达式等方法对文本内容进行了过滤。其次，需要使用分词工具对中文文本进行分词，以解决中文文本与英文文本相比缺乏空格的问题。最后，为了进一步处理文本数据中的无效词汇，需要依据停用词表对分词进行停用词删除、同义词替换和词性标记等操作。

（二）预实验

预实验采用 DTM（Dynamic Topic Model）模型基于 LDA（Latent Dirichlate Allocation）主题识别模型，引入时间因素，从而刻画舆论主题库主题随时间的动态演化。预实验使用 Python 用于机器学习的第三方库 Gensim[1]的模型包 ldaseqmodel 计算语料库主题分布。

经过反复尝试与调整，最终确定模型参数如下：（1）time_slice 语料库间隔：根据获取的数据时间分布情况，切割为 1225，325，255，225 四部分；（2）num_topics 主题数：3 个；（3）passes、inter 迭代次数：20 次 passes，50-100 次 em_inter。经过训练，获得主题的在不同时期的演变，结果如表 1。

表 1　主题在不同时期的演变

	1225（第一时期）		325（第二时期）		225（第三时期）		225（第四时期）	
	主题中心词	权重	主题中心词	权重	主题中心词	权重	主题中心词	权重
主题 1	女、男、舔、死等	0.171	女、男、舔、死等	0.172	女、男、舔、死等	0.173	女、男、死、舔等	0.173
主题 2	点、吃人血馒头、钱、现在等	0.101	点、吃人血馒头、钱、现在等	0.100	点、吃人血馒头、钱、现在等	0.101	点、吃人血馒头、钱、现在等	0.101
主题 3	胖猫、谭竹、好似、女等	0.116	胖猫、谭竹、好似、女等	0.118	胖猫、谭竹、好似、女等	0.119	胖猫、谭竹、好似、女等	0.119

由表 1 可知，三个主题的关键词在时间序列上基本保持不变，三个主题的权重在时间序列上基本保持不变。据此可以得出，三个主题在时间序列上不存

[1]　开源于：https://github.com/piskvorky/gensim.

在显著动态变化，即百度贴吧在该时间段内对“胖猫”事件的舆论主题基本保持一致，印证了研究背景中所提及的，百度贴吧舆论聚焦、延伸较少的特点。

预实验结果表示，主题关键词之间差异并不显著，即 LDA-DTM 模型组合训练得出的词向量表现、主题识别结果表现不良好，为最终选定 Word2Vec-Kmeans 模型组合进行主题识别在该舆论主题数据库的合理性提供了实验依据。预实验侧面验证了舆论主题在时间序列上的相对平稳性，为最终选定时间序列 ARIMA 模型提供了实验依据。

（三）主题识别模型的设计

在使用 Word2vec 模型将文本信息转化为词向量之前，我们首先通过 TF-IDF 模型提取特征词，以减少文本数据中无用信息和噪声的干扰。基于加权处理，我们从每条文本数据中提取出最重要的关键词，并使用它们作为词向量转换的数据集。其次，我们使用 Word2vec 中的 Skip-Gram 模型转换数据集的词向量，将文本转化为词向量，并使用常见的文本挖掘方法 K-means 算法进行文本聚类。根据文本之间的相似度，将提取的主题划分为不同的聚类，以实现文本主题的识别。最后，根据主题识别模型和 TF-IDF 词频表，使用聚类中心词组和聚类高频词组两类词组表示主题关键词。根据两组关键词，使用人工标签编码形成主题描述，以区分每个事件的文本数据。

（四）时间序列模型的设计

在获得舆论主题数据集后，统计百度贴吧发布时间与舆论数据集中的主题数量的关系，获得可以分析的时间序列数据。为了实现时间网络上的舆论主题分析，我们使用了 ARIMA 模型来分析舆论数据集。在获取舆论主题数据集的基础上，研究主要建立了一个以文本数量为纵坐标、时间尺度为横坐标的时间图，以观察时间尺度下舆论主题的演变。其中，获取基于主题的舆论时间序列数据的步骤如下。

首先，检查时间序列是否平稳，若不平稳则进行差分处理，直到序列平稳，再通过自相关函数（ACF）和偏自相关函数（PACF）确定 AR 和 MA 的阶数。其次，使用最大似然估计等方法估计模型参数。同时，通过贝叶斯信息准则（BIC）对模型可选参数进行穷举分析，对最大似然估计的结果参数进行验证，找到模型拟合最为良好的模型参数。最后，根据舆论主题数据集的发布时

间和文本数量建立二维坐标图，进行舆论主题的时间序列分析。

四、 研究发现

（一）数据描述性分析

在本文中，研究通过舆情研究平台“舆情通”获取相关数据。基于关键词“胖猫”和平台“百度贴吧”，获取了 2024 年 5 月 1 日至 2024 年 6 月 1 日之间百度贴吧上关于“胖猫”事件的帖子，这些帖子包括用户对事件本身的评论以及与其他人评论的互动（即回帖）。最终，研究选定回帖作为数据集，回收有效数据 2000 条。

首先，我们对获取的数据进行了初步浏览，删除了无效的帖子，通过正则表达式等方法对文本内容进行了过滤。其次，我们使用 Python 用于分词的第三方库 jieba[1]进行中文分词，使用的模型为默认的 HMM（隐马尔可夫）模型。最后，我们对开源的哈工大实验室停用词表 stopwords_hit.txt[2]进行微调，并据此对分词结果进行标注、合并同义词并删除了停用词。

同时，我们基于分词表，统计词语频率（如表 2），构建句-词-词频表，利用 Python 用于词云构建的第三方库 wordcloud[3]和用于绘图的第三方库 matplotlib[4]，绘制了词云图如表 7。

由频数表与词云图可初步推测，百度贴吧用户在该事件讨论中表现出以下特征：

1. 高度聚焦特定对象：“胖猫”一词以 149 的高频率成为绝对焦点，表明该事件或话题围绕“胖猫”这一核心元素展开，用户对此表现出极大的兴趣和讨论热情。

2. 性别议题显著：“女”、“男”分别以 134 和 116 的频数紧随其后，结合“龟男”、“舔狗”、“xxn”等词汇的出现，显示出讨论中存在明显的性别对立或性别角色相关的争议，反映了用户对于性别话题的敏感性和参与度。

3. 情感色彩浓厚：诸如“好死”、“吃人血馒头”、“骂”等词汇的高频出现，说明讨论中不乏激烈的情绪表达和道德评判，用户在交流中倾向于使用

---

[1] 开源于：https://github.com/fxsjy/jieba.
[2] 开源于：https://github.com/CharyHong/Stopwords.
[3] 开源于：https://github.com/amueller/word_cloud.
[4] 开源于：https://github.com/matplotlib/matplotlib.

强烈的情感词汇来表达不满、愤怒或批评。

4. 特定梗与网络文化：“吃人血馒头”、“龟男”、“舔狗”等具有特定含义的网络用语频繁出现，体现了百度贴吧用户群体内部独特的语言风格和网络文化认同，这些特定梗被用来快速传递复杂情绪或观点，增强了讨论的共鸣感。

5. 寻求真相与信息验证：“真的”、“知道”、“反转”等词的出现，表明用户在表达同情和愤怒的同时，也在寻找事件的真实面貌，有探索事件多面性、求证事实的倾向。

6. 群体归属与身份认同：“同情”、“好死”、“谭竹”、“姐姐”等词汇的使用，反映出用户在互动中进行道德判断和价值观的输出，既有对受害者的嘲讽或同情，也有对谭竹和“胖猫”姐姐的支持与否，展现了网络社区中道德观念的碰撞和讨论，以及用户的阵营划分与相互冲突。

表 2　词频图（部分）

词	频数
胖猫	149
女	134
男	116
好死	91
吃人血馒头	87
龟男	85
死	79
谭竹	75
真的	64
看	63
骂	62
钱	61
舔	53
舔狗	48
龟	43
发	43
知道	38
姐姐	37
同情	35
xxn	35
反转	34

… …

图 7　词云图

（二）基于主题识别的舆论数据集

首先，我们根据分词表，计算 TF-IDF 加权频数系数，据此从每一条句数据集中提取最多 12 个关键词形成句-词数据集，同时清除掉无关键词的无效句，最终留下 1882 条数据。其次，我们使用了 Python 用于机器学习的第三方库 Pytorch[1]，使用 Word2Vec 的 Skip-Gram 词向量模型，结合负样本采样方法，消除 Word2Vec 模型因忽略上下文之外的词语对中心词的语义影响而造成的误差，将文本数据转换为 embedding 向量值，形成句-词-词向量数据集。

训练模型后，以聚类的中心词-关键词组作为模型检验的选用词-参考词组，通过检索，计算词汇表所有词语与选用词的 embedding 向量值的余弦相似度（由于向量为多维而非二维，不使用欧氏距离，也无法作出散点图展示），取前 9 个最相似的词语及其相似度，得到选用词-相似词-相似度表。通过对相似词与参考词以及对相似度的比较，评估模型拟合度。

经过反复尝试与调整，最终确定模型参数如下：（1）K 负样本随机采样数量：18 词；（2）C 周围单词的数量：5 词；（3）NUM_EPOCHS 训练轮数：1 轮；（4）VOCAB_SIZE 词汇表大小：2063 词；（5）BATCH_SIZE 批处理大小 128 词 / 批；（6）LEARNING_RATE 学习率：0.4；（7）EMBEDDING_SIZE：300 维。经过训练，模型在第 101 次迭代时收敛，损失值为 131.689，将根据优化参数生成的词向量模型保存。其选用词-相似词-相似度表（部分）如表 3。

[1] 开源于：https://github.com/pytorch/pytorch.

表 3　选用词-相似词-相似度表（部分）

	相似词 1	相似度	相似词 2	相似度	相似词 3	相似度	相似词 4	相似度	相似词 5	相似度
一共	删	0.208	努力	0.204	草莓	0.197	发生	0.191	还以	0.177
高速	男	0.606	死	0.598	女	0.597	钱	0.586	好	0.585
死	女	0.952	胖猫	0.948	男	0.948	骂	0.925	谭竹	0.914

然后，我们使用 Python 用于机器学习的第三方库 scikit-learn[1]库中的多个模块包，采用 K-means 聚类方法实现主题识别。首先，我们通过通过轮廓系数确定最优簇数，绘制簇数 2 到 9 范围的轮廓系数折线图如图 8。根据手肘法确认最佳簇数为 3，即最佳主题分类数为 3。

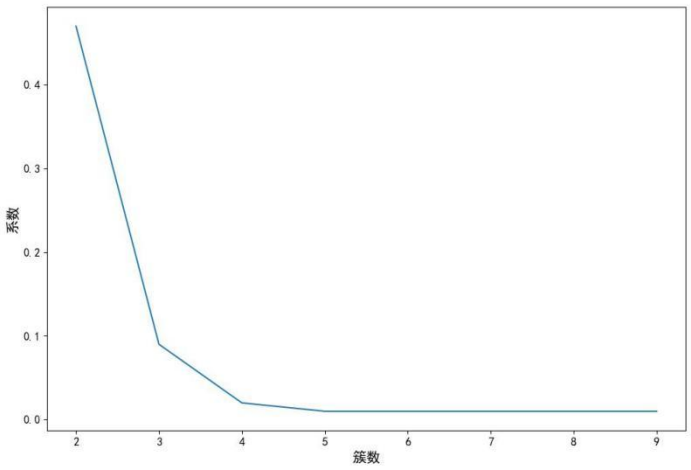


图 8　轮廓系数折线图

在该簇数下对词向量进行 K-means 聚类获得最终主题分类结果，形成句-词-词向量-主题表。同时，我们将 300 维词向量使用 TSNE 算法降维为双维和一维，形成句-词-词向量-主题-一维词向量，绘制双维聚类散点图如图 9。

[1] 开源于：https://github.com/scikit-learn/scikit-learn.

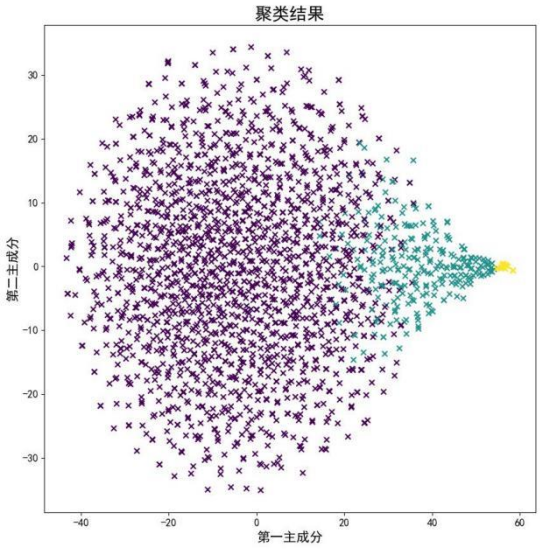


图 9　双维聚类散点图

由双维聚类散点图可得，聚类结果较好地区分了三类主题的分布，且三类主题具有鲜明特征、存在显著的热度差异。

最后，根据主题下词向量的余弦相似度查找得到有包含簇中心的 10 个中心词，根据主题下的词频查找得到 10 个高频词，使用标签编码人工为主题进行描述，构建主题-主题描述-中心词组-高频词组表如表 4。并且，我们通过句-词-频数表、句-词-词向量-主题-一维词向量表和主题-主题描述-中心词组-高频词组表分别通过词、主题进行内连接，形成舆论主题数据集（部分见表 5）。

表 4 主题-主题描述-中心词组-高频词组表（附词频总数、词数）

	主题描述	中心词组	高频词组	词频总数	词数
主题 1	关注经济问题、思考舆论操纵	一共 删 努力 草莓 发生 还 以 加害者 出去 男性 可悲	挣 拉 最 换 少 记录 流量 瓜 不好 纯	7060	1759
主题 2	对性别话题的评价与讽刺	高速 男 死 女 钱 胖猫 好 谭竹 龟男 看	真的 会 舔狗 发 姐姐 同情 xxn 出来 只能 点	3359	250
主题 3	议论事件本身、求证事实	死 女 胖猫 男 骂 谭竹 钱 好 看 现在	胖猫 女 男 吃人血馒头 龟男 死 谭竹 看 骂 反转	1117	15

表 5　舆论主题数据集（部分）

句	词	频数	词向量	一维词向量	主题	…
21 岁 年纪 12 精神 存 确实 牛 逼 低能	21	8	[0.0006494,……	1.597	1	…
	岁	9	[-0.0004473,……	-0.520	1	…
	年纪	4	[0.0009117,……	-2.551	1	…
…	…	…	…	…	…	…

由主题-主题描述-中心词组-高频词组表可得，三个主题的词频总数虽然差距显著，但是其数值都较大, 聚类合理。百度贴吧中对于“胖猫”事件展开讨论时，主要在经济问题、性别问题和事件本身上进行讨论。其中，经济问题讨论度最高，性别问题次之，事件本身讨论度最小。

（三）基于时间序列的舆论主题分析

根据舆论主题数据集，我们对性别主题与其他衍生主题在时间网络上的演变共性和差异展开探究。考虑到本研究中采用的公众舆论数据集的所有周期大约为一个月，不适合以天为单位进行切片。此外，为了进一步细化时间尺度，更直观地展示衍生舆论的演变趋势，我们最终选择以“三小时”为单位切割时间数据。

根据分析主体，选定主题回帖数和主题向量作为变量构建时间序列数据。我们统计了数据集中发布日期与主题回帖数的关系，最终构建了可分析的时间序列数据{时间，主题回帖数}。根据实验设计中的实验步骤，首先使用 Python 用于数据分析的第三方库 statsmodels[1]，通过 ADF 检验测量时间序列数据的平稳性如表 6。

由表可知，各时间序列均满足：ADF 在 95%的置信区间内，p 值远小于 0.001，说明时间序列数据均为平稳序列，可以直接将原始序列用于 ARIMA 模型分析，而不需进行差分处理。接下来，我们将时间序列数据{时间，主题回帖数}导入 ARIMA 模型进行时间-主题数量。

随后，结合 ACF 自相关图和 PACF 偏相关图确定 ARIMA 模型中的系数和系数的取值范围，并比较不同系数的 ARIMA 模型的结果，选取最佳系数来绘

[1] 开源于：https://github.com/statsmodels/statsmodels.

制每个主题在时间网络上的形成和演化图表。[1]每个主题下的 ARIMA 模型测试结果如表 6。模型的信息准则由 BIC 表示，根据模型的拟合度选择模型。此外，每个模型的残差不相关，时间序列数据的残差符合随机序列的分布，没有异常值，显示出良好的拟合效果。[2]

表 6　ADF 检验表（附 ARIMA 模型选用-BIC）

时间序列	ADF	ADF-pValue	模型选用	BIC
{时间，主题 1 回帖数}	-4.723<-3.507***	7.62e-05***	ARIMA（1,1）	797.62
{时间，主题 2 回帖数}	-4.822<-3.513***	4.93e-05***	ARIMA（1,1）	904.32
{时间，主题 3 回帖数}	-5.535<-3.515***	1.76e-06***	ARIMA（1,1）	560.64

最后，根据表所示的每个主题的 ARIMA 模型测试结果，我们建立了每个主题在时间网络上的形成和演化的测量和拟合图表。每个主题的实际值和预测值曲线如图 10(a)(b)(c)所示，其中横坐标是切片后发帖的时间，纵坐标是帖子数量。此外，蓝色曲线表示实际值，黄色曲线表示预测值。

由图 10(a)(b)(c)可知，三个主题下的实际值与预测值走势吻合，但对数值的预测存在差距，表明 ARIMA 模型捕捉性别议题的变化趋势的能力一般。三个主题下时间序列下的表现，呈现出相似的特点：三个主题在事件初期迅速达到高峰，随后逐渐回落并保持在较低水平，但在某些时间点上仍然有明显的波动，在 2024-05-29 左右三个主题迎来第二波峰，第二峰值均高于第一峰值。

经济主题的两波峰值差距较小，而事件主题的两波峰值差距较大，反映出对事件本体的讨论虽然占比较小，但在警方通报后占比提高。此外，值得注意的时性别议题在三个主题中更早到达第二峰值。

[1]　详见附录 2。
[2]　详见附录 2。

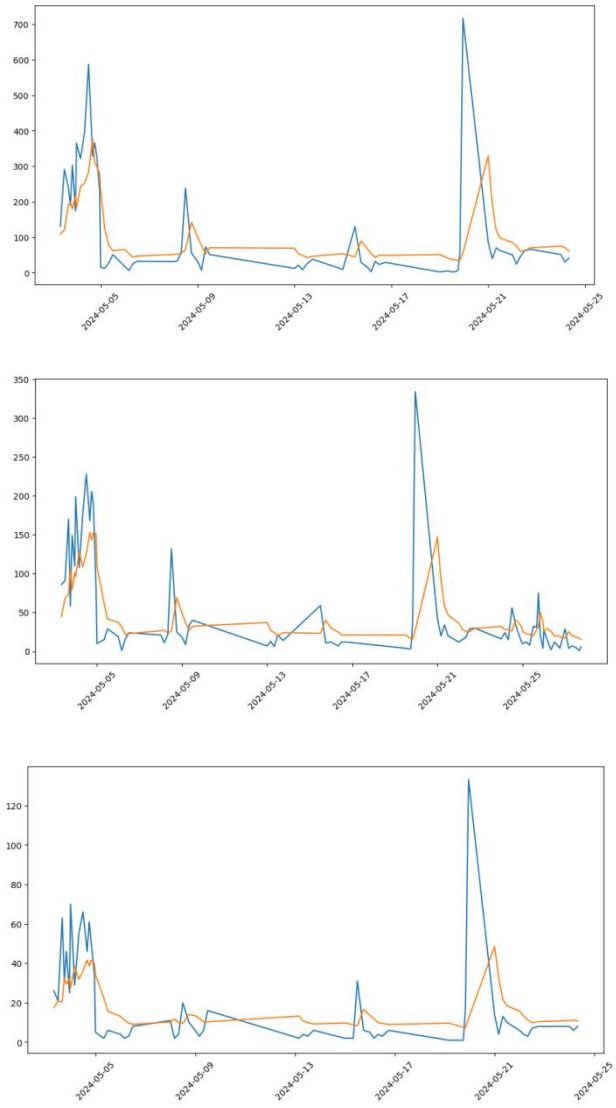


图 10(a)(b)(c)　主题 1、2、3 的 ARIMA 时间序列图

（四）主题演化规律

根据频数分析、主题识别和时间序列预测，可以总结百度贴吧对“胖猫”事件的主题演化规律：

早期讨论主要集中在事件的事实描述和初步反应，随后逐渐转向对性别角色和经济因素的深入探讨。经过时间序列分析，结合研究背景中百度贴吧舆论在该事件中的特殊表现，即互动声量在第二峰值才迎来爆发，可以认为性别议题实际上激发了百度贴吧用户展开互动讨论。此后，用户的讨论从对事件的表层反应逐渐深入到对社会现象和结构性问题的思考。不同主题在舆情时间序列中的出现频率和强度也反映了公众关注点的倾向。“经济问题”在时间序列上的讨论度始终高于其他主题，“性别问题”主题在事件初期和高潮时期显著增加，而“事件本体”则在事件后期讨论热度有所提升。

在讨论经济问题时，贴吧用户既涉及对舆论操纵的思考，还关注事件相关的经济问题。一方面，这表明在当前社会中，舆论与经济问题的相互影响越来越受到公众关注。词组如“挣”、“流量”等反映了网民对经济利益驱动下的舆论操控的敏感。另一方面，这体现出公众对情感经济交换的复杂看法。用户对“胖猫”在情感上进行巨额经济交换的合理性各抒己见，反映出公众在现实生活中对于性别关系、恋爱中的经济行为及责任分配的关注和心理需求。

在讨论性别问题时，贴吧用户关注对男性和女性行为的讨论，体现了性别议题在社交媒体上的高热度和复杂性。词组如“舔狗”、“xxn（小仙女）”等表现用户对性别角色和行为的刻板印象和调侃。值得关注的是，该主题的聚类中心词“高速”表现了部分用户对在性别问题下的持续争辩与冲突表示疲惫，希望公众应该更加关注“梅大高速塌方”事件等社会突发事件。

在讨论事件本身时，贴吧用户对事件中涉及的道德与公平问题表达了高度的关注。嘲讽“胖猫”和抵制散布谣言的行为，反映了网民对事件真实性的关注以及对不实信息传播的抵制。这一主题中的高频词组如“吃人血馒头”、“骂”等反映了网民对事件中涉及人物和行为的强烈情感反应或道德评判。“看”、“反转”等词语也反映了用户在参与讨论时不仅关注事件本身，还重视信息的真实性和完整性，有探索事件多面性、求证事实的倾向。

随着事件的发展，用户的情感态度从最初的震惊和愤怒逐渐演变为理性讨论和反思,反映着百度贴吧用户在讨论中并不仅仅只有激烈的情绪和对性别问题防御性，而是也会关照社会、政治、经济的整体情况，强调事实求证与社会公正。

五、 研究结论

（一）性别与经济舆情形成

“胖猫”事件产生的原因，从浅层次分析，是个人情感纠纷与经济利益冲突的直接爆发。男生“胖猫”与女友谭竹之间的情感纠葛、经济往来争议以及最终的悲剧结局，成为事件的直接导火索。从根本来说，事件发酵爆发的社会原因在于现代社会中性别角色认知的分歧、网络环境下信任缺失、以及公众对情感经济交换的敏感性，这些深层次因素为事件提供了肥沃的土壤。特别是在网络社交平台的放大效应下，个体事件迅速演变为具有广泛社会影响力的公共议题。

在事件发展的过程中，意见领袖如胖猫的姐姐和谭竹、知名人士如胡锡进、社交媒体如微博和抖音、主流媒体如人民日报等，他们的发声直接影响了舆论的走向，为各自的立场提供了支撑，或者加剧了公众的分化，或者统一着公众的集体诉求与意见。而百度贴吧等社区型平台，尽管不如微博等平台那样集中出现权威意见领袖，但用户通过帖子、回复等形式积极参与，形成了自发的站队现象，进一步推动了舆论的极化。平台内的讨论虽然分散，但同样促进了不同观点的碰撞与情绪的累积。

公众对于“胖猫事件”的事实性信息与相关价值判断的意见表现得极为分裂。一方面，许多人基于同情弱者的心态，对胖猫的遭遇表示深切同情，认为其为爱情付出巨大代价，是情感和经济双重受害者；许多人认为其心理脆弱、在情感中表现得过于软弱，对胖猫的选择进行嘲讽。另一方面，也有人质疑胖猫姐姐的动机，认为其可能借机获取关注与利益；此外，对谭竹的指责也反映了对“捞女”形象的负面刻板印象。这种分裂意见的交流中，议题逐渐超越了事件本身，扩展到了性别对立、家庭责任、网络暴力等多个社会议题，体现了公众对更广泛社会问题的关注与担忧。然而，当争论超越一定的界限，原本基于事实的讨论容易演变为对个人的人身攻击和道德审判，如性别对立的言论泛滥、对双方家庭成员的网络暴力等，这种过度的情绪化表达不仅损害了个体权益，也破坏了公共讨论的健康环境。

最终，在警方通报之后，形成了一种相对统一的“意见核”。这一“意见核”的特征体现在：首先，它凝聚了一定规模的社会成员共识，即无论事件的

具体细节如何，都需要通过法律程序来公正解决；其次，它稳定了意见结构，不同声音虽存，但都开始围绕法律框架下的事实认定展开；最后，确立了有序的意见表达，减少了非理性言论，有助于构建一个更为理性的舆论环境。这意味着，尽管事件本身依然存在诸多争议，但社会对于事件处理的态度和方式已经趋向成熟和理性，为解决类似问题提供了借鉴。

综上所述，“胖猫事件”作为一个复杂的社会现象，不仅关乎个人情感与经济纠纷，更折射出网络社会中性别对立、舆论极化、信息真实性危机等深层次问题。事件中，公众情绪与商业行为的交织、意见领袖的导向作用、以及网络平台的管理挑战，共同塑造了舆论的多元化面貌。警方通报和官方介入虽为事件带来一定秩序，但事件的长期影响凸显了网络环境下维护理性讨论、打击网络暴力及促进性别平等的重要性。这一案例再次警示，真相的探求与公共情绪的健康引导是现代社会治理不可或缺的部分。

（二）百度贴吧舆情特征

百度贴吧作为一个男性用户占主导的社交平台，其舆情传播具有独特的特征。男性用户的主导地位影响了讨论的方向和内容，使得其中的舆论主题在时间序列上表现一致。同时，某些性别偏见和刻板印象在讨论中被放大。在“胖猫事件”期间，除了对事件本身的讨论，百度贴吧的用户讨论集中在性别和经济问题上。这些讨论的主题在时间上表现出明显的演化特征，反映了用户对事件的关注点和情感态度的变化。

主题识别和时间序列分析显示，用户讨论从简单的事实陈述快速过渡到对性别角色、经济压力等深层次社会议题的剖析，表明百度贴吧用户在面对公共事件时，具备将个体案例引申至更广泛社会议题进行批判性思考的能力，反映了男性用户群体对于情感正义、性别平等、经济独立与网络诚信的深度关切。经济压力作为贯穿整个讨论的核心主题，凸显了当前社会环境下经济因素对公众心理和行为的深刻影响，以及用户对此的普遍关切。此外，用户情感的演变轨迹从情绪化的直接反应转为更为理性和多元的视角，尽管存在初期的情绪化批评和标签化行为，但后续讨论中不乏对事实的探究、社会公正的呼吁以及对性别问题的深度反思。一方面，这体现出百度贴吧用户讨论的复杂性和多维度。用户在批判之余，通过参与讨论，不仅在寻求事件的真相，也在表达对社会整体状况的关怀，对个人情感价值的捍卫，以及在复杂性别关系中的自我定

位。另一方面，这也显示着网络公众平台在促进社会议题讨论、增进公众意识觉醒方面的潜力。此外，这种参与既是对现实社会中某些问题的映射，也是在虚拟社区中寻求认同感和归属感的一种体现。

因此，本研究表明，百度贴吧作为一个男性用户占主导的社交平台，其舆论场域虽面临性别偏见放大的挑战，但也孕育了深入探讨社会、政治、经济议题的土壤，反映出用户群体对于社会不公的敏感度以及寻求改变现状的意愿。

（三）舆情中的政治经济

“胖猫事件” 在百度贴吧等网络平台上引发的舆论风暴，不仅是一起个人情感悲剧的公共讨论，更是社会结构、经济关系、权力运作以及文化意识形态在数字空间中的集中体现和互动结果。

事件中，“胖猫”姐姐和女方谭竹各自通过社交媒体发布信息，试图引导舆论走向，这体现了在数字时代，个人和组织可以利用粉丝基础、网络影响力等社会资本来影响公众舆论，从而在无形中争夺对事件叙事的控制权。这种争夺体现了网络空间中社会资本转化为实际权力的过程，即谁能更有效地动员和引导网络舆论，谁就能在一定程度上影响事件的舆论导向和后续处理。

事件中出现的性别对立言论，如 “捞女” 与“龟男”等标签的使用，反映了性别政治在当代社会中的持续作用。这些标签不仅加剧了性别之间的隔阂，而且在某种程度上掩盖了更深层次的阶级问题。事件中对经济纠纷的讨论，实质上触及了社会经济不平等和阶级差异，即不同社会阶层在恋爱关系中的资源分配不均，以及由此产生的社会心理影响。

网络平台上的信息过载与真伪难辨，使得 “胖猫事件” 中的真相变得扑朔迷离。这不仅考验了公众的信息甄别能力，也突显了数字时代媒介环境对真相建构的复杂性。一方面，信息的快速传播加速了事件的热度，另一方面，虚假信息和舆论操纵的盛行，反映出资本和权力如何利用信息不对称来引导公众情绪，维护或破坏特定利益。政府及平台方对网络暴力、性别对立言论的管理和处置，体现了在维护公共秩序与言论自由之间的平衡尝试。这不仅是一个法律和制度层面的问题，也是一个政治经济问题，即如何在保障公民言论自由的同时，防止网络空间成为非理性情绪、谣言和群体极化的温床，确保网络环境健康有序，是当前社会治理面临的重要课题。

（四）研究局限与展望

本研究主要基于百度贴吧的数据，可能存在单一性，未来研究可以扩展到其他社交平台和更广泛的用户群体，进行适当的比较分析，以更好地解释百度贴吧平台的舆论特征，探讨不同社交平台在类似事件中的舆情传播特征和用户反应差异。

其次，机器学习和深度学习方法在文本分析中的应用虽然有效，但在主题识别和情感分析上仍有改进空间，未来可以结合更多的自然语言处理技术提升分析精度。情感态度的变化不仅反映在文本内容中，也可以通过情感分析模型进行量化。

此外，由于收集到的数据集较小，数据的时间分布于整体规律不一致，本研究 ARIMA 模型的拟合精度较低，不能够很好地反映实际舆论演变，也无法在时间序列上进行有效预测，可以在补充数据的情况下再次展开探讨。

最后，本研究主要是基于机器学习的文本分析，后续研究中可以结合更多的社会学和心理学理论，通过问卷调查、网络民族志等调查方法，使用内容分析、话语分析等研究方法，对网络舆情中的性别和经济问题进行多维度分析，揭示其背后的深层次原因和影响机制。

六、 策略建议

对“胖猫”事件在百度贴吧舆情的时间序列研究，可以为理解社交媒体上的性别和经济问题讨论提供新的视角。这项研究不仅揭示了性别和经济问题在网络舆情中的演化规律，也为相关领域的研究提供了数据支持和分析方法。研究结果可以应用于社会治理和公共政策制定，为构建和谐社会和提升社会治理水平提供参考。

研究揭示了性别不平等和经济压力是事件讨论中的核心主题。这反映了公众对这些问题的高度关注，为政府和相关机构提供了重要的参考。公共政策制定需聚焦性别平等教育与经济压力缓解，通过广泛宣传教育与具体经济援助措施，改善社会认知结构与个体生活条件。

理性讨论和反思阶段的出现，表明公众在经历了情感宣泄后，逐渐趋向于理性思考和建设性讨论。这为社会治理提供了启示：通过引导和促进理性讨论，可以有效化解矛盾，促进社会和谐。

在社交平台治理方面，强化监管与促进正向互动是关键。这要求平台方既要严格执行规范，也要优化算法导向，保护在平台中占少数的女性群体的合理诉求，营造一个积极健康的网络交流环境。同时，需要避免个人和组织利用粉丝基础、网络影响力等社会资本影响公众舆论，不断加强主流媒体公信力和话语权，正确引导舆论，有效治理网络乱象。

此外，网络舆情监测中需要及时捕捉舆情热点，通过分析讨论主题和情感态度的变化，可以迅速识别出舆情热点并采取相应措施。例如，在事件初期的震惊和愤怒阶段，相关部门可以迅速做出回应，安抚公众情绪，避免事态恶化。关注讨论主题的演化，理解公众关切、积极回应公众疑问，并进行有针对性的政策引导和宣传。

七、 参考文献

[1] 韩运荣, 白岩冰. 大数据视角下网络舆论研判的原理与方法[J]. 现代传播: 中国传媒大学学报, 2015 (11): 153-154.

[2] 韩运荣，喻国明.舆论学原理：起源、方法与应用[M].北京：中国传媒大学出版社，2020.

[3] Cai M, Luo H, Cui Y. [Retracted] A Study on the Topic-Sentiment Evolution and Diffusion in Time Series of Public Opinion Derived from Emergencies[J]. Complexity, 2021, 2021(1): 2069010.

[4] 黄文森, 杨惠涵. 生态、圈层、可见性：社交网络舆情空间结构与平台逻辑[J]. 中国出版, 2024(6): 21-27.

[5] 晋良海, 王昕煜, 张文, 等. “4·29”特别重大房屋倒塌事件舆情主题聚类及演化研究[J]. 安全与环境学报: 1-12.

[6] 陈云松. 观念的“割席”——当代中国互联网空间的群内区隔[J]. 社会学研究, 2022, 37(4): 117-135+228.

[7] 胡杨涓, 胡千红. 虚拟社区中的用户特征与情绪表达——对“知乎”社区五类新闻议题讨论的实证分析[J]. 青年记者, 2019(33): 22-24.

[8] 赵艳娇. 网络空间的社群共同体——基于百度贴吧粉丝群的考察[J]. 北方民族大学学报(哲学社会科学版), 2019(5): 82-87.

[9] 走心发布丨贴吧垂类生态数据报告_兴趣[EB/OL]. [2024-06-28]. https://www.sohu.com/a/www.sohu.com/a/447474731_165158.

[10] Yan X, Guo J, Lan Y, et al. A probabilistic model for bursty topic discovery in microblogs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2015, 29(1).

[11] Li C, Sun A, Datta A. Twevent: segment-based event detection from tweets[C]//Proceedings of the 21st ACM international conference on Information and knowledge management. 2012: 155-164.

[12] Mathioudakis M, Koudas N. Twittermonitor: trend detection over the twitter stream[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. 2010: 1155-1158.

[13] Stieglitz S, Mirbabaie M, Ross B, et al. Social media analytics–Challenges in topic discovery, data collection, and data preparation[J]. International journal of information management, 2018, 39: 156-168.

[14] 董晨宇,林琦桁.“黏连剂”与“节拍器”：网络舆论议题的主题演化——对李佳琦“怒怼网友”事件的多维社会网络分析[J].传媒观察,2024(06):24-34.

[15] An L, Zhou W, Ou M, et al. Measuring and profiling the topical influence and sentiment contagion of public event stakeholders[J]. International Journal of Information Management, 2021, 58: 102327.

[16] Yan X, Guo J, Lan Y, et al. A biterm topic model for short texts[C]//Proceedings of the 22nd international conference on World Wide Web. 2013: 1445-1456.

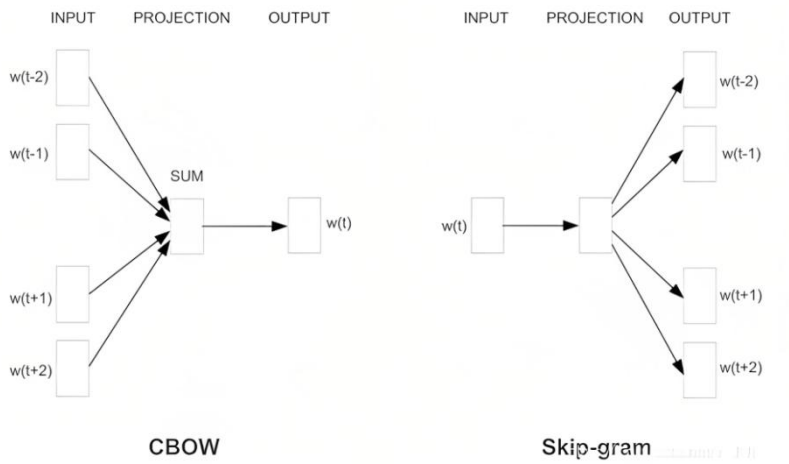
[17] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies. 2013: 746-751.

八、 附录 1 数学模型原理

（一）Word2Vec 模型

CBOW(Continuous Bag-of-Word)：以上下文词汇预测当前词，即用$w_{t-2}w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$去预测 w_t

SkipGram：以当前词预测其上下文词汇，即用w_t去预测$w_{t-2}w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$



首先，整个网络过程，我们需要做的是用输入的词去预测输出的词

其中输入层的单词w_i使用one-hot来表示的，即在上图中$x_1, x_2, x_3, \ldots, x_V$只有$x_k$为1，其余为0，其中$k$可以是输入的词在词汇表中的索引下标

之后就是经过词向量矩阵W连接输入层和隐藏层

其中由于X中只有一个1，因此经过与W相乘 (*相当于取出W中的的第k行，实际也就是输入单词的w_i的N维的词向量，使用v_{w_I}表示，来作为隐藏层的值*):

$$\mathbf{h} = W^T \cdot X = v_{w_I}^T$$

然后考虑从隐层的h到输出层Y，同样h经过矩阵W'相乘，得到一个$V \times 1$的向量u:

$$\mathbf{u} = W'^T \cdot h$$

其中u中的每个元素u_j，就是W'的第j 列用v'_{w_j}表示, 与h做内积得到：$u_j = v_{w_j}'^T \cdot h$，含义就是词汇表中第j 个词的分数

我们的目的就是要根据输入词w_i去预测输出的词，因此预测的词就取分数最高的即可

为了方便概率表示，使用softmax将u归一化到[0,1]之间，从而作为输出词的概率, 其实是一个多项分布, 也就是上图中的y:

$$P(w_j|w_I) = y_j = \frac{\exp(u_j)}{\sum_{k \in V} \exp(u_k)} = \frac{\exp(v_{w_j}'^T \cdot v_{w_I})}{\sum_{k \in V} \exp(v_{w_k}'^T \cdot v_{w_I})}$$

其中输入向量v_w与输出向量v'_w都称为词w 的词向量 (当然前面有说，一般使用前者输入向量作为词向量，而非后者)

至此前向过程完成，就是给定一个词作为输入，来预测它的上下文词，还是比较简单的，属于简化版的神经语言模型，这个过程中需要用到的参数有两个词向量矩阵W、W'

接着，明确训练数据的格式
对于一个训练样本(w_i, w_o)，输入是词wi的one hot编码，其维度定义为V的向量x，模型预测的输出同样也是一个维度为V的向量y
同时真实值w_o也是用one-hot表示，记为$t = [0, 0, 0..1, 0, 0]$，其中假设$tj^* = 1$，也就是说j^*是真实单词在词汇表中的下标，那么根据最大似然或者上面的语言模型，目标函数可以定义如下：

$$\begin{aligned} O &= \max P(w_O|w_I) \\ &= \max y_{j^*} := \max \log y_{j^*} \\ &= \max \log(\frac{\exp(u_{j^*})}{\sum \exp(u_k)}) = \max u_{j^*} - \log \sum_{k=1}^{V} \exp(u_k) \end{aligned}$$

一般我们习惯于最小化损失函数，因此定义损失函数:

$$E = -u_{j^*} + \log \sum_{k=1}^{V} \exp(u_k)$$

然后结合反向传播一层层求梯度，使用梯度下降来更新参数
先求隐层到输出层的向量矩阵W'的梯度:

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial w'_{ij}} = (y_j - t_j) h_i$$

这里面的y_j和t_j分别是预测和真实值的第j项，h_i是隐层的第i项
考虑：$\frac{\partial E}{\partial uj} = y_j - t_j$，直接对原始求导，如下：
先考虑E的对数部分:

$$\frac{\partial \log \sum \exp(u_k)}{\partial u_j} = \frac{\exp(u_j)}{\sum \exp(u_k)} = y_j$$

再看对u_j的梯度，综合求导$\frac{\partial E}{\partial uj} = y_j - t_j$，这个减法可以理解为输出层的第$j$项为预测值与真实值的差
因此梯度下降更新公式为：

$$w'^{(old)}_{ij} = w'_{ij} - \eta(y_j - t_j) h_i$$

然后结合反向传播一层层求梯度，使用梯度下降来更新参数
先求隐层到输出层的向量矩阵W'的梯度:

$$\frac{\partial E}{\partial w'_{ij}} = \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial w'_{ij}} = (y_j - t_j) h_i$$

这里面的y_j和t_j分别是预测和真实值的第j项，h_i是隐层的第i项
考虑：$\frac{\partial E}{\partial uj} = y_j - t_j$，直接对原始求导，如下：
先考虑E的对数部分:

$$\frac{\partial \log \sum \exp(u_k)}{\partial u_j} = \frac{\exp(u_j)}{\sum \exp(u_k)} = y_j$$

再看对u_j的梯度，综合求导$\frac{\partial E}{\partial uj} = y_j - t_j$，这个减法可以理解为输出层的第$j$项为预测值与真实值的差
因此梯度下降更新公式为：

$$w'^{(old)}_{ij} = w'_{ij} - \eta(y_j - t_j) h_i$$

整合为W'的列向量$\mathbf{v}'_{wj} = \{w'_{ij} | i = 1, 2, 3, \ldots, N\}$的形式如下:

$$\mathbf{v}'^{(old)}_{wj} = \mathbf{v}'_{wj} - \eta(y_j - t_j)\mathbf{h},\ j \in \{1, 2, 3, \ldots, V\}$$

也就是说对每个训练样本都需要做一次复杂度为V的操作去更新W'

最后，考虑隐层h的更新，其实也是输入层到隐层的矩阵W的更新，继续反向传播，跟神经网络的相同，输出层的V个神经元都会影响h_i:

$$\frac{\partial E}{\partial h_i} = \sum_{j=1}^{V} \frac{\partial E}{\partial u_j} \frac{\partial u_j}{\partial h_i} = \sum_{j=1}^{V} (y_j - t_j) w'_{ij} = W'_i \cdot P$$

其中W'_i是W'的第i行，这里为了方便书写，令$P = \{y_j - t_j | j = 1, 2, 3, .., V\}$，因此整合成整个隐层的向量$h$:

$$\frac{\partial E}{\partial \mathbf{h}} = W' \cdot P$$

得到一个N维的向量，上面已经介绍过，h就是词向量矩阵W的一行：$\mathbf{h} = W^T \cdot X = v^T_{wI}$，但是因为$X$中只有一 个1，因此每次只能更新的一行$v_{wI}$，其余行的梯度= 0，所以$v_{wI}$的更新公式为:

$$v^T_{wI} = v^T_{wI} - \eta W' \cdot P$$

（二）ARIMA 模型

1. 自回归模型**-AR**

p阶自回归过程的公式定义：$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$

y_t 是当前值　μ 是常数项　P 是阶数　γ_i 是自相关系数　ϵ_t 是误差

2. 移动平均模型**-MA**

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

3. 自回归移动平均模型 **ARMA**

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

4. 差分自回归移动平均模型

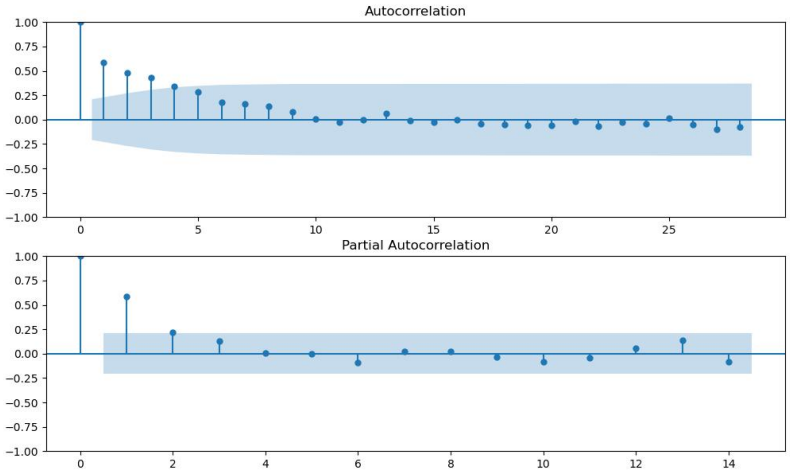
ARIMA(p，d，q)

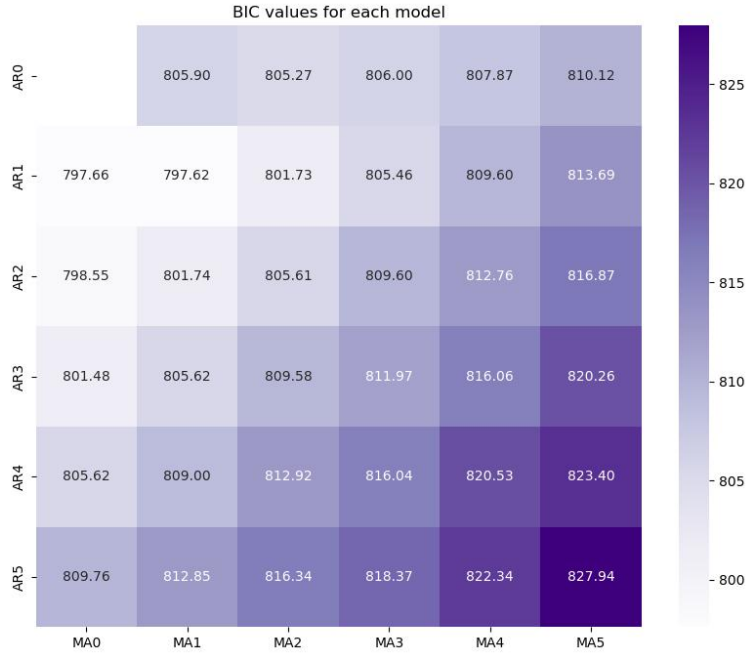
5. 自相关函数 **ACF(autocorrelation function)**

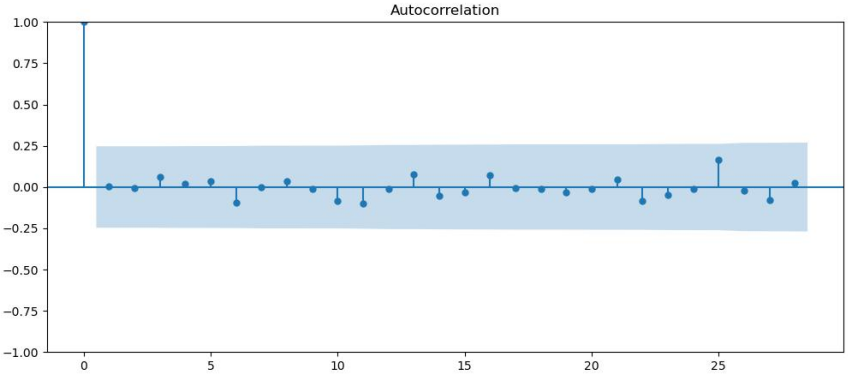
$$ACF(k) = \rho_k = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\mathrm{Var}(y_t)}$$

九、 附录 2 时间序列系数选择依据

（一）主题 1

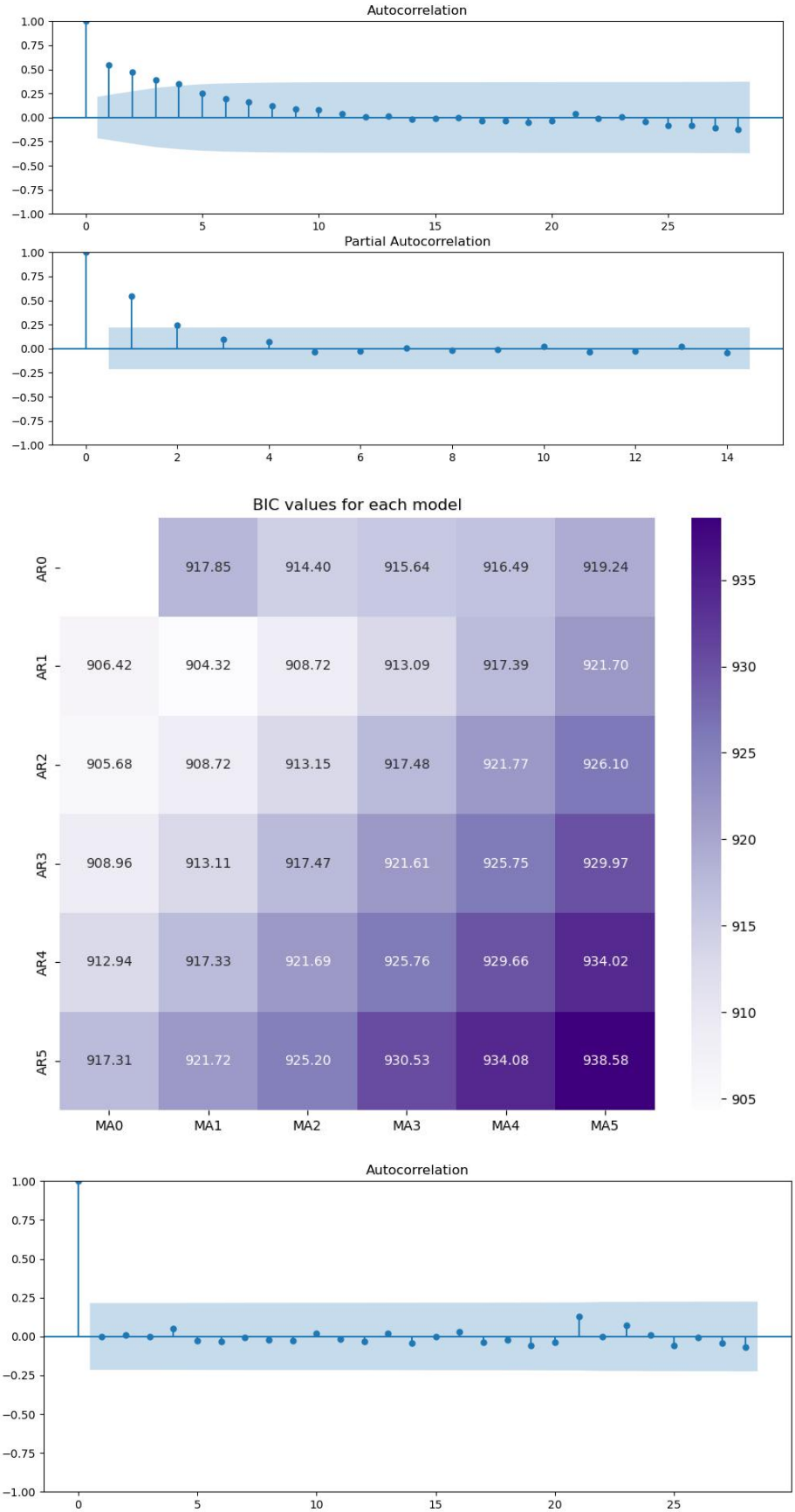






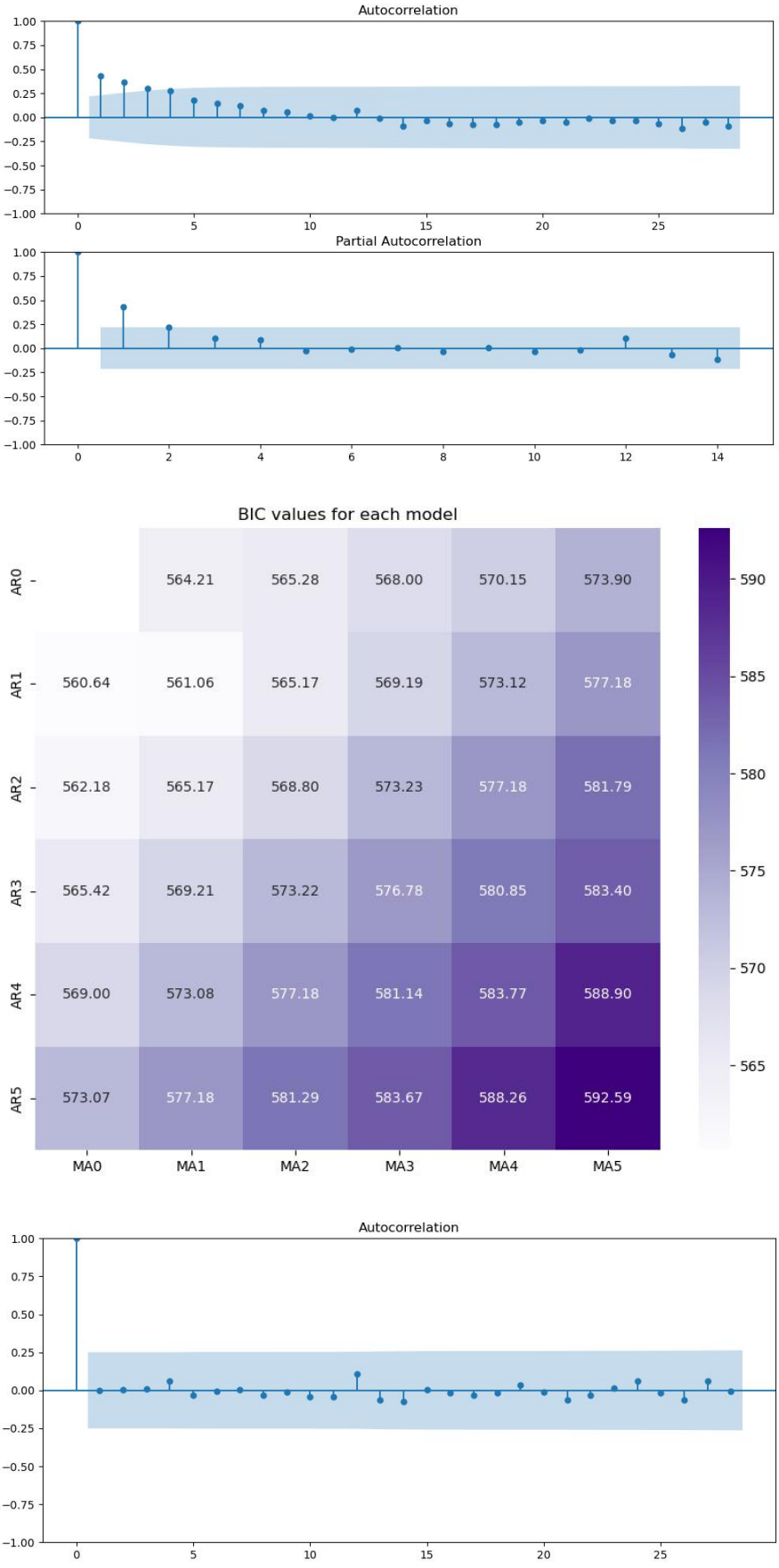
ACF 自相关图和 PACF 偏相关图、ARIMA 模型测试 BIC 热力图、数据残差图

（二）主题 2



ACF 自相关图和 PACF 偏相关图、ARIMA 模型测试 BIC 热力图、数据残差图

（三）主题 3



ACF 自相关图和 PACF 偏相关图、ARIMA 模型测试 BIC 热力图、数据残差图

十、 附录 3 其他附件

（一）舆情研究平台“舆情通”数据

2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”数据：

https://wish2333.fun/upload/舆论学-“胖猫”数据.zip

2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”-贴吧数据：

https://wish2333.fun/upload/舆论学-“胖猫”-贴吧数据.zip

2024 年 5 月 1 日至 2024 年 6 月 1 日“胖猫”-贴吧-2000 条内容源数据：

https://wish2333.fun/upload/舆论学-“胖猫”-贴吧-2000 条内容源数据.xlsx

（二）同义词表、停用词表

https://wish2333.fun/upload/舆论学-同义词表、停用词表.zip

（三）优化词向量模型

https://wish2333.fun/upload/舆论学-优化词向量模型.th

（四）数据处理代码

https://wish2333.fun/upload/舆论学-数据处理代码.zip

（五）舆论主题数据集

https://wish2333.fun/upload/舆论学-舆论主题数据集.xlsx