中国人民大學

RENMIN UNIVERSITY OF CHINA

“下沉的”视频平台与“上流的”文字平台？

——基于代表性视频平台与文字平台评论对性别问题典型事件的文本分析

孙子程 彭显威 钟东彤

孙大伟 杨天宇 黄丹晓 常炳骥

一.　问题的引入

1. 视频平台与文字平台的“刻板印象”

提起国内主流的社交平台，你对这些平台会有怎样的印象？如果你对这些平台都很熟悉，那脑海里一定会闪出各个平台的特点，诸如 B 站的二次元、混乱、宅文化、知乎的精致、高格调、精英主义……如果你对这些平台不甚了解，那也会在网络搜索中发现不同平台的刻板印象已经深入用户脑海，以至于出现下面这样的评价：

知乎——它的用户似乎是那些‘半高知’们，他们精英主义的论调，以高质量内容爱好者自居，所以被称作“一群从‘左’到‘右’再从‘右’到‘左’的‘半高知’们的聚集地”。

B 站——由于它的用户大多年轻，对自己感兴趣的事物表现出极高的热忱，但其言论往往并不成熟，所以又被称为 “皈依者”们的狂欢场。

上述这样的评价归根结底在于人们对视频平台和文字平台的刻板印象，伴随着近几年各种视频平台的兴起，人们总会为这种视频输出方式打上‘下沉’‘低龄’的标签。与此相对，以知乎为主的文字平台则充斥着各行各业高谈阔论的文章，早在 2010 年 10 月，知乎团队的一封内部邮件就写着：“我们相信一点，在垃圾泛滥的互联网信息海洋中，真正有价值的信息是绝对的稀缺品。知识——被系统化、组织化的高质量信息——都还存在于个体大脑中，远未得到有效地挖掘和利用。”不管如今的知乎是否真正实现知识有效地挖掘和利用，不可否认的是在各大社交平台中，知乎这类文字平台似乎总是更贴近人才、精英汇集的那一个。

就信息门槛而言，长图文>段子>音频>视频。门槛高的自然“精英化”，门槛低的自然“低俗化”，刻板印象就这样越来越根深蒂固了，如果真的存在此类刻板印象，那么不同平台的用户对于同一件事情显然应该持有符合其刻板印象的观点。现实是否真的如此，视频平台相比文字平台真的更加“下沉”？还是说刻板印象仅仅只是刻板印象？这就是我们小组这次主要研究的主题。

2. 选择代表性性别问题作为研究对象

本项目选取代表性性别问题“货拉拉女乘客坠车死亡事件”与“伊朗头巾事件”作为研究视频平台与文字平台用户可能的差异性。原因在于，近几年性别问题讨论度极高，不同的热点事件在各个网络平台均收到广泛关注，不论是视频平台还是文字平台都有针对该问题相关事件的内容与评论，事件的关注度不会因为平台不同存在显著差异。另一方面，性别问题作为一种与每个人切身相关的社会问题，相比于科技、经济、体育等方面的话题平台用户能够

参与讨论的门槛更低，各种职业、年龄、不同教育背景的人群在这一话题上都有相同发表意见的权利，并且也愿意参与该问题下的讨论。除此之外，人们在讨论性别问题时展现出了明显的对立观点，各个平台用户在性别问题相关内容与评论中往往存在鲜明的立场。本项目主要利用平台用户相关评论进行分析，所以性别问题下观点明确、立场鲜明的用户评论更有利于对比不同平台之间用户的差别。另外，本项目的是探究视频平台和文字平台是否如同刻板认知中前者的用户比后者更加“下沉”，而性别问题一方面与社会中每个人的生活息息相关，在男性与女性的交流与互动、婚姻等所有人都广泛参与的事件中实际就暗含了性别问题，几乎所有人都能够针对这些现象根据自身经验发表自己的看法；另一方面对性别问题的解读实际上涵盖了收入分配、社会权力结构与文化思潮变革等多方面视角，而不仅仅是普通人日常的生活经验和感受，这类型的解读需要有一定的专业知识支撑。所以考察性别问题时能够同时囊括网络平台中“下沉”用户的观点和具备更高知识水平用户的看法。

3.对性别问题代表性事件的简述

A. **货拉拉女乘客坠车死亡事件：**2021 年 2 月一名女子通过货拉拉服务进行搬家，在装车完成前该女子与货拉拉司机由于是否进行付费搬运服务产生争执，后在驱车前往目的地途中，司机通过货拉拉 APP 提前接下一单业务，并为节省时间私自更改行车路线。该女子在途中多次提出路线偏航，司机均为予理睬，在第四次提出质疑后，该女子将上半身探出窗外，结果导致从车窗坠车身亡。事件发生后涉事司机被警方以犯罪证据不足释放，但因社会舆论发酵，警方对案件再次勘验，最终判处涉事司机涉嫌过失致人死亡罪。期间社会舆论主要关注在该事件中女性与陌生男性单独乘车时路线偏航所导致的恐慌，直接造成其做出非理性行为，最终导致死亡。

B. **伊朗头巾事件：**2021 年 9 月，一名 22 岁的伊朗女子因为涉嫌“没有按照规定要求佩戴头巾”被伊朗警方逮捕，并在拘留期间死亡。该女子死讯在伊朗国内引起巨大影响，各地先是爆发以女性为主体的示威活动，而后示威人数持续增加，男性也加入示威人群中，逐渐演变为全国范围内的抗议甚至暴乱。人们呼吁要求废除女性在公共场所必须佩戴头巾的规定，废除对女性的一些偏见，并赋予女性更多权利。

二. 相关平台数据爬取与预处理

针对视频平台，本项目主要爬取“bilibili 弹幕网（以下简称 b 站）”与“货拉拉女乘客坠

车死亡事件”和“伊朗头巾事件”相关视频评论及弹幕。针对文字平台，本项目主要爬取“知乎”平台上述两件事件相关回答下一级评论与二级评论以及评论时间，并将“虎扑”相关数据中的正文作为补充。

平台	B 站	知乎	虎扑
爬取评论数量	40478	3683	81
评论类型	视频评论与弹幕	问答一、二级评论	帖子正文
单条评论长度	100 字内	100 字内	1000 字内

表 1:平台数据统计

之后将不同平台数据首先汇总于一张表格，并且利用 jieba 库对评论进行分词处理，去除评论中的停用词，并将每条评论的分词结果列为全新一行。接着将数据按照平台、事件两种分类方式分别汇总到不同表格中。

	comment	cut_comment
0	从重你麻麻，司机有什么错，✓玩意儿，你是不是车家请的✓水军啊	从重 麻麻 司机 错 ✓ 玩意儿 是不是 车家 请 ✓ 水军
1	阳光的杀猪盘女生，怕被找上门屡次搬家的色情软件HR[思考]	阳光 杀猪 盘 女生 找上门 屡次 搬家 色情 软件 HR 思考
2	我觉得老蔡讲的是最好的，乘客死亡的唯一原因是跳车，无论再往前司机的这些行为都是无法导致乘客死...	觉得 老 蔡 讲 最好 乘客 死亡 唯一 原因 跳车 往前 司机 行为 都 无法 导致 乘客...
3	法律要求司机：开车时不得吃喝、用手持设备打电话、玩手机，摸副驾大腿\n\n法律要求乘客：不要...	法律 要求 司机 开车 不得 吃喝 手持 设备 打电话 玩 手机 摸 副 驾 大腿 法律 要...
4	如果性别互换，肯定全都骂跳车的	性别 互换 肯定 全都 骂 跳车
...	...	...
44229	表达伟大的革命理念	表达 伟大 革命 理念
44230	不要试图和反动派讲道理	不要 试图 反动派 讲道理
44231	马超[尴尬]	马超 尴尬
44232	时无英雄，使马超成名[尴尬]	英雄 马超 成名 尴尬
44233	马超在某些方面确实是一匹快马，至少比某些媒体要更快更高[大笑]	马超 确实 一匹 快马 至少 媒体 更 快 更 高 大笑

44234 rows × 2 columns

图 1:所有数据分词处理结果

三. 整体数据与分事件的主题建模与词云分析

1. 针对整体数据的分析:

在讨论平台的差异性之前，首先对于整体数据应该有清晰的认识。所以本项目首先对总体数据进行分析。

A. 词云分析:

首先通过词云甄别在代表性男女问题事件中平台用户关注的重点。通过分词与数据预处理的表格，使用 Wordcloud 库进行词云图的绘制。

图 2:总体数据词云图

本项目在数据爬取过程中对于两个事件没有明显倾向，即爬取评论的不同事件相关内容没有明显数量上的差异，而以上词云分析表明所收集的数据中 “货拉拉事件”和“伊朗头巾事件” 用户评论数量接近，证明两个事件受到用户关注度也比较接近，用户对于两个事件没有存在明显的偏好。

B. 整体数据的主题聚类

为了实现降低特征空间维度和加深语意理解、挖掘文本信息的目的，本项目对总体数据进行了 tSNE 方法聚类，用主成分分析将数据降到二维，做出散点图将聚类结果可视化。

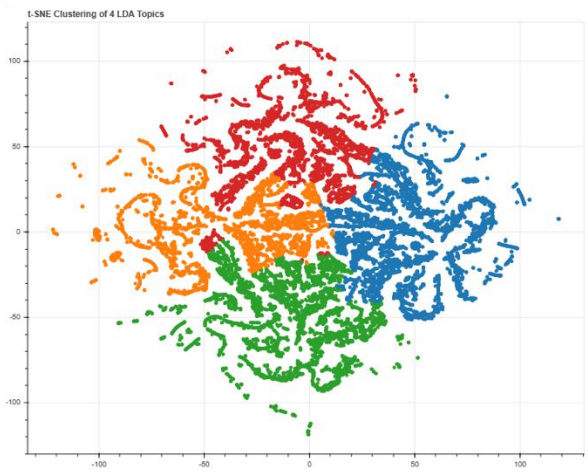


图 3:tSNE 聚类图

C. 主题建模与向量文本化分析

在完成整体数据等词云分析后，项目试图考察整体数据的主题分类情况，将数据中只出现过一次的词语删除并进行文本数据的结构化，最后利用 LDA 模型进行主题建模，指定模型将数据分为 4 个主题，结果大概如下：

主题 0：'0.012*"支持" + 0.011*"戴" + 0.011*"头巾" + 0.010*"5" + 0.008*"伊朗" + 0.007*"都" + 0.007*"司机" + 0.006*"拉拉" + 0.006*"跳车" + 0.006*"货"'),

主题 1：'0.020*"哈哈哈" + 0.009*"哈哈哈哈" + 0.006*"漂亮" + 0.006*"太" + 0.006*"司机" + 0.005*"好看" + 0.005*"直接" + 0.005*"呵呵" + 0.004*"都" + 0.004*"很"'),

主题 2：'0.008*"1" + 0.006*"确实" + 0.005*"加油" + 0.005*"司机" + 0.005*"自由" + 0.004*"没有" + 0.004*"偏激" + 0.004*"病" + 0.003*"哪来" + 0.003*"伊朗"'),

主题 3：'0.008*"笑" + 0.007*"伊朗" + 0.007*"都" + 0.006*"死" + 0.005*"宗教" + 0.005*"沙特" + 0.005*"美国" + 0.005*"好家伙" + 0.005*"错" + 0.005*"理解"')

在完成 LDA 主题模型建立后，对其进行可视化处理，分别绘制了不同主题的词云图并还通过 pyLDAvis 进行可视化处理。从不同主题的词云图来看，可以注意到由于数据以各个平台的短评论为主，因此存在大量口语化与表达情绪的用词。在基于总体数据构建的 LDA 模型中，可以认为 Topic0 主要倾向对两起事件的描述；Topic1 主要倾向于表达评论者的主观感受，含有大量语气词；Topic2 主要为围绕伊朗头巾事件的争论，并且争论语气较为激烈，内容较为尖锐；Topic3 主要倾向讨论伊朗头巾事件的历史文化与国际政治背景。从 pyLDAvis 的结果来看，除与词云相类似的结果以外，还能看出前两个主题有交集，其中一个主题主要是对两起事件的描述，另一个则为评论者的主管感受，证明了用户在描述事件客观事实时在一定程度上会结合自身的主观感受。

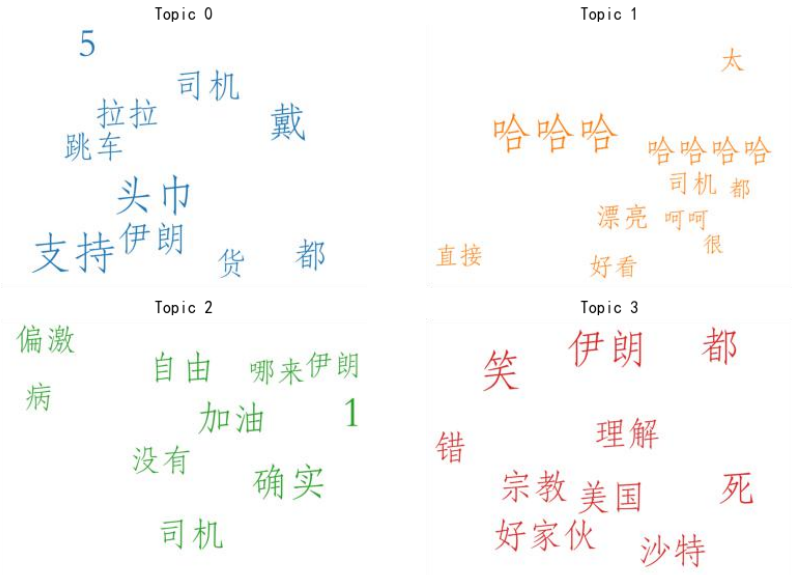


图 3:总体数据主题词云图

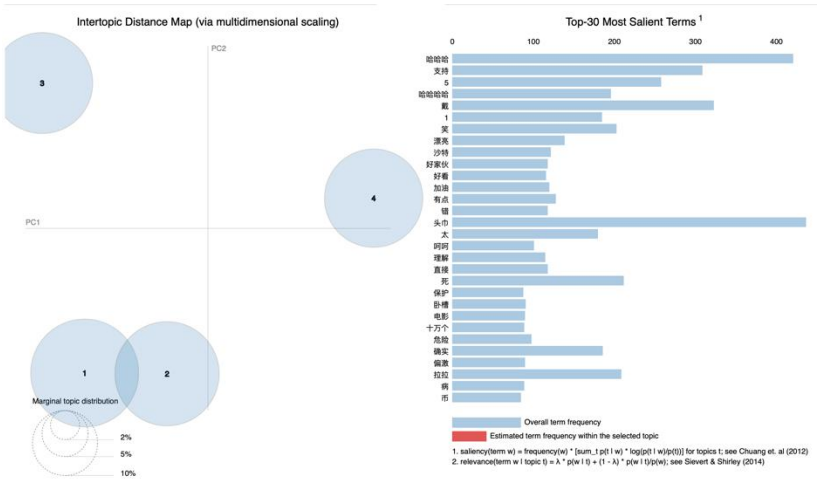


图 4:pyLDAvis 可视化结果

2. 分事件的主题建模与词云分析

在对整体数据进行研究后，考虑到两个事件虽然都是男女问题的代表性事件，但是两则事件的发生背景分处国内与国外，并且“伊朗头巾”还包含了伊朗本国的历史与文化，属于国际事件，这可能导致平台用户在对两个事件发表看法时存在差异性。

与整体数据分析类似，本部分通过分词与数据预处理的表格，使用 Wordcloud 库进行词云图的绘制。

图 5、6:“伊朗头巾”与“货拉拉事件”数据词云图

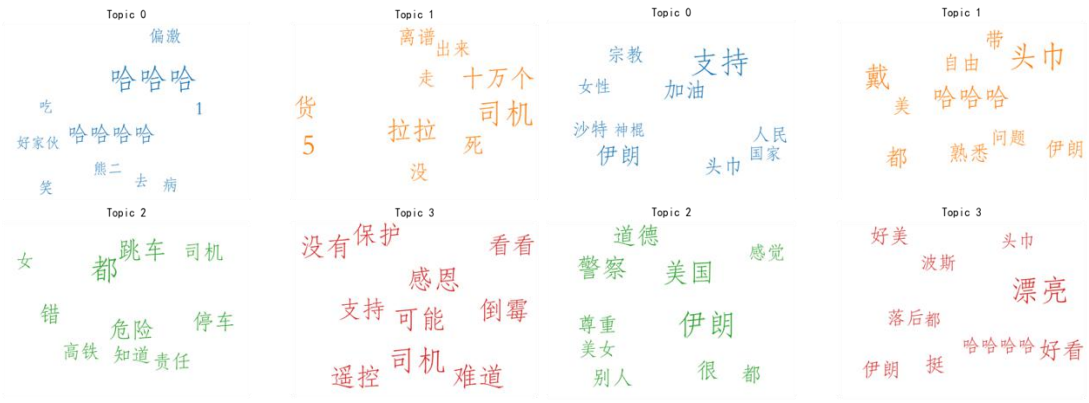


图 7:“货拉拉事件”主题词云图

图 8:“伊朗头巾”主题词云图

从上述结果来看，在基于货拉拉评论数据构建的 LDA 模型中，可以认为 Topic0 主要倾向表达评论者当下的直接感受，评论内容较为随意，倾向于表达情绪的语气助词；Topic1 主要倾向于认为该事件比较荒唐；Topic2 主要倾向于对货拉拉事件中司机行为的讨论，具体一些是讨论司机对于女生是否保护到位；Topic3 主要倾向于对表达对司机的情况的同情。在基于伊朗头巾评论数据构建的 LDA 模型中，可以认为 Topic0 主要倾向对中东传统宗教与文化势力的讨论，并且表现出了对世俗化与以人为本的体制的支持；Topic1 主要倾向对围绕头巾话题对个体自由的讨论，并带有一定讽刺意味；Topic2 主要倾向于讨论事件发生与美国之间的联系，以及伊朗的道德警察；Topic3 主要倾向于对伊朗女性的赞美（除了直接的对外貌的赞美，应该也包含了精神层面的赞美）。

另外不难看出，用户对两个事件的关注点存在差异。相比于“货拉拉事件”，“伊朗头巾”的国际事件的属性更加明显，用户不仅仅集中于对事件本身做出评论，“美国”、“沙特”、“国家”等词语表现了用户更发散性地讨论与国际政治有关的话题而不仅仅局限于性别问题本身。但是两事件都展现了用户对于男女问题的关注以及看待问题的观点存在差异，所以两事件可以作为探究“探究视频平台和文字平台是否如同刻板认知中前者的用户比后者更加‘下沉’”的数据材料。

四. 分平台主题模型与词云分析

在确定了上述事件可以作为分析材料后，本项目开始尝试对不同平台的评论文本进行分析。与上文中对整体数据研究方法相同，首先通过词云甄别在代表性男女问题事件中平台用户关注的重点，通过分词与数据预处理的表格，使用 Wordcloud 库进行词云图的绘制。其次考察整体数据的主题分类情况，将数据中只出现过一次的词语删除并进行文本数据的结构化，

最后利用 LDA 模型进行主题建模，指定模型将数据分为 4 个主题。

1. “知乎”的分析结果

主题建模结果：

主题 0：'0.030*"十万个" + 0.016*"看看" + 0.016*"好好" + 0.015*"司机" + 0.015*"br" + 0.014*"死" + 0.014*"拉拉" + 0.012*"货" + 0.012*"没有" + 0.012*"去"'

主题 1:'0.030*"遥控" + 0.029*"难道" + 0.019*"电影" + 0.019*"跳来跳去" + 0.019*"每次" + 0.019*"保护" + 0.018*"错" + 0.015*"遇到" + 0.015*"博主" + 0.015*"女拳"'

主题 2：'0.031*"偏激" + 0.031*"病" + 0.019*"缺乏" + 0.017*"常识" + 0.014*"能力" + 0.013*"理性" + 0.011*"群众" + 0.011*"朋友" + 0.011*"身边" + 0.011*"无穷的"'

主题 3:'0.029*"危险" + 0.027*"高铁" + 0.026*"停车" + 0.016*"跳车" + 0.016*"理解" + 0.015*"团结" + 0.015*"激化矛盾" + 0.015*"和谐" + 0.015*"少数" + 0.014*"都"'

图 9:“知乎”数据词云图

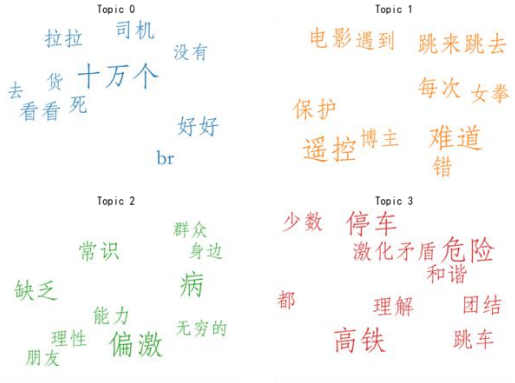


图 10:“知乎”主题词云图

在基于知乎评论数据构建的 LDA 模型中，可以认为 Topic0 主要倾向讨论事件本身；Topic1 主要倾向于较为激烈地支持司机，并且对部分对立面言论进行攻击；Topic2 主要倾向于质疑这次讨论中人们与舆论不够理性，显得比较偏激；Topic3 主要倾向于担忧是否会导致社会进一步撕裂。

2. b 站的分析结果

主题建模结果：

主题 0：'0.023*"好好" + 0.018*"高铁" + 0.018*"停车" + 0.016*"保护" + 0.016*"br" + 0.016*"直接" + 0.015*"知道" + 0.015*"缺乏" + 0.014*"危险" + 0.012*"常识"'

主题 1:'0.023*"微博" + 0.022*"看看" + 0.016*"豆瓣" + 0.016*"半点" + 0.016*"评论" + 0.015*"没有" + 0.015*"

司机" + 0.013*"跳下去" + 0.013*"整笑" + 0.013*"有点"'
主题 2:'0.034*"十万个" + 0.025*"遥控" + 0.024*"偏激" + 0.024*"病" + 0.024*"难道" + 0.017*"每次" + 0.017*"电影" + 0.017*"跳来跳去" + 0.017*"错" + 0.016*"都"'
主题 3: '0.024*"都" + 0.018*"去" + 0.018*"跳车" + 0.018*"理解" + 0.016*"激化矛盾" + 0.016*"团结" + 0.016*"和谐" + 0.016*"少数" + 0.014*"坏人" + 0.013*"完"'

图 11:“b 站”数据词云图

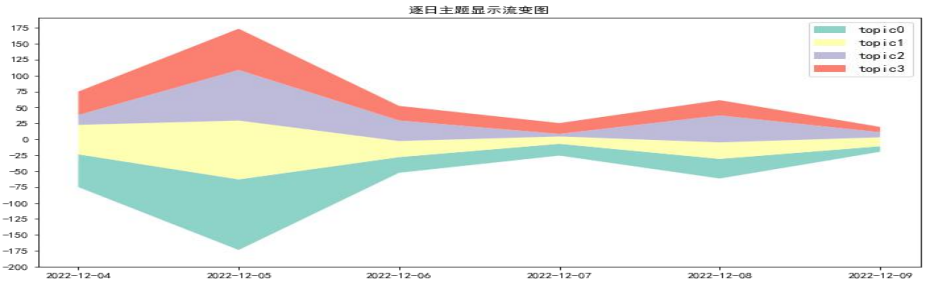
图 12:“b 站”主题词云图

在基于 b 站评论数据构建的 LDA 模型中，可以认为 Topic0 主要倾向讨论事件本身，认为女生没有对高速行驶的车上跳下的危险性有足够的意识；Topic1 主要倾向于讨论其它平台上对这件事的看法，其中豆瓣与微博正是我们传统认为的“更加支持女性的平台”；Topic2 主要倾向于批评舆论比较病态且偏激；Topic3 主要倾向于担忧是否会导致社会进一步撕裂。

3. 关注度时序性河流图

针对主题模型结果本小组对“知乎”各个主题在时间顺序上受关注程度进行探究，尝试探究是否存在用户对于不同主题关注度随时间变化的情况，这可能会揭示同一问题下人们关注的侧重点会有所改变。

将不同日期所对应的话题关注度用话题权重来表示（同一天内四个主题加总为 100，受关



注的主题权重更高），结果如下：

注意到随着时间的推移所有主题以相对均匀的方式变化，没有出现某一主题在某一时刻收到更多关注的情况。另外各个主题的变化可以证明事件发生第二天热度达到高峰，随后呈现下降趋势。

五. 分平台情感分析

1. 语料库介绍

情感分析是在研究人们针对某一事件或主题的主观意见和情感，情感分析的过程是对带有情感色彩的主观性文本进行分析研究。将情感分析应用于本小组所选取的‘伊朗头巾’和‘货拉拉’事件中，有利于分析平台网友对于事件表现出的态度或情绪倾向性。

鉴于本项目本次收集数据的评论性，在情感分析的语料库选择方面，使用的是simplifyweibo_4_moods 数据集，该语料库数据量大，情感类别丰富，共包含 36 万多条评论示例，分为四种情感：喜悦、愤怒、厌恶、低落，其中喜悦约 20 万条，愤怒、厌恶、低落各约 5 万条。利用 simplifyweibo_4_moods 语料库已标注好的数据搭建模型（SVC），利对所探究的评论进行文本情感识别。

图 13：“b 站”语料库基本情况

2. 文本评论情感识别

我们首先对文本进行初步清洗，去除掉一些停用词，然后分别进行词频统计和向量化操作，构建出评论分词模型，接着使用已标注的语料库数据集作为训练集对模型进行训练，最后利用模型对爬取评论内容的情感倾向进行预测，从而得到每个评论的文本情感识别。汇总

数量结果和比例结果如下表：

	B 站 货拉拉	B 站 伊朗头巾	虎扑 货拉拉	虎扑 伊朗头巾	知乎 货拉拉	知乎 伊朗头巾
0（喜悦）	781	1960	135	73	189	165
1（愤怒）	4119	5217	1130	376	1632	426
2（厌恶）	878	1071	281	77	428	111
3（低落）	1744	1079	386	90	616	114
总计	7522	9327	1931	616	2865	816
	B 站 货拉拉	B 站 伊朗头巾	虎扑 货拉拉	虎扑 伊朗头巾	知乎 货拉拉	知乎 伊朗头巾
0（喜悦）	10.38%	21.01%	6.99%	11.85%	6.60%	20.22%
1（愤怒）	54.76%	55.93%	58.52%	61.04%	56.96%	52.21%
2（厌恶）	11.67%	11.48%	14.55%	12.50%	14.94%	13.60%
3（低落）	23.19%	11.57%	19.99%	14.61%	21.50%	13.97%
总计	1	1	1	1	1	1

表 2:各平台情感识别结果

为了便于分析，我们将结果可视化为饼状图如下图

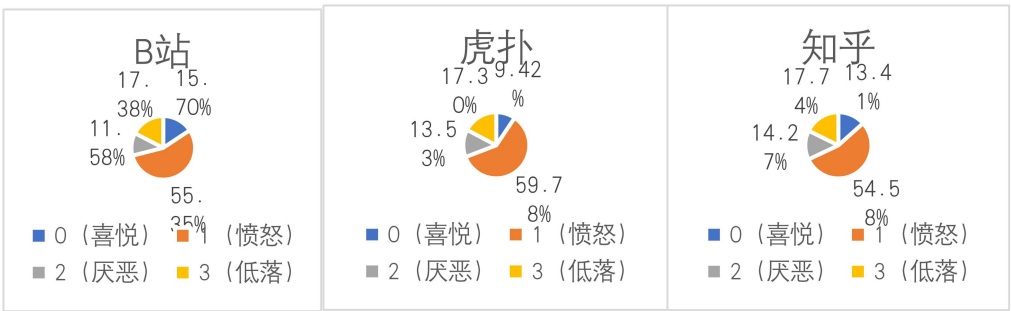


图 14:情感分析可视化结果

首先我们对四种情感在该事件语境下的意义做出解释：愤怒表示对货拉拉事件或伊朗头巾事件中对当事人一方或双方行为的极度反感；厌恶表示对事件中行为的反感；低落主要表示出对事件中受害人的同情和事件发生的悲哀；喜悦表示对事件中某些行为的愉快高兴心理，

从具体评论中可以发现不少幸灾乐祸、事不关己的言语。

从情感比例来看，三大平台愤怒情绪的占比都是最高的，均超过 50%，这一点是由于事件本身的性质决定的，无论是货拉拉中仅因口角导致的坠车悲剧，还是伊朗头巾事件中仅因未带头巾就遭受警察压迫至死的伊朗女性，从情理上说都应该更引起人们极度反感的情绪。于此相对，各大平台中的喜悦评论虽然占比小，但是是有一定比例的，除了少数确实表达心情愉悦的评论，还有一些奇特的评论 因为模型对自然语义识别性存在一定不足，导致也被收录到喜悦情绪中，比如“哈哈哈哈哈，这都给我录上了，置顶。”这种评论。

其次我们来看一下不同平台的比例差别，比如虎扑平台中表达愤怒的比例要高于另外两个平台 4-5 个百分点，而 B 站和知乎中语义分析被识别为喜悦的比例要远高于虎扑。这可能主要由平台差异化的评论语言组织形式和不涉及事件的噪音评论比例有关，也就说明了不同平台给予用户的不同内容呈现形式和不同评论形式会对事件的情绪比例产生影响，比如 B 站的弹幕区，还有 B 站对于事件的视频呈现不同于知乎对事件的文字呈现。

3. 情感时序分析——基于知乎平台的伊朗头巾评论数据

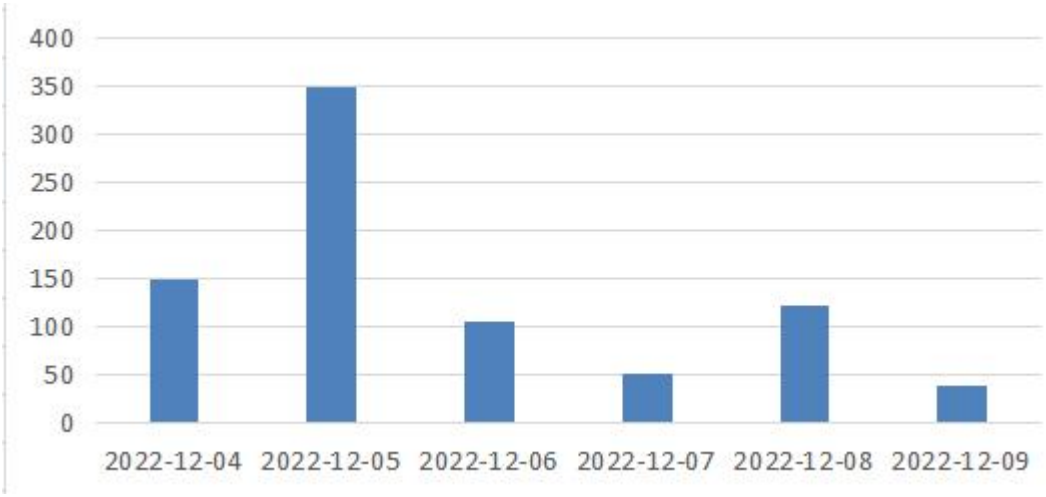
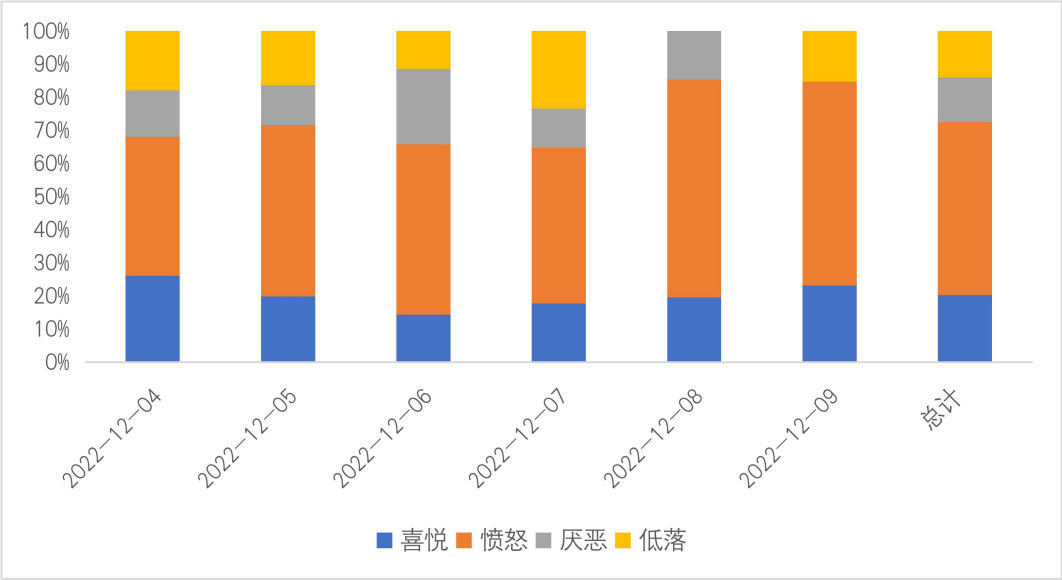


图 15:“知乎”数据事件分布

与前面主题模型分析时一样，知乎平台的评论数据具有较好的时序性，于是我们针对该部分数据进行情感分析。从评论时间来看，分布差异较大，可以观察到其有两个波峰，大波峰出现在 12 月 5 日，而后热度大幅下降，但 12 月 8 日又迎来一个上升。

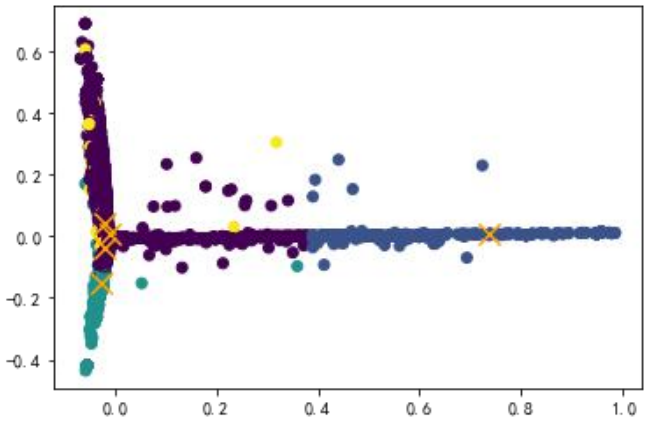


对各个时间段的评论进行情感识别，从各个时间段的情感分布变化来看，四种情感都是处在不断的波动中，并无较为明显的变化趋势，其中识别为愤怒的评论数量总体由少到多在12月8日达到峰值，可能与事件的关注度和发酵度随事件积累有关。

为了更好地展示该时间段知乎用户对于两个事件的评论内容，我们也绘制了依托情感识别结果的词云图，如下图用‘笑脸’图形表示的喜悦评论词云图和用‘火苗’表示的愤怒评论词云图。

六. 主题聚类

为了实现降低特征空间维度和加深语义理解、挖掘文本信息的目的，本项目对总体数据进行了 Kmeans 方法聚类，用主成分分析将数据降到二维，做出散点图将聚类结果可视化。



七. 总结

1. 通过上述分析，本报告得出结论：虽然两个事件本身存在差异，用户关注点有所不同，但是对于男女问题的观点，视频平台与非视频平台没有显著差异，主流用户观点都认为男女问题存在激化社会矛盾的嫌疑，并且讨论偏激。用户的关注重心在于维护社会团结，而并不仅仅是集中在问题本身进行讨论。
2. 研究的不足：
   a) 在数据爬取时没有获得非常理想的数据，比如平台数较少、时间标签数据较少，因此研究的方向受到了一定限制。
   b) 在进行 LDA 主题聚类时直接选择了主题数为 4，并没有针对这个超参数进行网格搜索。
   c) 由于 LDA 模型自身限制，聚类结果具有一定的随机性，而且对短文本的处理能力有限，导致在一定程度上无法提取出较稳定的主题。