面向新闻评论的情感分析

李锦阳，王雨秾，曾晓雨，武欣

基本介绍

随着“互联网+”时代的到来，人们获取信息的手段逐渐向移动互联网和社交网络方向倾斜，微博、微信，等社交工具蓬勃发展，而且网易新闻、腾讯新闻等相继推出手机新闻客户端。网络新闻逐渐成为人们获取新闻的主要方式，而**广大用户也积极通过网络评论来表达自己对新闻的观点和情绪**，为及时了解广大用户的情感偏向并分析舆情，**面向网络新闻领域的评论情感分析**具有重要意义。

2017 年 10 月 8 日，明星鹿晗和关晓彤通过微博公布恋情，引发了一场新浪微博用户们(尤其是女性用户)之间的轩然大波，相关的转发和评论甚至导致新浪微博瘫痪。

我们针对关于鹿晗关晓彤公布恋情的新闻评论，分别使用 lstm 方法和词典法分析网友们对于二人恋情的态度（支持与否）。

鹿晗关晓彤恋情新闻评论的赋分方法：

三种分值：-1，0，+1。

总体原则：是否支持鹿晗与关晓彤的恋情，支持为+1，不支持为-1。如果评论没有标明是否支持，认为与话题无关，记为 0 分。

根据话题的特殊性，以下是一些常见评论的赋分方式：

1. 与迪丽热巴有关的评论：心疼、抱走迪丽热巴等：-1。恭喜迪丽热巴（支持鹿关恋情，支持热巴退出话题）：1。
2. 吃瓜看戏等都是 0。
3. 骂网友，骂新浪等：0
4. 鹿晗不关心粉丝，脱粉等：-1
5. 同性恋，二人年龄差大，二人身高差等：-1

但是由于该新闻的特殊性，有很多评论无法明确划分分值，需要根据语境判断。例如：

欠我小姐姐一个道歉，凭什么我小姐姐提关挡了大半年的锅　-1

鹿晗和关晓彤公开恋情，跑男我也不需要看了。　0

估计又是为新戏炒作　0

别人的恋爱是别人的事，不要再喷了。没有谁配不上谁，只要两人愿意就行了呗。　1

最终结论：由于我们的标注方法与传统的情感极性分析略有差异，词典法效果不佳，lstm 方法效果更好。

组员分工：

武欣：数据爬取

曾晓雨：数据清洗整理
王雨秾：lstm 方法
李锦阳：词典法
数据标注：四人共同完成

Python 基于词典的分析

一、基于词典的方式通常用于文本情感极性分析

「情感极性分析」是对**带有感情色彩**的主观性文本进行分析、处理、归纳和推理的过程。按照处理文本的类别不同，可分为基于新闻评论的情感分析和基于产品评论的情感分析。其中，前者多用于舆情监控和信息预测，后者可帮助用户了解某一产品在大众心目中的口碑。
目前常见的情感极性分析方法主要是两种：基于情感词典的方法和基于机器学习的方法。

基于词典的情感分析常用方式是通过情感打分的方式进行文本情感极性判断。
给每一个词汇赋予一个情感分数，并对文本句子进行拆分，将句子中的每一个词汇的分值按照一定方式计算，得出整个句子的分值。score > 0 判断为正向情感，score < 0 判断为负向情感。

二、将基于词典的情感分析运用于本课题

1. 基本思路：

对评论进行分词，根据句子中的词语的分值来计算句子的得分，所有句子分值之和为评论最终得分。

2. 具体细节：

（1） 由于社交网络的特殊性，很多评论中的词汇在日常生活的词典中很少见到，需要手动添加。例如“祝 99”,“吃瓜”等。

（2） 由于话题的特殊性，在分词的词典中需要手动添加一些与话题相关的词汇，例如“迪丽热巴”“胖迪”“热巴”“鹿晗”“杨树苗”。

（3） 由于话题的特殊性，除过常见词汇的分值之外，需要给某些词语赋予特殊的分值。例如给“迪丽热巴”赋一个负数分值。

三、代码结构

1. 数据准备

分词：使用结巴分词 Jieba，外加手动添加的词典 userdict.txt，内含特殊词汇如“迪丽热巴”
停用词典：科院计算所中文自然语言处理开放平台发布了有 1208 个停用词的中文停用词表。
否定词词典：不、没、无、非、莫、弗、勿、毋、未、否、别、休、难道等。文件：notDict.txt

情感词典及对应分数：
词典来源于 BosonNLP 数据下载的情感词典，来源于社交媒体文本，所以词典适用于处理社交媒体的情感分析。同时手动添加一些该主题的特殊词汇。文件：sentiment.txt

使用爬虫程序得到的 10301 条新闻评论。文件 comments.xls

2 数据预处理

（1）分词

即将句子拆分为词语集合，结果如下：

e.g. 这样/的/酒店/配/这样/的/价格/还算/不错

Python 常用的分词工具：

结巴分词 Jieba

Pymmseg-cpp

Loso

Smallseg

此处使用结巴分词，并去除停用词

（2）构建模型

将词语分类（情感词，否定词）并记录其位置：

将句子中各类词分别存储并标注位置。

（3）计算句子得分

简化的情感分数计算逻辑：所有情感词语组的分数之和

定义一个情感词语组：两情感词之间的所有否定词与这两情感词中的后一情感词构成一个情感词组，即 notWords + sentiWords，例如不是很交好，其中不是为否定词，很为程度副词，此处被忽略，交好为情感词。假设交好的分数是 1.2，那么这个情感词语组的分数为：

Score = (-1) ^ 1 *1.2。

我们忽略了程度副词，因为我们只考虑正负性，（即是否支持），不考虑程度。

最终一个句子的得分是所有情感词组得分之和。

伪代码如下：

finalSentiScore = (-1) ^ (num of notWords) * sentiScore

finalScore = sum(finalSentiScore)

（4） 结果：使用±1，±2，±3，±4 作为分界的结果分别是：

#ratio1:0.5081553398058253

#ratio2:0.5213592233009708

#ratio3:0.5133980582524272

#ratio4: 0.502233

使用±2 效果更好一些，最终正确率约为 50%

四、代码调试，结果评价与反思

1. 情感词的分数大部分在-1 到 1 之间，极少数达到+10 或-10。分数绝对值不超过 10。

2. 通过观察部分评论发现了很多评论的内容集中于几个话题，例如关于迪丽热巴，关于鹿晗粉丝，关于网友多管闲事等。根据相关的娱乐新闻以及评论的得分规则，可以设置一些特殊的情感分数：

- 迪丽热巴，胖迪等关于迪丽热巴的词：-5。因为提到迪丽热巴的评论大部分是关于鹿晗和迪丽热巴二人前期的炒作，网友表示更喜欢鹿、迪 cp 组合，因此提到迪丽热巴的评论大部分是对鹿关恋情持否定态度。
- 关于李易峰：-5。提到李易峰的意思质疑是关晓彤曾说过的话，并表示对鹿关恋情的否定。
- 祝福，祝 99：10。很多评论前半部分表达自己作为粉丝的伤心，最后一句话送祝福，因此需要给祝福赋予一个很高的分值，才能让最后的总分为正。
- 大七岁，一米八等：-1。很多评论从年龄差，身高差出发表达对恋情的不满。
- 关于同性恋：-3。很多评论由此出发抨击二人恋情。

3. 词典法缺点 1：由于评论的得分规则是基于态度支持与不支持，与此无关的评论得分为 0。但是基于词典的分析方法以词汇的情感倾向为基础，很难分辨评论是否相关或无关。例如，有的评论涉及其他与此事无关的明星，如汪峰，易烊千玺；有的评论关于当时其他社会热点事件，例如红黄蓝幼儿园；还有的评论完全是商家广告，这些评论应该得零分。但是最终程序无法通过限制人名或者事件来区分出这样的无关评论。

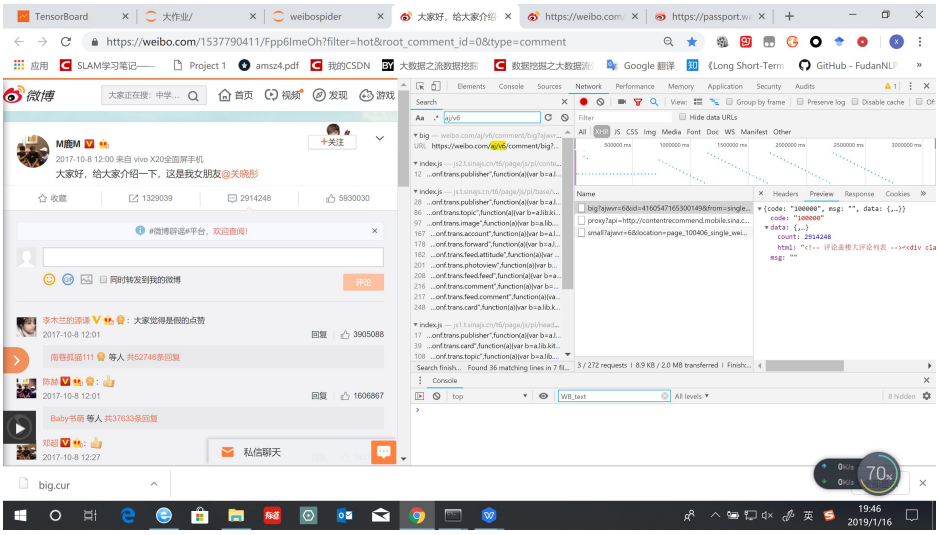
4. 词典法缺点 2：由于社交网络用语不同于生活用语或书面用语，再加上中文的博大精深，同义词汇在不同语境下意义可能完全不同。而且有些网络词汇表达的内容与事件背景有关，这给使用词典法带来了困难。例如：天长地久在生活中和书面用语中是表达祝福的词汇，但是在社交网络上，有时被用于负面的词汇，这取决于具体的语境。【欠我小姐姐一个道歉……】这个评论实际指的是与迪丽热巴的炒作，但是词典法无法识别，因为“小姐姐”指代对象无法提前预知。

数据爬取

目的：爬取微博网页端上的微博评论信息

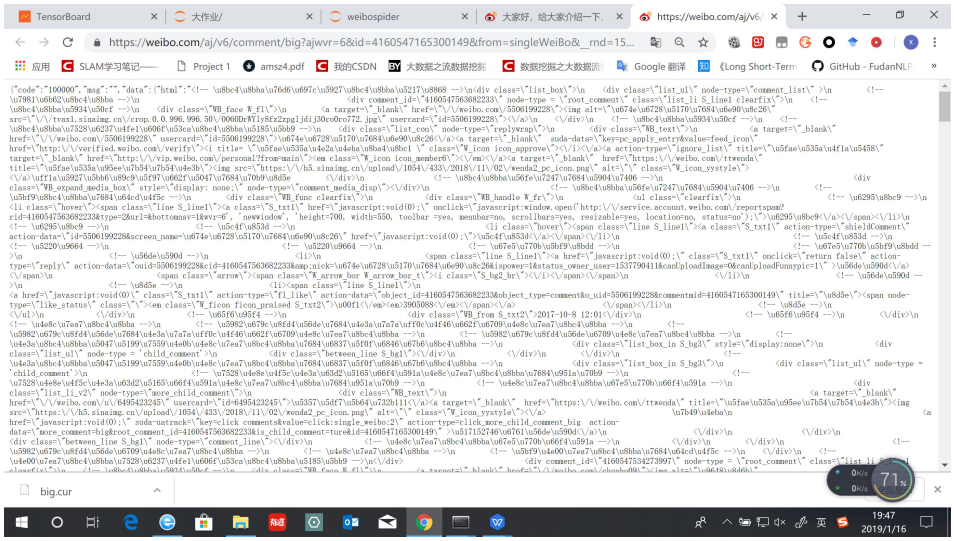
首先要模拟用户登录，即用 requests 发送 get 请求，传入带有 cookies 的 header 登录微博。cookies 就是自己微博登陆的 cookies，可以使用 chrome 开发者工具获得。

由于微博评论采用异步加载，必须点击或者下拉滚动条才会加载出更多的图片或视频，在源码中只有空白。为了提高用户的体验，许多网站都使用这种方法。究其根本，就是将这部分请求放在了后台。为了查看隐藏评论对应的 json 格式，按 F12 然后 F5 刷新页面，在开发者工具的 Network 标签下就会显示。根据 json 数据我们编写了自己的爬虫代码。



原评论 url:（鹿晗公布恋情微博）

https://weibo.com/1537790411/Fpp6ImeOh?filter=hot&root_comment_id=0&type=comment



异步加载 url:

https://weibo.com/aj/v6/comment/big?ajwvr=6&id=4160547165300149&from=singleWeiBo&__rnd=1547636454812

从原评论 url 中可以获得异步 url 中需要的 id,再按如下格式重建异步 url,调整 page 后面的值就可以获取微博每个页面的 json 数据。

```
url="https://weibo.com/aj/v6/comment/big?ajwvr=6"\
            +"&id="+str(self.id)\
            +"&page="+str(page) \
            + "&filter=hot" \
            +"&from=singleWeiBo"
```

我们注意到微博对字符串信息进行了 Base62 编码，所以在提取的时候要进行 Base62 解码。Base62 举例如下：

原文:但使龙城飞将在 不教胡马度阴山

Base62:LpaUdHtXIzYCI3huilByODItxjcJPHjKgoDgrV7VtukwXT9kKbscJHWmFV

解码后从 json 数据中提取评论，用 xpath 函数即可。注意首先要用 etree.HTML()将可能不完整的 json 数据构造成 XPath 解析对象。将获取的评论放入.txt 文件中，如果数据量太大还可以放入数据库中。

LSTM 方法

(一)、体系架构

要得出正向、负向、中性三个情感，需要用到三分类。我们采用典型的文本三分类方法 LSTM 建立模型，用标定的正向、负向、中性评论集合各 1 个，结合起来按 8：2 的比例随机划分训练集和测试集。

首先将语料库转换为词向量。要完成这一步，需要对语料进行分词，并建立词典。采用 jieba 分词工具对语料进行简单的分词，然后用 gensim 提供的 word2vec 训练词向量。

得到词向量之后，需要定义网络模型，进行 LSTM 训练。我们采用 Keras 提供的 Sequential 模型搭建网络。Sequencial 是一个简单的，单输入单输出的序贯模型，各个层之间呈相邻关系。需要加入嵌入层，将上一步得到的索引转换成词向量维度的向量。然后就可以一次加入 LSTM 循环层，Dropout 层，全连接层和激活层。之后就可以进行编译、训练、评估，最后将得到的网络模型转换成.yaml 文件保存起来。

(二)、技术路线选择

1.预处理

将评论整理成正向、中性、负向各一份，由于.xlsx 文件在读入中会出现问题，所以将它们转换成.csv 格式，采用 pandas 提供的函数 read_csv 读取语料。

读取后，用 numpy 提供的库函数 concatenate 分别整合三份评论及它们的标签，用 numpy 提供的函数 ones 和 zeros 可以很方便地得到标签矩阵。

之后采用 jieba 分词工具对整合之后的内容进行分词，定义了一个很简单的分词器，用

空格替换换行符。分词的模式采用了 jieba.lcut 默认的精确模型分词，因为全模型的分词方式会产生大量不必要的词语，不利于训练词向量。

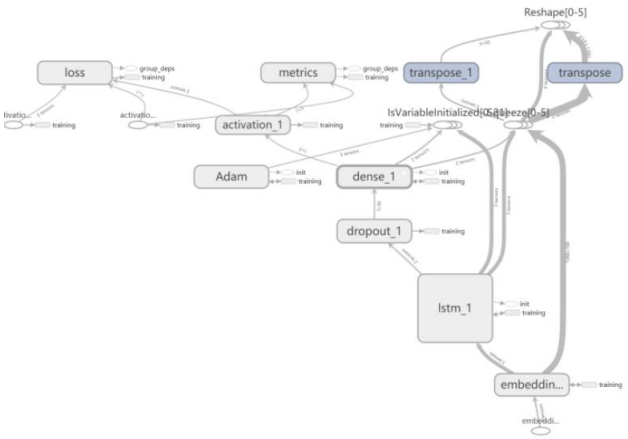
采用 gensim 提供的 word2vec 模型训练词向量，最小词频取为 10，小于 10 的词语可以忽略不计，对结果没什么影响。将训练好的词向量用.pkl 文件保存起来。

用得到的模型为整合文件创建词语字典，返回每个词语的索引、词向量，以及每个句子所对应的词语索引。使用 gensim 的 Dictionary 中的函数帮助创建词典，转为词典，得到单词与索引的对应关系和单词与词向量的对应关系，从而得到每个单词的索引和词向量，进一步计算出整个语料中每个句子所含词语对应的索引。

2.搭建网络模型

采用 Keras 提供的 Sequential 模型搭建网络。按照 8:2 对整个语料随机划分训练集和测试集。先加入一层嵌入层把索引转换成与词向量同等维度的向量。用 LSTM 作为循环层训练分类器。Dropout 层来防止过拟合，全连接层输出维度设为 3，激活层选用 softmax 作激活函数。用 add 函数将上述网络层加入 model 中，然后用 compile 函数设定损失函数为 categorical_crossentropy，优化方法为 adam，用 accuracy 作指标。

最后调用 fit 训练网络，用 evalueate 查看每一轮训练的 loss 和 accurracy。将训练完成的网络结构转换成.yaml 文件保存起来。



(三)、训练结果

用 tensorboard 查看了训练的结果，损失函数与准确率随轮数变化的曲线如下：

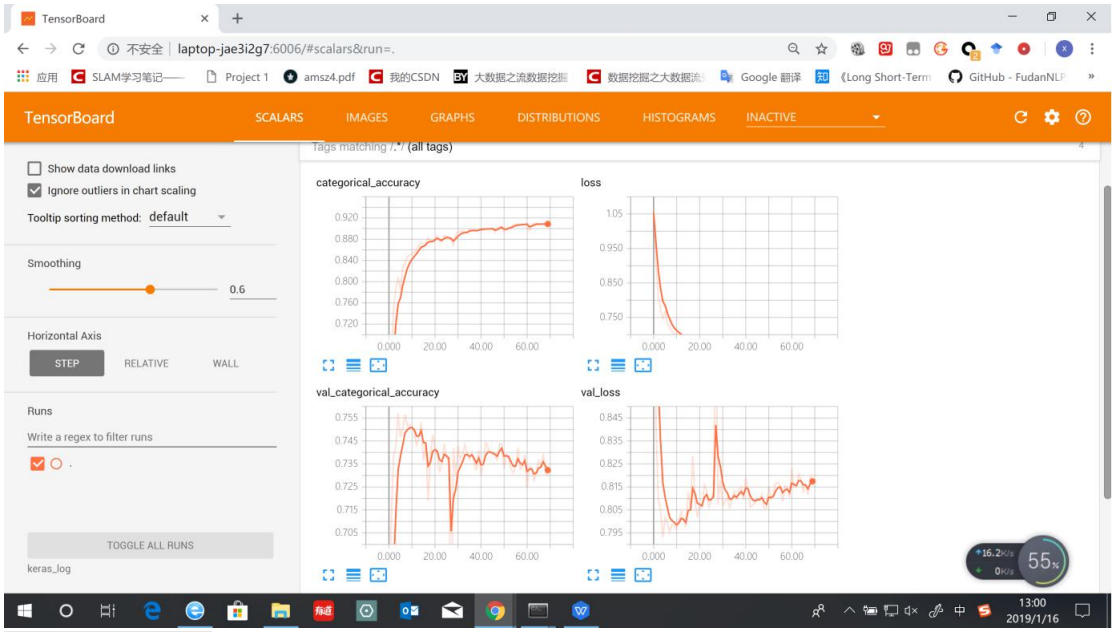


图 1：学习率固定

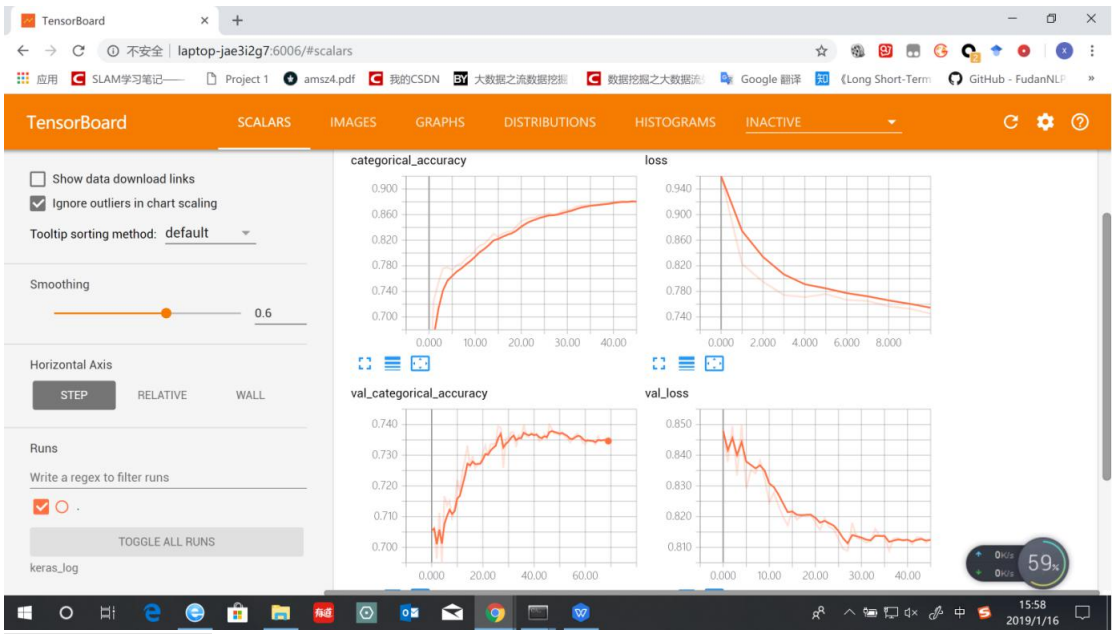


图 2：动态学习率

最终，我们采用了动态学习率，准确率为 73.99%。

对评论的分析结果

1.情感倾向随时间的变化

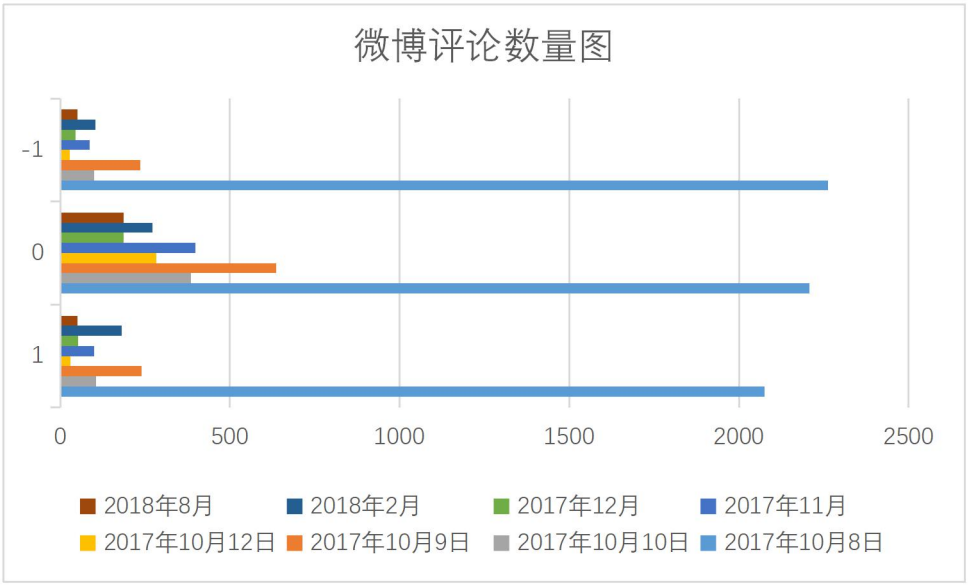


图 3：微博评论情感倾向的变化及评论数量分布

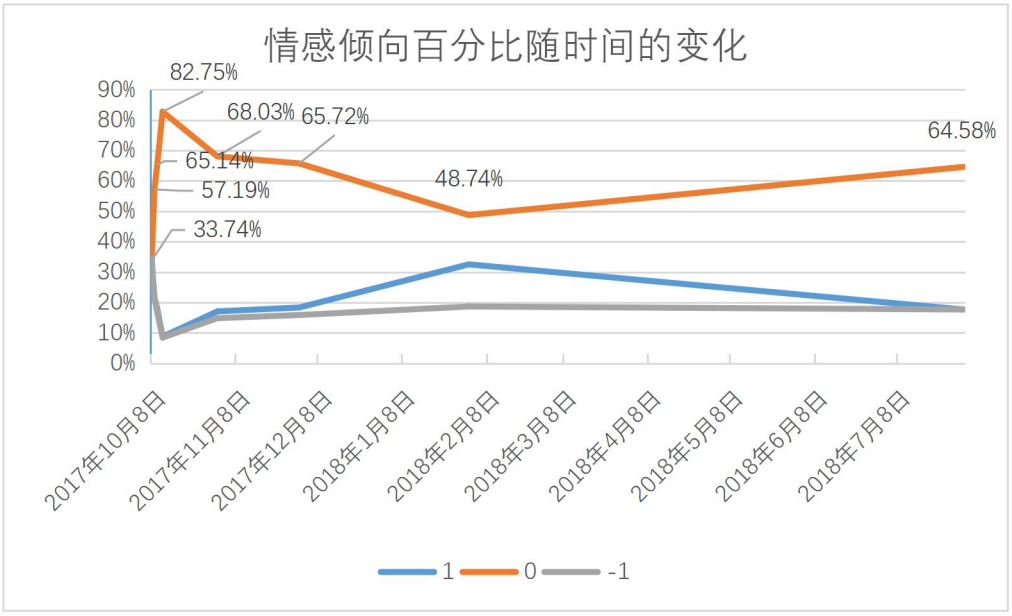


图 4：情感倾向所占占百分比岁时间的变化

2.网易、微博、腾讯的情感倾向对比

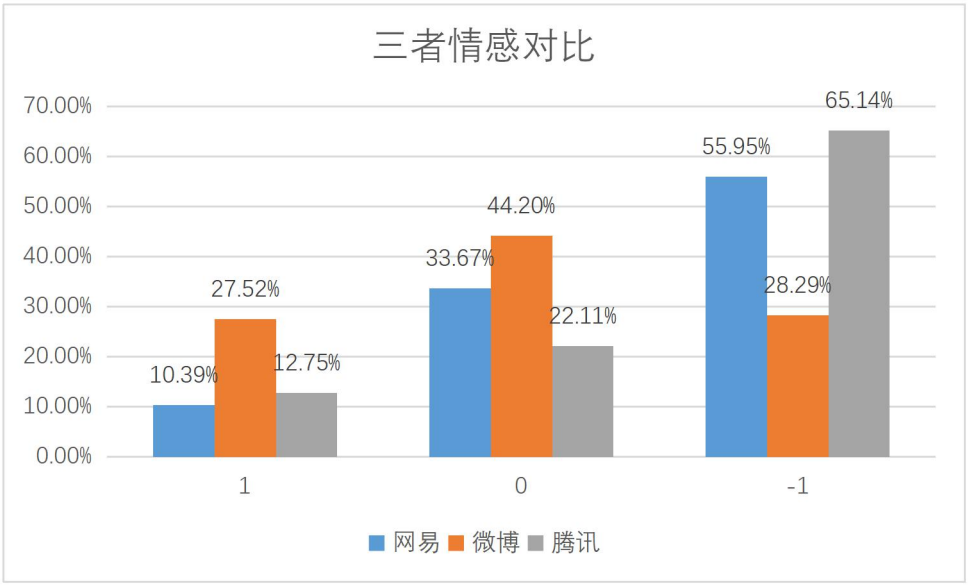


图 5：网易、微博、腾讯不同情感倾向的对比