@深圳天气微博分析

1. 基本情况

@深圳天气总发博量为46587条，粉丝数为1894191人。

通过对@深圳天气1-3月微博进行爬取，共获得微博1464条，平均每天发博16条，每条微博的平均点赞数为55个，平均转发10次，平均评论56条。

1. 内容分类

采用手工分类方式，对1464条微博进行了内容分类，其中气象预告预警类910条，科普内容21条、气象美景（视频、图片）20条、其他（搞笑、感悟、抽奖等）513条；原创1137条和转发327条

|  |  |  |
| --- | --- | --- |
| 类别 | 微博条数 | 占比 |
| 气象预告预警 | 910 | 62.15% |
| 科普内容 | 21 | 1.43% |
| 气象美景 | 20 | 1.37% |
| 其他 | 513 | 35.04% |

|  |  |  |
| --- | --- | --- |
| 类别 | 微博条数 | 占比 |
| 原创 | 1137 | 77.66% |
| 转发 | 327 | 22.34% |

1. 情感分类

将1464条微博进行分词处理，对分词结果进行分析，通过提供的情感词汇本体文件，将每条微博分词结果与情感词汇进行比对，如果分词结果出现在情感词汇中，则将该词语对应的情感分类保存下来到情感分析.txt。将情感分析.txt导入到excel中，对情感分类进行中文大类替换，如PA替换为乐，PD替换为好。替换完成后，每条微博有0~5个词有情感分类，取分类出现次数最多的为分类结果。分类结果如下：

|  |  |  |
| --- | --- | --- |
| 类别 | 微博条数 | 占比 |
| 好 | 564 | 38.52% |
| 乐 | 242 | 16.53% |
| 怒 | 4 | 0.27% |
| 哀 | 28 | 1..91% |
| 惧 | 36 | 2.46% |
| 恶 | 171 | 7.32% |
| 惊 | 6 | 0.4% |
| 无 | 413 | 28.21% |

4、语言风格

常用语气词有'的','了','么','呢','吧','啊'这几个，以及'！','？','。'判断句式类别，通过对每条微博的分析，得到文件：语气词和句式类别.xslx。从该文件中可得到以下结果：

|  |  |  |
| --- | --- | --- |
| 标点符号 | 句数 | 占比 |
| ！ | 206 | 12.26% |
| ？ | 216 | 12.86% |
| 。 | 966 | 57.50% |
| 总数 | 1442 | 85.83% |

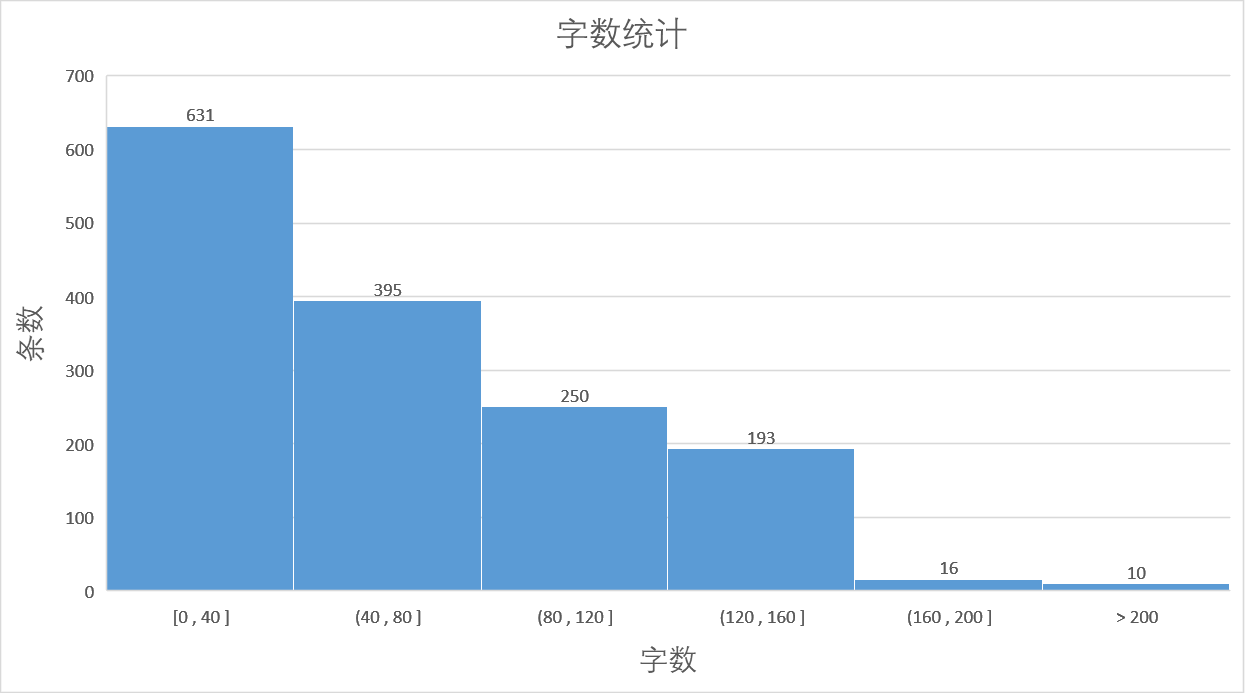
由于每条微博不一定只有一句话，所以上表数据中句数不等于微博数，占比也使用了在句数中的占比。总句数为1680句，其中有的句子没有出现上表中3种标点符号，出现这三个标点符号的句数为1442句，占总句数的85.83%。

|  |  |  |
| --- | --- | --- |
| 语气词 | 句数 | 占比 |
| 的 | 21 | 1.25% |
| 了 | 501 | 29.82% |
| 么 | 148 | 8.81% |
| 呢 | 48 | 2.86% |
| 吧 | 106 | 6.31% |
| 啊 | 44 | 2.62% |
| 总数 | 868 | 51.67% |

共有868句使用了语气词，其中语气词“的”只计算了句末出现的次数，句中出现的不算。

1. 长短句

以下是直方图统计：

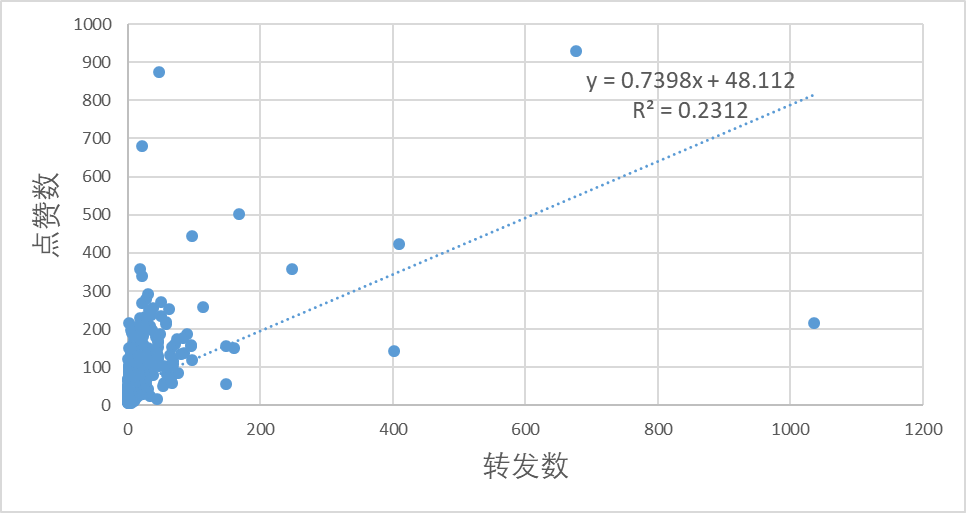


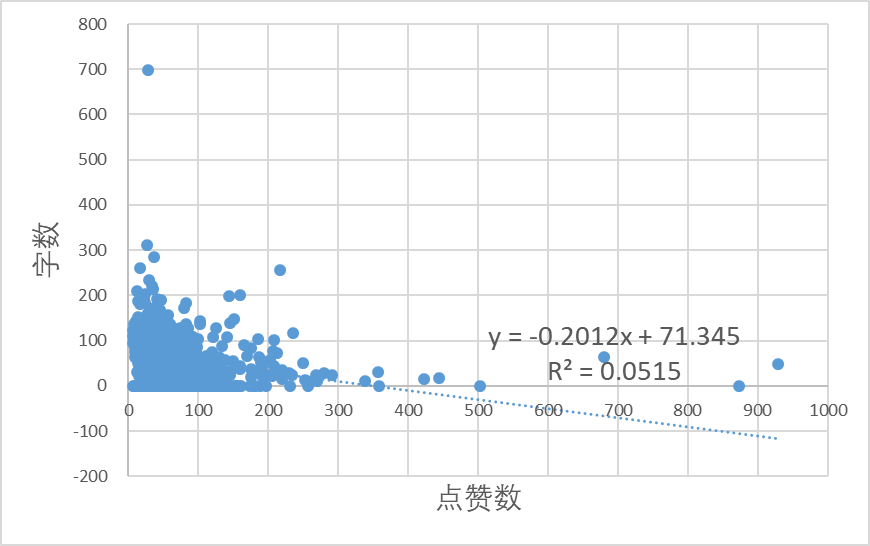
可以看出，40个字以内的微博最多，有631条。总体而言，字数越多的微博条数越少。

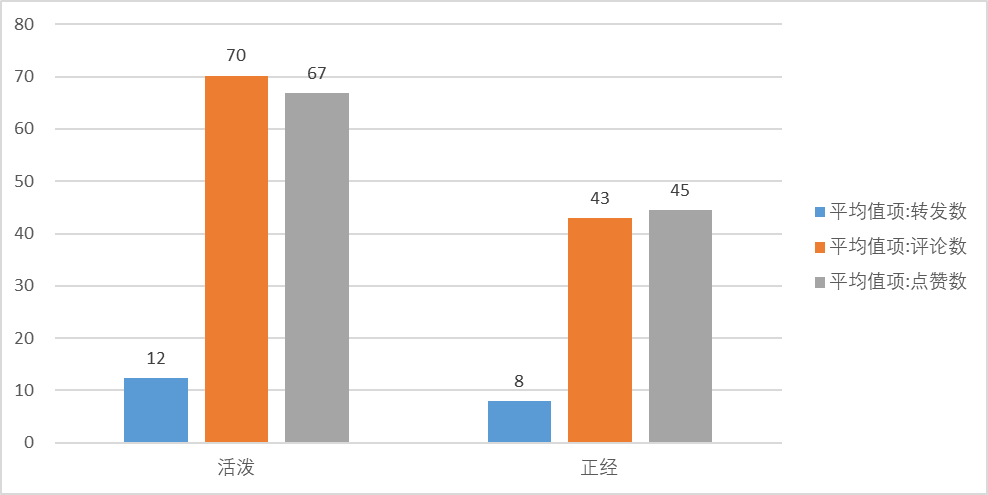
1. 人称

基本上为第一人称，1464条微博中有420条微博包含了“我”这个字，占比28.69%。

1. 相关性分析







将带有“？”“！”或语气词的微博分类为活泼，带有”。”和其他的分类为正经微博，活泼微博共726条，正经微博共738条。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类别 | 平均点赞数 | 平均转发数 | 平均评论数 | 微博条数 | 占比 |
| 正经 | 45 | 8 | 43 | 726 | 49.59% |
| 活泼 | 67 | 12 | 70 | 738 | 50.41% |