第四小组

微博转发点赞预测分析

指导教师：鲁凌云
项目组长：吴竞琦 17301169
项目成员：刘嘉欣 17301128
张雨桐 17301144
段怡冰 17301153
黎秀凤 17301156

项目报告修订版

1.项目背景

微博用户数量快速增长:数据显示，2018 年中国微博用户规模为 3.37 亿人，与 2017 年末相比增长 3456 万人，在整体网民数量中微博用户数比例达到 42.3%。，微博数据分析的需求呈现出快速上升的趋势,与此同时,市场是有效的数据分析结果显示表的极大的兴趣,也希望通过的结果显示,获取有效的商业或社交信息，从而实现更多的利益。

微博行为的目标包括营销、公关、客户服务、调研、产品曝光、活动发起、舆情监测等。事实上，实现这些目标需要大量的普通用户。因此，有必要对微博内容、微博用户、微博使用情况进行深入挖掘，实现定量和定性监测。现有的搜索引擎很难跟踪、分析和提取这些数据，因为它们太机械，本质上是信息检索。然而，跟踪微博上的各种互动，通过技术手段进行解释，将各种信息进行关联，然后给出结果作为决策依据，这与用户的心理有关。微博的数据分析结果主要提供给微博用户，用户可以通过微博数据的显示结果了解自己微博的发展变化。但是如何通过有效的设计来实现这样的数据表示，将对未来用户的关注和行为产生重要的影响。因此，运用科学有效的数据分析方法，以及更加直观人性化的设计展示就显得尤为重要。

2.业务分析

2.1 社会公共事件舆论分析

通过公共事件的‘关键词’词汇，如“肺炎、口罩、武汉、地震”等等，获取全平台热门微博，从中分析事件发展轨迹、舆论焦点、挖掘消息来源、识别中心传播者/机构以及风险预警通知。

- 采集范围：微博疫情“肺炎”关键词的微博内容
- 应用领域：高校课题研究《关于社会公共事件的微博舆论导向》

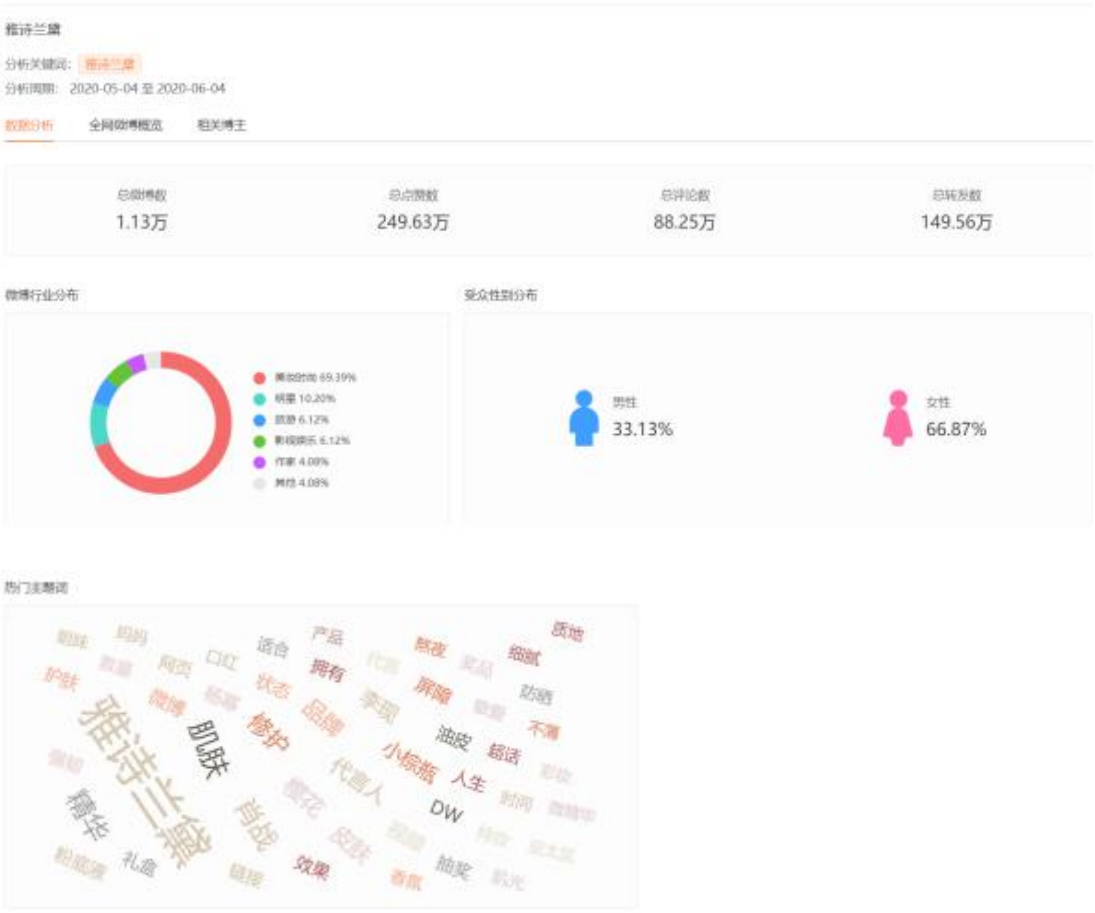
微博热点词云图

以乐清顺风车司机杀人案进行更直观地分析，通过博文转发评论点赞数据分析，捕捉公众对此类事件关注的焦点。

2.2 品牌舆情舆情监控

——以品牌雅诗兰黛 2020-05-04 至 2020-06-04 与品牌维多利亚的秘密 20020-04-22 至 2020-05-20 的微博数据为例进行监控分析

- 采集范围：微博“品牌名称”关键词下的网友评论
- 应用领域：品牌口碑情感倾向与消费者满意度/好感度评估



维多利亚的秘密

分析关键词：维多利亚的秘密

分析周期：2020-04-22 至 2020-05-20

数据分析　全网微博概览　相关博主

总微博数	总点赞数	总评论数	总转发数
1,006	81.26万	13.41万	60.19万

微博行业分布

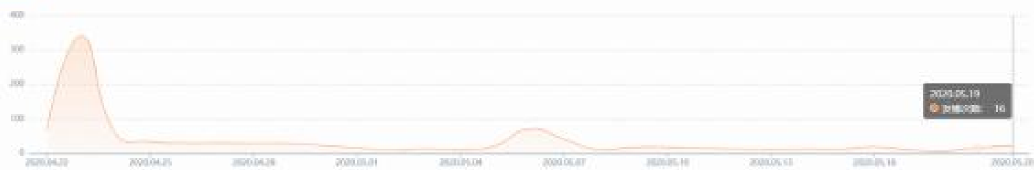
受众性别分布

男性 42.85%

女性 57.15%

微博走势

发博次数　点赞　转发　评论



内容分析

TOP10博主

排行	博主	类目	粉丝数	发博数	平均评论数	[illegible]	[illegible]	操作
01	杨幂	明星	10766.29万	4,096	21.94万	1,935.06	2,650.13	详情
02	时尚芭莎	美妆时尚	1793.11万	3.03万	1,344	1,105.72	1,700.62	详情
03	MadameFigaro中文版	影视娱乐	421.95万	4,424	1,774	1,075.56	1,790.92	详情
04	程铭玥YvonneChing	美妆时尚	611.80万	807	853	919.57	1,097.84	详情
05	梨视频	资讯	2245.36万	6.48万	321	814.27	1,154.11	详情

2.3 市场调研与营销咨询

疫情期间，泡面市场又重新复苏。通过搜索“泡面”关键词获取所有和泡面有关微博评论以及微博用户的基本信息（地区、年龄、职业等）。再通过用户画像统计与人工智能文本分析，描绘疫情期间购买泡面的用户人群、用户情感的正负面分析、关注焦点等。

- 采集范围：根据客户的品牌词、关联品类词采集微博内容与用户信息
- 应用领域：品牌形象调研、消费者市场喜好调研、用户画像制定

2.4 时事热点追踪、信息源获取

采集汇总微博各大官方权威机构发布的公告通知、最新最热的时事事件、焦点人物动态跟踪等，将信息源同步至企业内部平台进行标签筛选或深度加工，再通过企业 APP 或专题页面分发至对应的人群。

- 采集范围：“卫健委”系列、“央媒地方媒体”系列官方最新微博动态
- 应用领域：丰富企业自己的新闻平台的内容来源，保证对时事热点的实时动态追踪、对数据源的深度加工形成专题页报道等

2.5 广告投放、营销

2020 年 2 月 26 日—中国领先的社交媒体微博公司（NASDAQ：WB）公布截至 2019 年 12 月 31 日的第四季度及全年未经审计的财务报告。

“2019 年我们在用户和商业化方面继续保持增长。”微博首席执行官王高飞表示，“微博用户规模和活跃度的强劲提升，主要得益于我们针对产品的优化和内容消费体验的提升。在商业化方面，品牌广告收入在 2019 年依旧保持稳健增长，得益于整体流量的强劲增长和微博品牌广告的独特价值。2020 年我们将继续完善平台生态系统，进一步强化核心竞争力。在商业化方面，继续专注于帮助广告主提升在微博的营销价值的同时拓展多元化变现模式，以实现未来可持续化增长。”

2019 年全年企业业绩：2019 年全年净营收 17.7 亿美元，较 2018 年的 17.2 亿美元增长 3%。:2019 年全年广告和营销营收 15.3 亿美元，较 2018 年的 15.0 亿美元增长 2%。2019 年全年来自大客户和中小企业的广告和营销营收为 14.3 亿美元，较 2018 年的 13.8 亿美元增长 4%。2019 年全年微博增值服务营收为 2.367 亿美元，较 2018 年的 2.193 亿美元增长 8%，其增长主要得益于 2018 年第四季度收购的直播业务所产生的营收。

以上数据材料不难看出微博广告营销给资方带来的巨大利益，特别的 2020 年再度刮起一股大风的带货直播，微博自然也想要分一笔羹。

3.可行性分析

3.1 文化政策可行性

• 文化背景

当前社会发展情况下，互联网的使用越来越频繁，随着微博的出现和发展，越来越多的人熟练使用微博，并成为稳定用户，对微博的依赖程度较高。因此社会的发展趋势以及用户的微博使用情况使得分析微博转发点赞预测有了可能性。

• 社会背景

微博对人们生活的影响逐渐增大，从公共事件舆论、品牌舆情监控、市场调研与营销咨询、时事热点追踪、信息源获取，到广告投放、营销等，微博都发挥了不可忽视的作用。因此对微博转发点赞的分析预测是非常有意义的。

3.2 技术可行性

数据：从网上查找数据，得到较为准确的微博转发点赞原始数据。通过数据清洗，将无效数据进行剔除，获得有代表性的有效数据。之后提取特征值，用于下一步的算法分析。

算法：可以 KNN、Kmeans 等多种算法，从多个角度进行分析数据，可以更加全面的对数据进行分析，得出更加准确的结论。

3.3 时间可行性

项目的开始时间为 2020 年 5 月 4 日，在经历了小组成员选题、了解项目、查找资料资源等过程后，正式实施开始进行的时间为 5 月 21 日，项目交付时间为 6 月 8 日，有充足的时间进行项目实践。

3.4 经济可行性

项目的可持续性，本项目的成功预测，不论是公共事件舆论、品牌舆情监控还是市场调研与营销咨询、广告投放、营销，都是很重要的，更好的时间点、更有趣的方式等等，都能带来更好的效益。

3.5 对人和社会的影响

对国家社会来说，更好的了解当前有效的舆情传播方式有利于了解民众的想法，对社会的发展有很大的帮助。对想要做推广的公司来说，进一步了解微博转发点赞等方面，有利于寻找更合适的方法进行推广宣传，让更多的人了解本公司，从而能够达到公司推广的目的，将公司的利益最大化。

4.制定需求分析框架和分析计划

4.1 分析框架

• 目标变量的定义：用户发布微博后，一周之内得到的转发量、点赞量

分析思路

通过搭建 KNN 分类模型，根据训练集中每一条数据的特征值（发帖时间、内容关键字等）与目标值（发布后一周的转发量、点赞量）的对应关系。输入没有目标值的新数据后，将新数据的每个特征与样本集中数据对应的特征进行比较，根据距离公式求出二者之间的距离， 然后根据距离大小进行排序，提取样本集中特征最相似数据（最近邻）的分类标签。 一般来说，我们只选择样本数据集中前 k 个最相似的数据，这就是 K 近邻算法中 k 的出处，通常 k 是不大于 20 的整数。 最后，选择 k 个最相似数据中出现次数最多的分类，作为新数据的分类来比较准确且有效地来预测微博发布后一周的转发量和点赞量。

• 分析样本的数据抽取规则：

我们采用天池新浪微博互动预测比赛提供的数据集。数据集来源于从 2014 年 7 月 1 日到 2014 年 12 月 31 日的新浪微博数据，包括发布时间，微博内容、发布后一周的转发量、评论量、和点赞量，为了保护用户隐私，用户 ID 和微博 ID 均已加密。

• 潜在分析变量：

发布时间，可从中提取月份、星期、小时等特征。

内容，即发布微博内容，可从中提取关键字进行分析。

• 项目风险及应对策略：

如果训练的模型效果不好，出现欠拟合的现象，则对数据做进一步的处理，挖掘未提取的特征，再进行训练。如果 KNN 分类算法被证明不可行，则通过阅读文献、请教老师、小组研讨来找到新思路。

• 项目的落地应用价值分析和展望

模型投入应用后提前预测微博引流能力，从而可以使运营方有针对性地开展推广、服务等运营工作；可以将建模过程中发现的有价值的、最可能影响微博流量的重要字段和指标选择性地提供给运营方，用于制定运营方案和策略的依据和参考；针对影响微博流量的核心指标和字段，可以提供给相关业务方，以作为提高微博运营盈利的依据和参考线索。

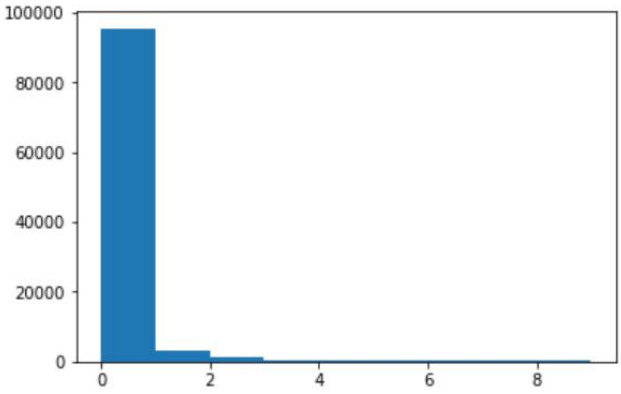
4.2 分析计划

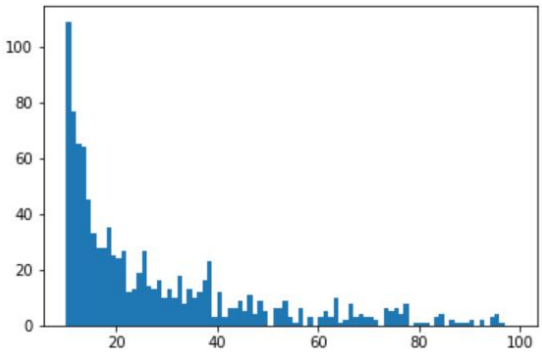
时间	分析进度
5 月 4 日——5 月 14 日	数据的抽取和摸底阶段
5 月 15 日——5 月 22 日	数据的前期分析阶段
5 月 23 日——5 月 31 日	建模阶段
6 月 1 日——6 月 5 日	模型验证评估，撰写分析总结和运营方案建议

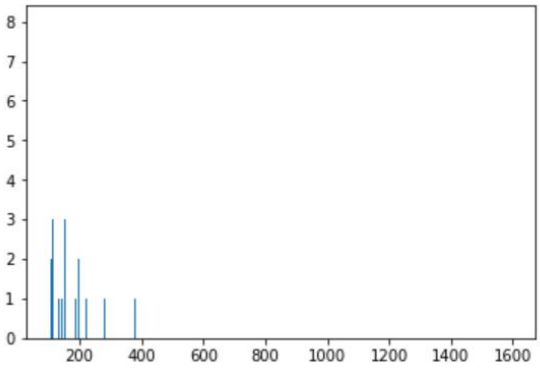
5.抽取样本数据、熟悉数据、数据清洗和摸底

（1）目标变量分布情况：

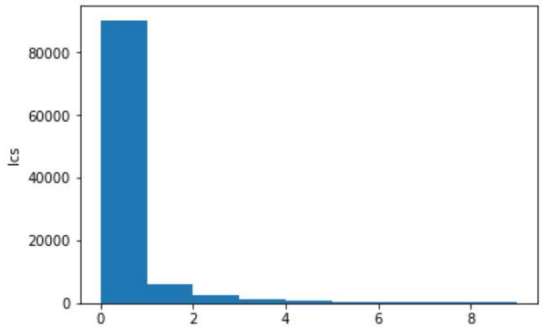
• 转发量分布情况：大多数微博转发量为 0，只有一小部分微博转发量突破 10，突破 100 只有零星的几个。

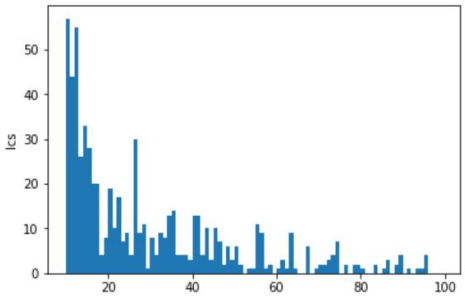


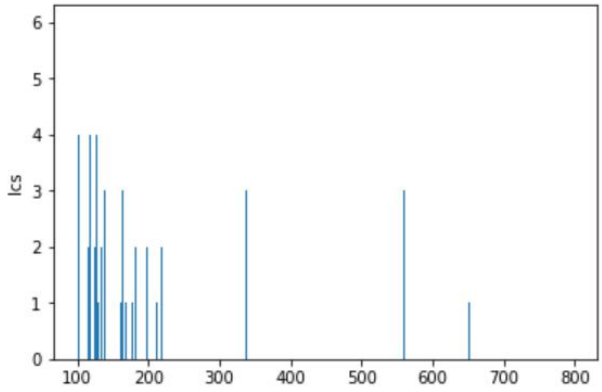




• 点赞量分布情况：大多数微博点赞量为 0，一部分微博转发量在 10-100，突破 100 只有零星的几个。







（2）划分训练集，测试集

- 训练集，有： 1225088 行 7 列
- 测试集，有： 245018 行 7 列

（3）提取一天中的发帖时间，范围 0-23 如：

原来的时间格式：2015-08-19 22：44：55 转化成：21

（4）提取内容关键字并编号：

使用结巴分词' jieba.analyse.extract_tags() '提取每条微博的关键词，并对其进行编号，范围 0-184806

tag	fid
韩国菜	0
洋葱味	1
萨金	2
新班咯	3
乌山	4
受不得	5
曹辉	6
了图	7
板需	8
唯心论	9
氮化	10

6.按计划初步搭建挖掘模型

6.1 进一步筛选模型的输入变量

通过调研并查阅《基于能量优化的微博用户转发行为预测》等相关论文，我们发现一条微博的转发量和点赞量很大程度上受发布时间和内容影响，于是我们初步选用一天中的发布时间（0-23）和微博内容的关键字编号(0-184806)作为模型的输入变量。

6.2 尝试不同的挖掘算法和分析方法

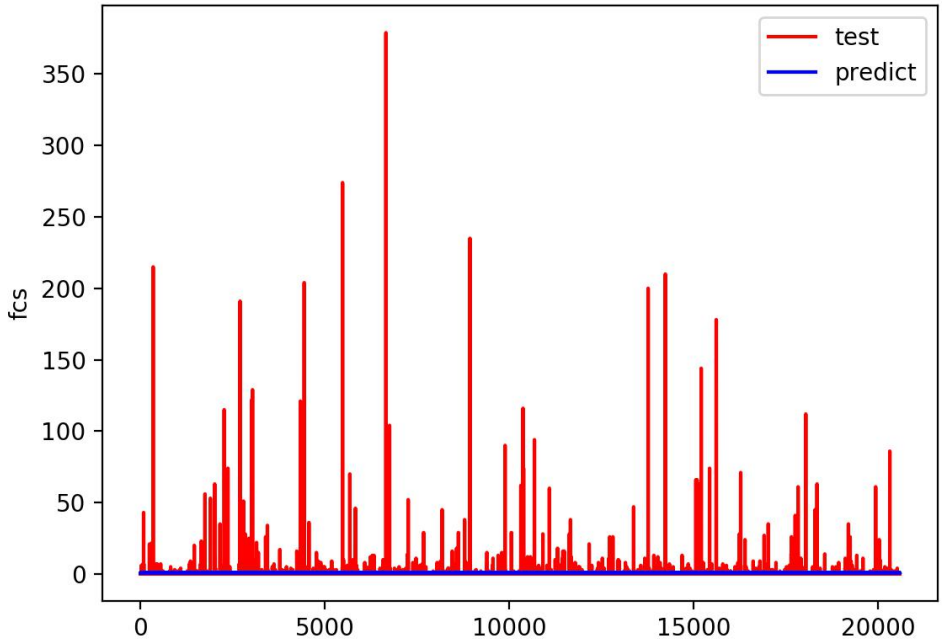
我们分别用多元线性回归、K-means、KNN 算法对数据进行挖掘分析。

• 多元回归分析

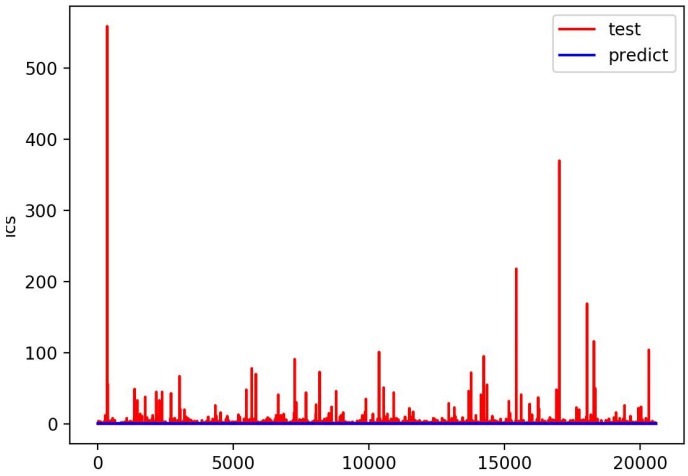
使用多元线性回归算法进行分析时，输入变量发布时间（tid）和微博内容的关键字编号(fid)，调用的第三方库 sklearn 的 LinearRegression。

根据预测值画出的拟合曲线与实际曲线相差甚远，欠拟合现象特别严重，由此我们判断不适合采用线性模型进行分析。同时，我们从下图可以看出数据有聚类现象，因此下一步，我们尝试使用 K-means 算法进行分析。

对转发量进行分析的大的拟合曲线：

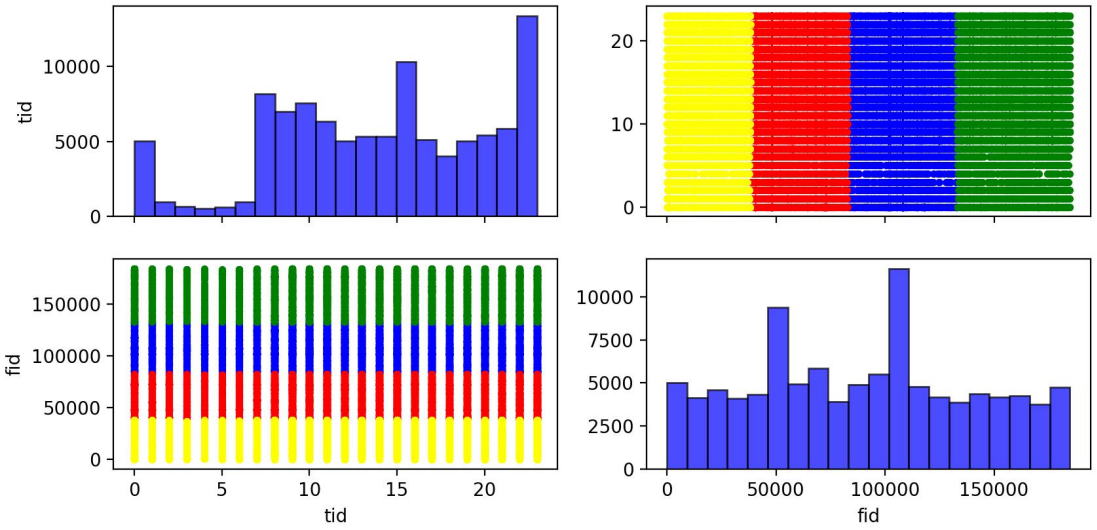


对点赞量进行分析得到的拟合曲线：



• **K-means 算法分析**

使用 K-means 算法进行分析时，输入变量发布时间（tid）和微博内容的关键字编号(fid)，调用的第三方库 sklearn 的 K-means，设定 K=4。得到下图结果，由图可以得出，模型的分析效果不错，聚类结果主要受 fid 的影响。

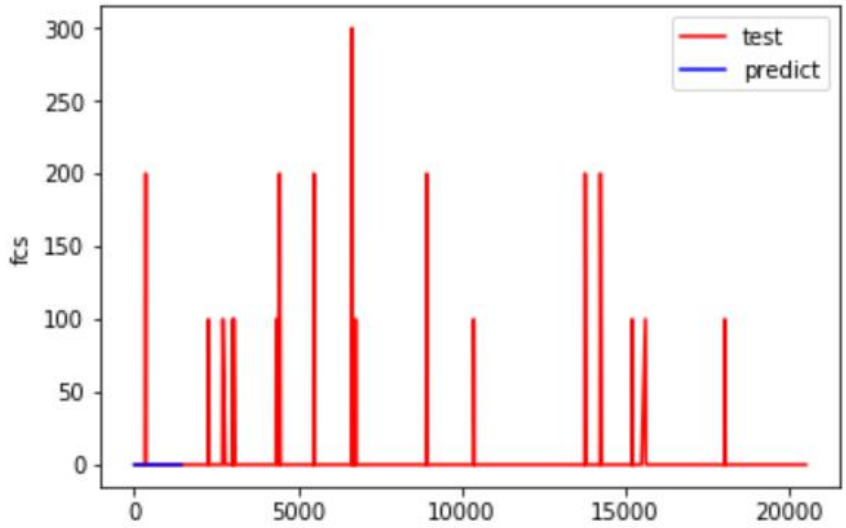


• **KNN 算法分析**

使用 KNN 算法对转发量进行分析时，输入变量发布时间（tid）和微博内容的关键字编号(fid)，调用的第三方库 sklearn 的 KNeighborsClassifier，默认 n_neighbors=5。

将训练数据的转发和 100 整除，得到 10 个类别[0，100，200，300 400，500，600，1000，1400，1500]，用 Sklearn 的 model.score 评估模型得到评分为 0.9983。

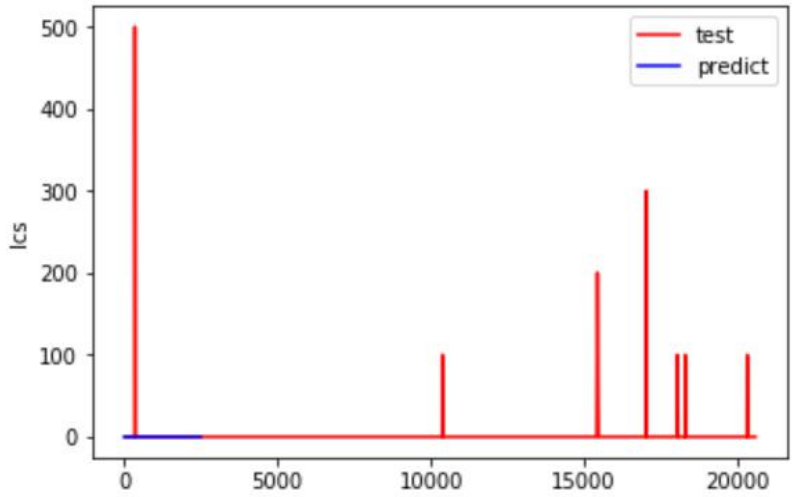
考虑到大部分数据转发量都为 0，我们将测试数据中转发量为 0 的数据去掉后，重新测试，得到的模型评分为 0.9769。评分虽然不低，但从下图预测值和真实值的对比中，我们发现，模型对转发量的预测过于悲观。



使用 KNN 算法对点赞量进行分析时，输入变量发布时间（tid）和微博内容的关键字编号(fid)，调用的第三方库 sklearn 的 KNeighborsClassifier，默认 n_neighbors=5。

将训练数据的转发和 100 整除，得到 7 个类别[0，100，200，300，500，600，700]，用 Sklearn 的 model.score 评估模型得到评分为 0.9995。

考虑到大部分数据点赞量都为 0，我们将测试数据中点赞量为 0 的数据去掉后，重新测试，得到的模型评分为 0.995。评分虽然不低，但从下图预测值和真实值的对比中，我们发现，模型对点赞的预测过于悲观。



7.讨论模型的初步结论，提出新的思路和模型优化方案

- 模型中已考虑的自变量

名称	含义	范围
tid	一天中发帖时间	0-23
fid	内容关键字编号	0-184806

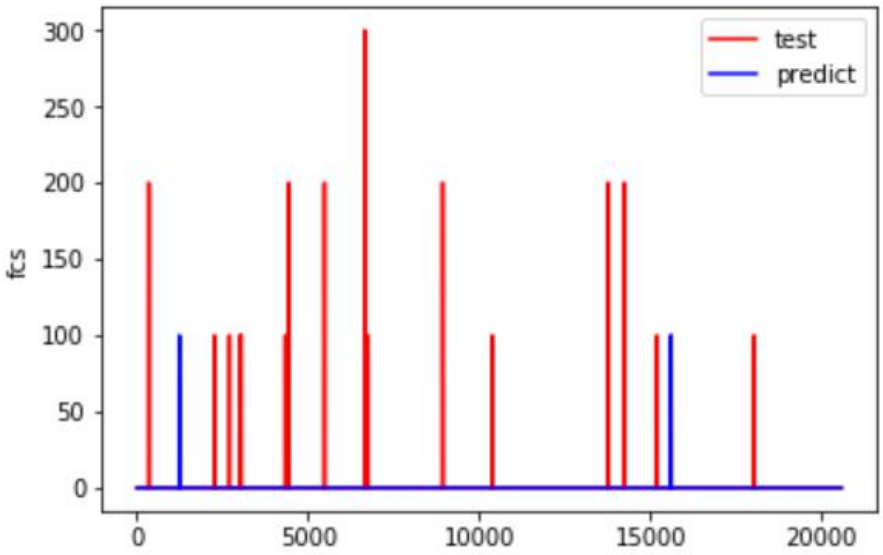
• 优化方案
进一步从数据中提取的特征：对用户重新编号，发帖的月份，周几发帖。

名称	含义	范围
uid	对用户重新编号	0-21112
month	发帖的月份	0-11
wday	周几发帖	0-6

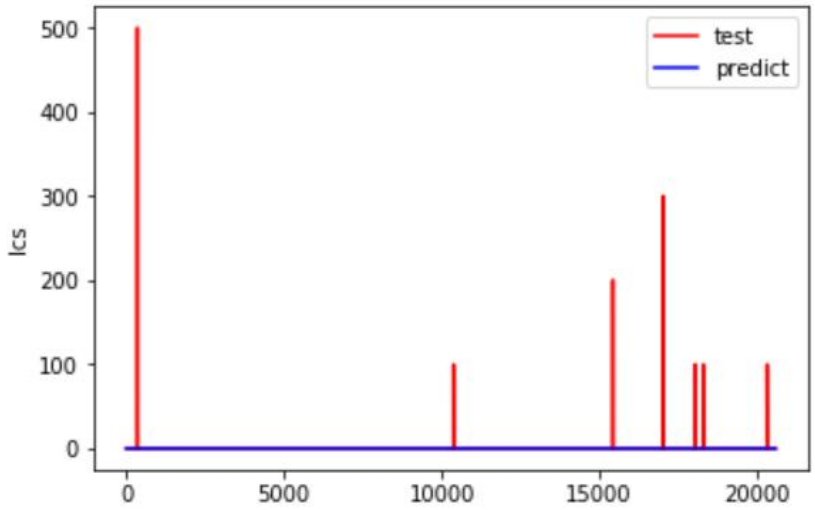
清理僵尸用户数据：清除发布多条微博的转发量、评论量、点赞量都为 0 的用户。

8.按优化方案重新抽取样本并建模，提炼结论并验证模型

重新抽取样本后，还是用原来的 KNN 模型进行训练，得到的模型评分虽然没有显著提高，但在对转发量的预测方面，悲观倾向有所改善。
转发量预测结果



点赞量预测结果



结论：

KNN 算法在对微博的转发量、点赞量的预测上表现良好，但对高质量微博的捕捉的灵敏度有所欠缺，后续将继续挖掘高质量微博的特征，并在这一方面对算法做进一步改进。

9.落地应用建议

9.1 品牌数据监控

通过对进行推广的博主广告微博转赞评互动数据进行分析，分析该博主推广商业价值。

9.2 广告投放

通过对目标博文现有数据或同类信息进行转赞评分析，分析感兴趣的人群，通过年龄、性别、地域、兴趣、是否 wifi、安卓还是 IOS 等多个维度选择投放用户，比如你是一个做经期管理的 app，可以投放给北上广、年龄在 17-45 岁的女性用户，推广的这些条件广泛而精准，所以投放效果也大大提升。

通过高精度的推广能够实现：1.海量覆盖，微博庞大的活跃用户量，保证了推广目标客户的广泛度。2.黄金的广告位，与很多广告联盟的广告出现在侧边栏不同，将推广直接出现在微博的信息流中，提高触达效率。

*参考文献

基于能量优化的微博用户转发行为预测_ 王伟 张效尉 任国恒 秦东霞 刘琳琳

面向不平衡微博数据集的转发行为预测方法_赵煜

基于随机森林的微博互动特征分析_于澍

基于 Python 的新浪微博用户数据采集与分析_高雅 苏艳 席方园

融合多种转发习惯的微博转发预测_徐建民 韩康康 何丹丹 吴树芳

基于大数据分析方法的微博热点建模与预测_王 哲 刘贵容 彭润亚