上 海 财 经 大 学

毕 业 论 文

题目 基于 GBDT 的电子游戏评论情感分析研究

姓　　名 徐晓杰

学　　号 2016110715

学　　院 统计与管理学院

专　　业 应用统计学

指导教师 王刚

定稿日期　　2020 年 5 月 15 日

基于 GBDT 的电子游戏评论情感分析研究

摘　要

电子游戏作为彰显国家软实力的载体之一，是近年来增长迅速的一大产业。在一款游戏的生命周期中，游戏本身是需要不断打磨的。玩家评论中包含的情感倾向是今后改进的重要方向。正向情感指出了开发者需要保持的地方，负向情感指出了开发者需要改进的地方。

本文基于 Taptap 平台游戏评论数据，运用 TF-IDF 方法进行分词并收集评论字数、手机型号等额外特征，建立了决策树情感倾向分析模型。而后利用 Boosting 集成学习思想，构建了 GBDT 模型。结果证明 GBDT 模型获得约 83% 的准确率，F1 值达到约 81%，显著高于决策树模型。此外，本文还对有额外特征与无额外特征的模型进行了比较。结果证明，额外特征对模型预测能力的提升具有显著效果。

关键词: 电子游戏，用户评论，情感分析，机器学习，集成学习，决策树

Sentiment Analysis of Game Reviews Based on GBDT

Abstract

As one kind of media reflecting domestic soft power, video games have become one of the fastest growing industries in China. During the life cycle of one video game, the game itself needs to be continuously updated and improved. Players' sentiment in their reviews leads the further improvement. Positive sentiment points out what developers need to persist, while negative sentiment points out what they need to make better.

Based on reviews data in Taptap, using TF-IDF for word segmentation and collecting additional features such as phone type on each review, this paper develops a decision tree model to classify players' sentiment inclination. Furthermore, based on boosting ensemble learning ideas, this paper develops a gradient boosting decision tree (GBDT) model. The result shows GBDT model reaches 83% accuracy and 81% F1-score, obviously better than the decision tree model. In addition, this paper compares models with additional features and models without additional features. It turns out that additional features contribute a lot to the performance of the models.

Keywords :video games, game reviews, sentiment analysis, machine learning, ensemble learning, decision tree

目录

1 前言

1.1 研究目的与意义

随着近年国民经济的快速发展，人民物质生活水平大幅度提高，精神和文化生活也不断丰富。电子游戏逐渐成为人们消遣娱乐、放松心情、丰富生活的新兴方式。电子游戏作为我国新崛起的一大文化产业，不仅带来了巨量的经济效益，也为我国文化软实力的提升提供了巨大的力量。

自 2014 年 1 月，国务院办公厅发布关于上海自贸区的通知，我国长达 13 年的游戏机禁令正式解除。据资料显示，游戏机禁令仅 2013 年造成的直接损失就在 830 亿元以上。而禁令的解除无疑为中国电子游戏市场带来了空前的机遇。2019 年中国游戏市场实际销售收入 2308.8 亿元，同比增长 7.7%。观察历史数据，可以发现，自 2014 年以来，中国游戏市场收入平均年增长率超过 20%，每年的增长率超过 5%。（见图 1.1）

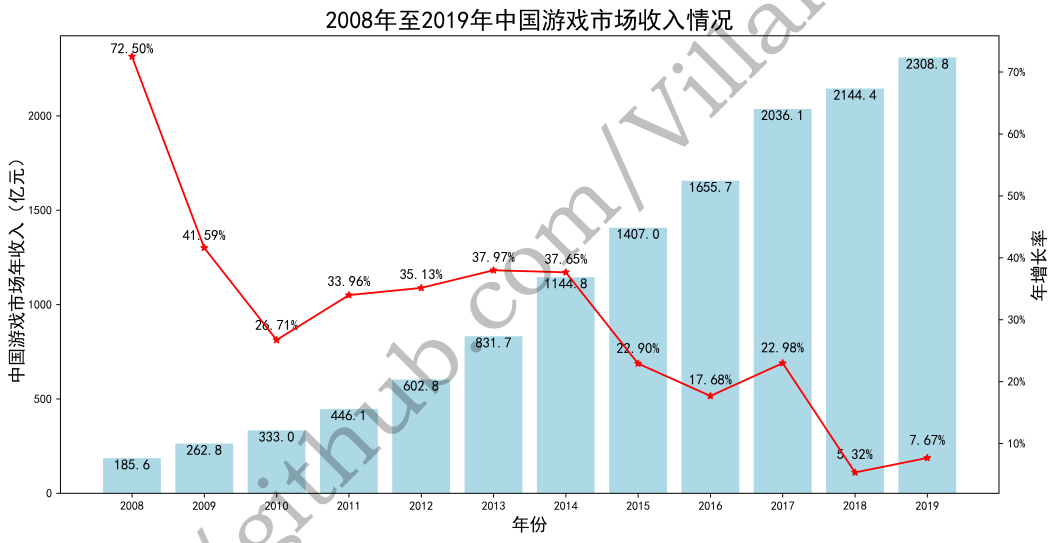

图 1.1: 2008 年至 2019 年中国游戏市场收入情况

电子游戏的广阔市场激起了广大中国开发者的创作激情。2014 年以来出现了许多优秀的国产电子游戏，受到国内外玩家的好评。例如，2019 年，网易游戏和故宫博物院合作开发的一款《绘真·妙笔千山》手机游戏，以传世名画《千里江山图》为蓝本，讲述了一个中国背景的故事，游戏自上线以来获得多个国家和地区的应用商店推荐；又如，2017 年，腾讯旗下北极光工作室自研的《无限法则》收获了电子游戏权威评分机构 IGN8.5 的高分。禁令解除后的种种成果都证明了电子游戏对我国文化软实力提升的显著效果。

然而，由于我国电子游戏产业起步较晚，且有一段较长的停滞期，开发过程并不一帆风顺。当游戏开发初步完成并上市后，其本身还有一些地方值得打磨，而开发者、发行商由于自身的视角问题，可能并不能发现。这时，玩家评论就显得尤为重要。

1) 有的玩家对游戏爱不释手，称赞有加，如“作为国内的一款音游我觉得 OK！别看评论了兄弟萌赶快买 GKD!!!! 早买早开心!!!!”。这无疑会间接带动新的潜在玩家。

2) 有的玩家则会对游戏时产生的一些问题而感到不满意，如“曲风太单一了，感觉大部分都是节奏很快的电音可爱风……”；再如“发现了两个 bug，一个是时不时的闪退一下，这很烦人，另

一个是暂停键不能离开有反应……”。如果开发者不及时予以重视，这些玩家便很有可能会流失。

不难看出，评论中往往带有玩家自身的情感倾向，有的积极，有的消极。如上文所说，不同的情感倾向会产生不同的影响。本文尝试对电子游戏评论数据进行情感分类，准确把握玩家对于游戏的情感倾向，以帮助开发者及时对游戏做出调整与完善，使我国电子游戏的改进革新之路更具效率。

1.2 领域研究状况

1.2.1 情感分析发展历史

情感分析 (Sentiment Analysis)，又称意见挖掘 (Opinion Mining)，是自然语言处理 (NLP) 的研究领域之一。此概念最早分别由 Nasukawa 等人与 Dave 等人于 2003 年同年提出。顾名思义，情感分析的目标在于分析给定文档中所蕴含的用户情感。而文本情感分类作为情感分析最直接的一项任务，需要判断给定文档的情感倾向，如积极、消极等。

早在 2002 年，虽然尚未对情感分析进行定义，Bo Pang 等人使用朴素贝叶斯，最大熵分类和支持向量机三种传统机器学习算法，采用一元分词模型 (unigrams) 和二元分词模型 (bigrams) 作为特征，对电影评论情感进行分类，取得较为良好的结果。作者认为，相比于人类定制规则自行判断，标准机器学习模型在情感分类任务上的表现更为优秀。然而限于当时网络并不发达，数据有限，且计算机算力存在瓶颈，该领域并没有得到大规模的研究。

直到近年来随着社交网络的发展与大数据的兴起，情感分析这一领域开始得到重视，越来越多学者将情感分析应用于不同领域并提出新颖的方法。2010 年，Alexander Pak 等以社交平台推特 (twitter) 的文章作为语料库，运用支持向量机 (SVM)、条件随机场 (CRF)、朴素贝叶斯三种传统机器学习模型，同样采用 n-gram 模型用于构造特征，构建了一个判断文档情感极性的三分类情感分类器（积极、消极、中立）。2013 年，Richard Socher 与 Christopher Potts 等人提出了循环神经网络模型 (RNN)，使用深度学习方式对文档情感倾向进行判断。2015 年 Kai Sheng Tai 等人将句法结构元素加入长短期记忆网络模型 (LSTM) 中，在句法分析的结果上进行语义组合，在句子级情感分类任务上取得了良好的效果。

目前文本情感分类研究方法主要有两类。一类是基于情感词典，一类是基于机器学习方法。

基于情感词典方法，根据构建好的情感词典，在对文本进行分词后筛选抽取情感词，而后根据情感词典计算该文本的情感倾向。中文开源情感词典较为有名的有知网 Hownet 情感词典，台湾大学 ntusd 情感词典。该方法的效果取决于词典的有效性。然而中文语义博大精深，上述两款词典尽管收录了一部分有效的情感词，但每个领域都有属于自身的特殊词汇，例如在网络领域，除了考虑传统的情感词语外，还需考虑许多流行语。故使用此方法进行分类仍需要对所研究领域的术语、特定语等词汇有一定了解。基于机器学习方法，运用词模型 (n-grams，TF-IDF，word2vec 等) 对原文本进行特征构造，将其输入机器学习模型进行训练。

1.2.2 电子游戏领域与情感分析的结合

在上述诸多学术成果的支持下，人们进一步将情感分析与各个领域相结合，产生了许多商业应用。主要分为：商品服务评论分析、社交网络分析、情感机器人。

将情感分析与电子游戏领域结合，正是属于商品服务评论分析应用。其目的在于挖掘顾客与用户的游戏需求与期望，并使游戏公司与开发者能从先前的成功抑或失败中获取经验。

2016 年 Dorinela 等人以亚马逊网站上 9500 条游戏评论为语料，运用词数指示器、主成分分析、多元方差分析，得到了 5 个对用户评论情感倾向解释占比极大的变量，用于解决游戏评论积极、消极、中立的三分类问题，其 DFA 分类器的准确率达到了 55%。

2017 年 3 月 Kiran 等人以推特上的电子游戏最新评论为语料，训练支持向量机，贝叶斯模型，最大熵模型，对评论进行积极、消极、中立三分类判别。同年 Rohan 以 steam 评论数据及评论附带的其他元数据作为输入，训练朴素贝叶斯、支持向量机、逻辑回归与无监督方法相结合的情感倾向二分类器。同年 4 月 Bjorn 以两款已发行电子游戏《Dragon Age》与《Mass Effect》的用户评论为语料，通过词频分析抽取出频率较高的词，并对这些词进行了基于方面（aspect）的情感分析，发现评论的整体情感倾向与这些方面的情感倾向有着强关联性。

1.3 分析方法与流程

本文通过分析游戏评论数据，提出了基于决策树的有附加特征与无附加特征的情感倾向预测模型，之后使用基于 Boosting 的 GBDT 集成学习模型进行改良。结果证明附加特征能使模型对用于情感倾向的预测能力得到提升，集成模型的预测能力明显优于单一模型。本文的研究流程如图 1.2所示。

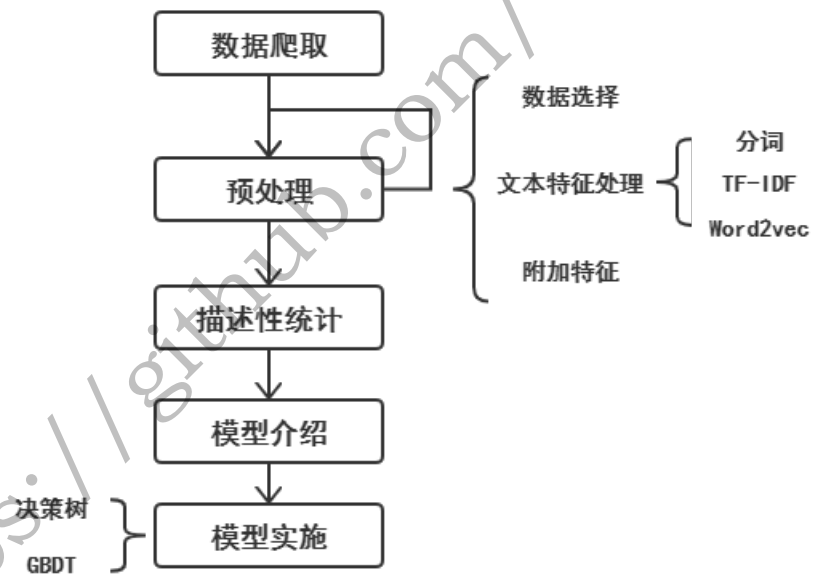

图 1.2: 本文研究流程

1.4 本文创新点

在国外，大多游戏领域与情感分析结合的研究者都采用传统统计学习算法，所得到模型精确度并不高。在国内，该领域的研究并不多。因此，在中文游戏评论情感分析领域有许多内容值得研究。

1）本文在使用决策树模型的基础上，运用梯度提升的集成思想，进一步对模型进行强化。在不失模型解释性的同时保证了模型的性能。

2）本文通过使用 word2vec 方法，基于原始语料构建了电子游戏领域中文情感词典。

3）本文在使用评论本身的同时，加入其他相关信息作为附加特征进行训练，最后结果说明，这些附加特征对判断玩家情感倾向的准确度具有一定提升效果。

1.5 论文基本框架

本文基于 python 爬虫程序爬取近 1300 款手机单机游戏的评论数据，利用决策树模型以及基于梯度提升的决策树集成模型对用户评论进行情感分析。通过特征筛选与参数调试，得到了效果较为良好的情感倾向预测模型。

本文分为如下部分：

第 1 章为引言，介绍研究目的与意义、领域研究现状、研究方法总体框架及创新点。

第 2 章介绍本文的变量设计及预处理过程。

第 3 章展示预处理后数据的描述性统计。

第 4 章介绍本文使用的相关机器学习模型理论，以及集成方法思想。

第 5 章介绍模型训练情况与预测效果，结果表明该模型具备较强的预测能力，且模型本身具备良好的解释性。

第 6 章为总结与展望，总结本文，并对后续改进工作进行展望。

2 数据获取与预处理过程

2.1 数据获取

TapTap 由易玩（上海）网络科技有限公司开发，是主要以广告为收益的正版游戏交流社区。该平台聚集了大量国内移动游戏开发者与用户群体，实时同步全球各大平台的游戏排行，能与全球玩家共同探讨游戏内容与质量。

TapTap 平台具有基数广大的优质用户，据 2018 年 5 月统计，TapTap 活跃用户占游戏平台类应用第二位，数量达到 738.7 万，活跃人数行业渗透率为 20.8%，仅次于 4399 游戏盒，是一个为开发者与玩家、玩家与玩家之间提供高品质游戏交流、分享、下载的游戏社区。各玩家的评论，也对游戏日后的品质改进起到了推波助澜的作用。

本文以单机移动游戏为研究目标，使用 python 程序爬取 TapTap 平台上近 1300 款游戏，共计近 140 万条评论，出于对数据应用目的的考虑，设计爬虫程序时制定了如下策略：

1）限于硬件条件，本课题仅考虑爬取“单机”类型的手机游戏作为研究。

2）爬取评论时，由于近期评论倾向于反应最新的问题，故优先收集近期的评论。

2.2 爬取数据变量设计

通过 python 程序获取的原始数据情况如表 2.1，为方便下文的说明，该表将部分原始变量分为“文本特征”与“附加特征”两类。

表 2.1: 原始变量说明

名称	说明	数据类型	含有空值	备注
game_id	游戏编号	整数	否	同 user_id 构成主键
user_id	用户编号	整数	否	同 game_id 构成主键
issue_time	评论发表时间	日期	否	
content	评论内容	字符串	否	文本特征
content_len	评论字数	整数	否	附加特征
phone_type	手机型号	字符串	是	用户手机具体型号，附加特征
fun	对此评论表示有趣的人数	整数	是	附加特征
up	赞同此评论的人数	整数	是	附加特征
down	不赞同此评论的人数	整数	是	附加特征
play_time	游玩时间	整数	是	以分钟为单位，附加特征
user_score	用户对游戏的评分	整数	否	共一星至五星五类

2.3 数据选择

在日常生活中，我们往往会发现许多满分好评下写着类似“五星好评不解释”、“习惯五星好评”等评论内容，而且诸如此类的满分评论不在少数。虽然不能断言所有满分评论都是如此，但可以确定的是，在满分评论论中，确实有许多评论是用户在无意识，无明显情感倾向的情况下所发表的。这样的数据对情感倾向的分类会产生影响。

而当用户未给出满分评价时，可以推测此时用户的情感倾向是更加明确的，评论内容是更加具有意识的。例如：

- 当用户给出四星评论时，用户倾向于具有“虽然有一些小瑕疵，但是总体上较为满意”的正向情感。
- 当用户给出一星、二星评论时，用户的倾向于具有“问题比较多，总体上并不能满意”的负向情感。

在上述假设下，对数据进行筛选，取出一星、二星的数据标记为负类 (0 类)，取出四星的数据标记为正类 (1 类)。最终得到近 33 万条数据，正负类比为 2:1。

2.4 文本特征的预处理

2.4.1 分词

国内有许多优秀的分词工具。本文使用北京大学的 pkuseg 工具对评论内容进行分词。pkuseg 是基于 CRF 模型，辅以 ADF 训练方法的分词工具，在各个开源数据集合上具有良好的分词结果。初步完成分词后，抽样查看样本分词结果，记录未能正确分割的词语，产生本文的领域词表。在建立领域词表后，将该词表与停用词表一同导入 pkuseg 工具，对原始文本特征重新分词。

2.4.2 TF-IDF 方法

众所周知，统计模型只能处理数字，故分词结果尚不能直接作为模型的输入。对于如何将词数字化，一种直观的想法便是对分词结果中出现的每一个词构造一维特征，将每条评论表示为由各个词出现频数所构成的向量，这便是词袋模型。该想法很直观，但其缺陷也很明显，因为该模型仅仅考虑词频，而没有考虑上下文的关系、全文的语境，因此会丢失一部分文本的语义。

TF-IDF(Term Frequency-Inverse Document Frequency) 模型作为词袋模型的改良。该方法为每一个词计算 TF-IDF 值，在一定程度上弥补了词袋模型的缺陷。TF-IDF 是一种用于资讯检索与文本挖掘的常用加权技术。作为一种统计方法，TF-IDF 值评估单个词对于一个语料库中的一条语料的重要程度。单词的重要性随其在语料中出现的次数成正比增加，但同时会随其在语料库中出现的频率成反比下降。

以“游戏剧情不错，游戏玩法也不错”作为单条语料为例子。分割结果为“游戏 | 剧情 | 不错 |，| 游戏 | 玩法 | 也 | 不错”。“游戏”一词在该评论中出现了两次，频率较高，而由于本文研究领域是游戏领域，故“游戏”一词在其他评论中也会高频出现，据此认定“游戏”一词对该评论的重要性并不高。“不错”一词在该评论中也出现了两次，同“游戏”相同，但是表达评价的词有许多例如“很好”、“垃圾”、“赞”等，由此可以推断“不错”一词在其他评论中出现的频率比“游戏”低，说明“不错”一词相比“游戏”更具识别度，据此认定“不错”一词对该该评论的重要性要比“游戏”高。

通过上面的例子，不难推知 TF-IDF 的主要思想：如果某个词或短语在一条评论中出现的频率高，并且在其他评论中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

TF-IDF 的计算方式如式 2.1、式 2.2、式 2.3所示。

$$TF(w) = freq(sample_w) \tag{2.1}$$

$$IDF(w) = log\frac{N}{N(w)} \tag{2.2}$$

$$TF - IDF(w) = TF(w) * IDF(w) \tag{2.3}$$

其中 $freq(sample_w)$ 表示词 w 在评论 $sample$ 中的频数，N 表示评论总量，$N(w)$ 表示所有评论中包含词 w 的评论数量。

本文使用 TF-IDF 方法对分词数据进行数字化处理。

2.5 领域情感词典的建立

2.5.1 必要性

电子游戏领域是一个小众领域。在该领域中，存在许多特殊用语，类似于“行业黑话”。例如“祖安人”一词，普通人可能会认为该词指的是来自某个地域的人群，然而在电子游戏语境下，该词特指那些在游戏中喜欢嘲讽他人的玩家。又如“告辞”、“呵呵”等词，在电子游戏语境中含有厌恶、反感的情绪。有的词本身是从该领域中新产生的。有的词在普通语境与电子游戏语境中有着截然不同的意思。

上述的这些词包含玩家的主观情感，对玩家情感倾向的判定具有重要的作用。而现有的中文情感词库并不适用于该领域。因此，有必要构建针对电子游戏领域的中文情感词库。

2.5.2 Word2vec 方法

如果从分词结果中逐个地去寻找相关的情感词语，其人工成本是高昂的。为降低工作量，可以作如下设想：

先在分词结果中人工寻找一部分典型的领域情感词语，再利用某种方法让机器寻找出与这些词语意思相近的词，这些新的词同样表示了用户的某种情感。

使用什么方法？本文认为使用 Word2vec 方法是一种比较好的选择。

Word2vec 方法是一种词的分布式表示方法，旨在将每个词通过人工神经网络映射成低维稠密向量。下设两个模型，分别为：

- 跳字模型 (skip-gram)：根据当前词语最大化上下文词语出现的概率。
- 连续词袋模型 (CBOW)：根据上下文词语最大化当前词语出现的概率。

Word2vec 的核心思想是：在一句句子中，一个词的意义是由其周围的词决定的 (CBOW)，而周围的词的意义也是由这个词决定的 (skip-gram)。基于这种“两个词上下文相似，则语义也倾向于相似”的想法，Word2vec 词向量可以很好地表达不同词之间相似与类比的关系。因此，使用 Word2vec 方法训练词向量，而后根据余弦相似度筛选出与典型情感词相似的词，便可以构成领域情感词典。

作为示例，表 2.2展示了分别与“真香”、“辣鸡”最相近的 3 个词语。

表 2.2: word2vec 所得情感词语实例

真香	辣鸡
rua	垃圾
美妙	烂
追定	恶心人

最后，本文构建了由 535 个积极情感词与 569 个消极情感词所组成的领域情感词典。

2.6 附加特征的预处理

附加特征共 6 个，分别为 content_len、phone_type、fun、up、down、play_time。首先使用零值对空值进行填补，而后对含有不规则值、异常值的样本进行筛除。

特别地，对于特征 phone_type，首先统计出手机的主流品牌，而后使用对应的代号进行替代，手机品牌及对应代号如表 2.3所示。

表 2.3: 手机品牌代号说明

phone_code	phone_type
0	未填写
1	小米
2	OPPO
3	华为
4	Vivo
5	一加
6	三星
7	魅族
8	乐视
9	iPhone
10	其他

3 描述性统计

图 3.1为本文获取评论数据的时间分布。评论时间范围在 2016 年 4 月至 2020 年 1 月。其中 2016 年至 2017 年数据总量相较于 2018 年至 2020 年的数据总量低。这与本文优先选取近期数据的爬取策略相一致。从图中可以看出，每年 12 月至次年 1 月，以及每年 6 月至 8 月的评论数量都比其他月份高。这也可以理解，因为该时间段是人们放假休息的时期，所以游玩电子游戏以放松心情的人会较其他月份的人更多。

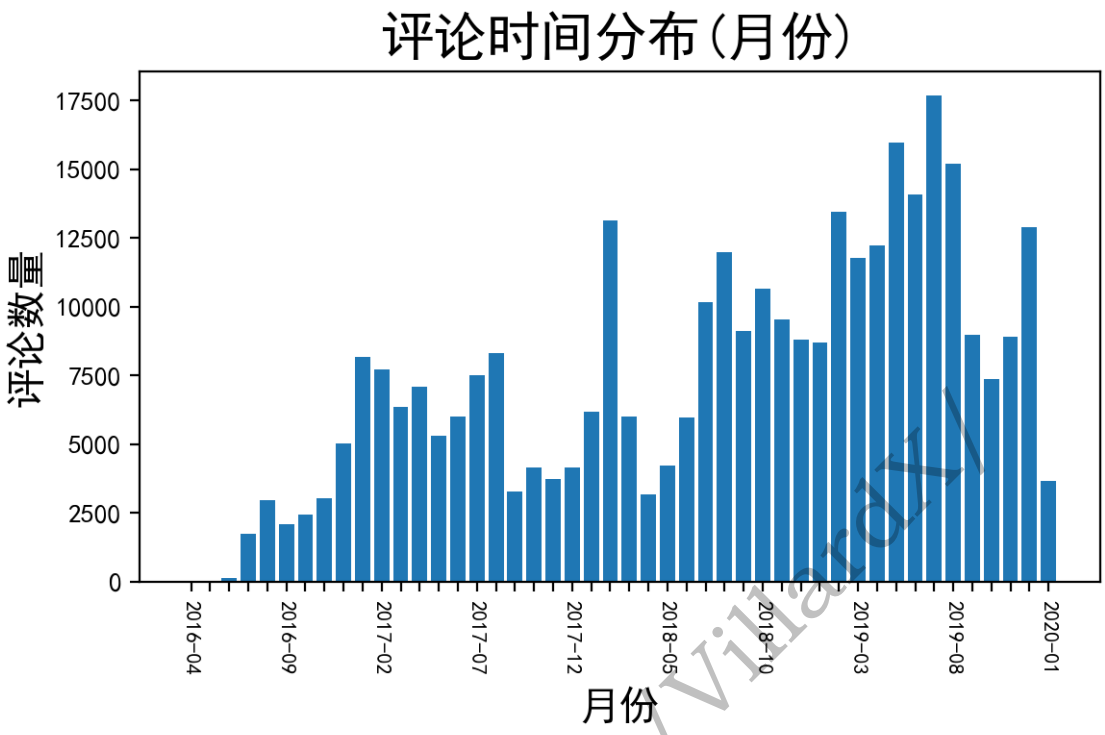

图 3.1: 评论时间分布（月份）

表 3.1展示了互动特征 up、down、fun 的零值占比。可以看出，各个附加特征的零值率较高，均超过 50%。可能的原因是该平台的玩家虽然愿意发表对游戏的评论与看法，但玩家与玩家之间缺乏合适的互动，也可能是不同玩家对同一款游戏的看法不尽相同，所以缺乏认同感，也有可能是由于玩家并没有注意到该互动功能。

表 3.1: 互动特征零值情况

特征名称	零值数	零值率
up	207253	61.09%
down	267257	78.78%
fun	296650	87.45%

图 3.2给出了游戏时长与评论字数的对数箱线图，从图中可以看出，总体上，正类玩家平均比负类玩家花更多的时间游玩游戏；正类游玩时间的中位数也要明显高于负类，这说明玩家倾向于在自己给出好评的游戏上花费更长时间。正类玩家的平均评论字数高于负类玩家评论字数，可以说明喜欢该游戏的玩家更愿意花费笔墨给予游戏好评。

可见，不同情感倾向的游戏评论在游戏时长与评论字数上的分布区别较大，故猜测加入评论字数与游戏时长两大特征可以使得情感预测更加有效。后续的模型也证明了这一结论。

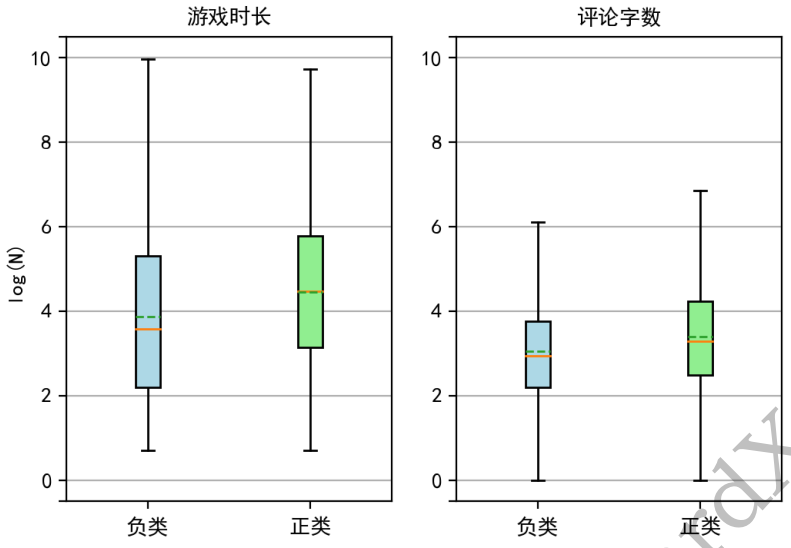

图 3.2: 游戏时长与评论字数对数箱线图

图 3.3给出了不同手机品牌用户的情感倾向分布。可以看出各个手机品牌用户对游戏的正向评论与负向评论所占比相差不大。华为、小米用户数量占比明显超过其他手机品牌用户。OPPO 与 Vivo 手机用户占比处于中等水平，而其他手机用户则占比较少。

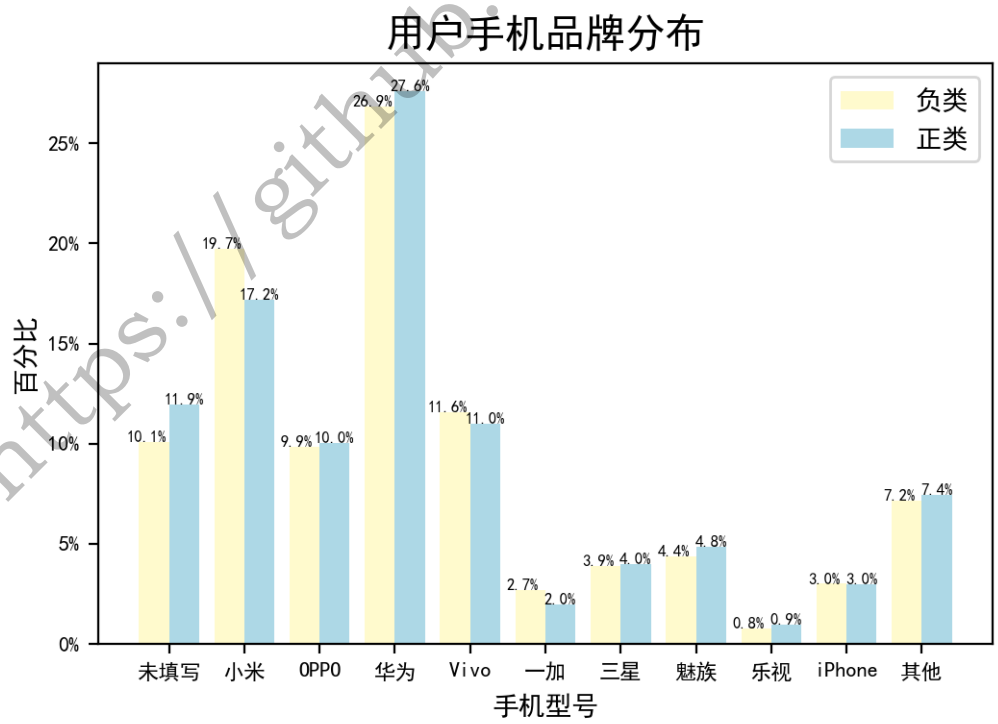

图 3.3: 用户手机品牌分布

4 预测模型介绍

4.1 模型选择

基于 TF-IDF 方法的文本情感分类算法有多种。例如逻辑回归，支持向量机，朴素贝叶斯，神经网络等。本文选用决策树模型及其基于 Boosting 思想的衍生模型作为文本情感分类器。原因在于：

- 决策树模型本身非常直观，具有强有力的解释性，且容易将决策树的路径转化为判别规则。
- 可以同时处理计量数据、计数数据与属性数据. 本文在预处理过程中，每个词的 TF-IDF 值为计量特征，赞同、不赞同评论的人数等为计数特征，而手机型号为属性特征。决策树模型无需额外的预处理，便可输入这些特征。
- 训练效率较高，在相对短的时间内能够对大容量的数据做出可行且效果良好的结果，本文所使用的样本数将近 33 万，决策树训练的时间成本较低。
- 决策树的分裂法则易于理解，选择此模型，可以在训练结束后，筛选出区分度排名较为靠前的特征以供后续的研究。

综上所述，本文使用基于决策树的模型作为文本情感分类器。

4.2 决策树模型介绍

决策树，顾名思义，是树状的统计模型。其思想非常直观，以分类任务为例，该模型把特征空间按照各特征取值划分成多个不同的区域。根据每个区域所落入的样本，按照一定策略，定义该区域样本所属的类。

CART 树是一种典型的决策树模型，由 Breiman 等人于 1984 年提出，该模型既可以完成回归任务，也可以完成分类任务。

CART 树由有向边和结点两部分组成。其中，结点有两类：

- 非叶子结点：该结点将当前区域根据一定决策条件再分为两个新的区域。
- 叶结点：每个样本最终都会属于一个叶子节点；在同一叶子节点中的数据说明有大概率属于同一类；每个叶子所属的类需要一定方法确定，例如投票法。

在生成 CART 分类树时，采用基尼指数 (Gini Index) 最小化准则。

基尼指数衡量了一个集合中各个样本取值的不确定性。其定义如式 4.1所示。

$$Gini(D) = 1 - \sum_{k=1}^{K}(\frac{|C_k|}{|D|})^2 \tag{4.1}$$

其中，$|D|$ 表示数据集 D 的样本容量，$|C_k|$ 表示数据集 D 中属于第 k 类的样本数量。

如果数据集 D 根据特征 Q 是否取值为 q 被分成了两部分：$D_1 = \{(X,y) \in D|Q = q\}, D_2 = D - D_1$ 那么在该条件下，集合 D 的基尼指数变为式 4.2。

$$Gini(D,Q) = \frac{|D_1|}{|D|}Gini(D_1) + \frac{|D_2|}{|D|}Gini(D_2) \tag{4.2}$$

可以看出，随着 CART 树的生长，数据集会不断细分，使得基尼指数变小，模型对训练集分类结果的预测会变得更为准确。可以想见，如果 CART 树分将数据集分得足够细致，使每个叶结点只

有一个样本，或每个叶结点下的所有样本属于同一类，此时的基尼系数变为 0。尽管对训练集的训练效果是极优的，但是必然会过拟合，这并不是理想的结果。性能良好的模型应当拥有较好的泛化能力。因此，在训练 CART 树前，应当先设置树的最高层数、最大叶结点树等参数，以防止树模型生长过深而导致过拟合。

决策树的生长方式主要有两种，leaf-wise 与 level-wise。前者遍历当前树的所有叶结点，找出能够使得损失函数减少最大的叶结点进行分裂。后者则不做选择，对当前所有叶结点进行再分裂。两种生长方式如图 4.1所示。

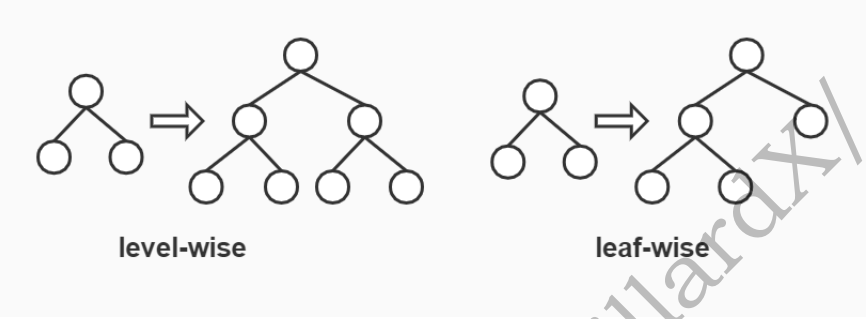

图 4.1: level-wise 与 leaf-wise 的区别

很明显，level-wise 的生长方式非常适合计算机，但其缺点也非常明显，计算的空间代价大，容易过拟合，产生不必要的分裂。尽管 leaf-wise 可能会长出较深的树，且计算时间可能会增加，但其准确率应当比 level-wise 高，且其缺陷可以通过提前设置参数避免。因此，本文采取 leaf-wise 的生长方式。

本文决策树算法如算法 4.1所示。

算法 4.1 决策树训练

输入：训练集 D，最大叶子结点数 C

输出：决策树 $T_C(x)$

for $m = 1 \to C$ do
    寻找最佳分裂 $(p_m, f_m, v_m) = FindBestSpilit(X, y, T_{m-1}, L)$
    进行分裂 $T_m(x) = T_{m-1}.split(p_m, f_m, v_m)$
end for

其中，寻找最佳分裂点的算法如算法 4.2所示。

算法 4.2 寻找最佳分裂点 $FindBestSpilit$

输入：训练集 D，当前树模型 $T(X)$，损失函数 $Gini$，当前树所有叶子结点 $Leaves$，特征 $Features$，所有可取的分割阈值 $Thresholds$

输出：最佳分割 (p_m, f_m, v_m) ▷ 注释：p_m 为选取的叶结点，f_m 为选取的特征，v_m 为选取的阈值

$\Delta Gini = 0$
for $p \in Leaves$ do
  for $f \in Features$ do
    for $v \in Thresholds$ do
      使用 (p, f, v) 分割当前树 $T(X)$，得到新的树 $T_{new}(X)$
      $\Delta Gini_{now} = Gini(D, T(X)) - Gini(D, T_{new}(X))$
      if $\Delta Gini < \Delta Gini_{now}$ then
        $\Delta Gini \leftarrow \Delta Gini_{now}$
        $(p_m, f_m, v_m) \leftarrow (p, f, v)$
      end if
    end for
  end for
end for

4.3 Boosting 集成思想介绍

当单个机器学习模型的效果并不十分理想时，人们会自然而然地想到训练多个模型，并将其联合在一起，达到的效果是否会有所改进。这种方法便是机器学习中的集成思想。集成学习的指导思路有许多种，Boosting 便是集成学习中重要的一种。可将 Boosting 模型看作是一系列模型的线性组合，如式 4.3所示。

$$F_M(X) = f_0(X) + f_1(X) + ... + f_M(X) \tag{4.3}$$

在训练过程中逐步确定每一个子模型 $f_i(X)$，叠加至复合模型中来，使损失函数随着子模型的增加而减少。

Boosting 模型的训练方法主要有两种。以分类问题为例，一种通过改变样本的分布（即各个样本的权值）来加强对分类错误样本的关注度，典例有 Adaboost 模型。一种通过改变训练目标来加强对分类错误样本的关注度，典例为梯度提升模型（Gradient Boosting）。

根据梯度提升模型训练的目标：使损失函数随着子模型的增加而减少。一个合理的想法便是新加入的子模型 $f_i(X)$ 可以使得复合模型 $F_i(X)$ 沿着损失函数关于 $F_i(X)$ 负梯度方向变化，这就是梯度提升模型的任务。

4.4 GBDT 模型介绍

当使用决策树模型作为梯度提升模型中的子模型 $f_i(X)$ 时，便得到了 GBDT 模型。

GBDT 分类模型本质上是 CART 回归树的堆叠。在二分类问题中，类似于逻辑回归模型，将复合学习器 $F_m(X)$ 的输出看作是对事件 $(y = 1|X)$ 与事件 $(y = 0|X)$ 的概率对数优比，即 $F_m(X) = log\frac{p}{1-p}$，其中 $p \doteq Prob(y = 1|X)$，则 $1 - p = Prob(y = 0|X)$。换算可以得到 $p = \frac{1}{1+e^{-F_m(X)}}$。

GBDT 模型的思想是：

- 若原数据 $(X_i，y_i)$ 的实际标签 $y_i=1$，则每增加一棵树，输出所得概率 $\hat{p_i}$ 能更加接近 1。
- 若原数据 $(X_i，y_i)$ 的实际标签 $y_i=0$，则每增加一棵树，输出所得概率 $\hat{p_i}$ 能更加接近 0。

基于上述假设，可以得到 GBDT 模型的对数极大似然函数，如式 4.4所示。

$$lnL=\sum_{i=1}^{N}[y_i ln\hat{p_i}+(1-y_i)ln(1-\hat{p_i})] \tag{4.4}$$

其中 N 为样本容量，$\hat{p_i}$ 为样本 X_i 输入当前模型 $F_m(X)$ 后所得的概率值 $\hat{Prob}(y_i=1|X_i)$。稍作修改，便可以得到 GBDT 的损失函数，如式 4.5所示。

$$Loss=-\frac{1}{N}\sum_{i=1}^{N}[y_i ln\hat{p_i}+(1-y_i)ln(1-\hat{p_i})] \tag{4.5}$$

同样，也不难得出单个样本 (X_i,y_i) 在模型 $F_m(X)$ 下的损失，如式 4.6所示。

$$\begin{aligned} J(y_i,F_m(X_i)) &= -[y_i ln\hat{p_i}+(1-y_i)ln(1-\hat{p_i})] \\ &= y_i ln(1+e^{-F_m(X_i)})+(1-y_i)ln(1+e^{F_m(X_i)}) \end{aligned} \tag{4.6}$$

在建立第 $m+1$ 棵树 $f_{m+1}(X)$ 时，我们希望新模型 $F_{m+1}(X_i)$ 的各个样本 (X_i,y_i) 的损失 $J(y_i,F_{m+1}(X_i))$ 能尽可能地小。所以子模型 $f_{m+1}(X)$ 的训练目标可以由式 4.7表示。

$$r_{m+1,i}=arg\min_{r} J(y_i,F_m(X_i)+r) \tag{4.7}$$

直观的想法是对 $J(y_i,F_m(X_i)+r)$ 求关于 r 的导函数，使导函数等于 0 以反向推导出 r 的值，然而该导函数的显式解并不可求。针对此问题，Friedman 利用损失函数的负梯度值作为提升树算法中 r 的近似值，从而得到子模型 $f_{m+1}(X)$ 的训练目标，如式 4.8所示。

$$\begin{aligned} r_{m+1,i} &= -(\frac{\delta J(y_i,F(X_i))}{\delta F(X_i)})_{F(X)=F_m(X)} \\ &= y_i-\frac{1}{1+e^{-F(X_i)}} \\ &= y_i-\hat{p_i} \end{aligned} \tag{4.8}$$

综上所述，GBDT 算法流程总结如算法 4.3所示。

算法 4.3 GBDT

输入：训练集 D，训练轮数 M，每棵树叶子结点数 K，总体损失函数 $Loss$，单样本损失函数 J

输出：GBDT 分类器 $F_M(X)$

初始化 $F_0(X) = ln\frac{\sharp(y=1|y\in D)}{\sharp(y=0|y\in D)}$

for $m = 1 \to M$ do

  for $(X_i, y_i) \in D$ do

    计算各个样本的残差作为第 m 棵树的训练目标：

$$r_{m,i} = -(\frac{\delta J(y_i, F(X_i))}{\delta F(X_i)})_{F(X)=F_{m-1}(X)}$$

  end for

  将 $(X_i, r_{m,i})$ 作为训练集训练子模型 $f_m(X)$，得到 K 个叶子结点

  for $k = 1 \to K$ do

    计算每个叶子结点的最佳取值：

$$c_{m,k} = arg\,\min_{c} \sum_{X_i \in R_{m,k}} Loss(y_i, F_{m-1}(X_i) + c)$$

▷ 注释：$R_{m,k}$ 为 f_m 第 k 个叶子结点中包含的所有样本

  end for

  更新分类器：$F_m(X) = F_{m-1}(X) + \sum_{k=1}^{K} c_{m,k} I(X \in R_{m,k})$ ▷ 注释：$I(\bullet)$ 为示性函数

end for

得到最终分类器：$F_M(x) = F_0(X) + \sum_{m=1}^{M} \sum_{k=1}^{K} c_{m,k} I(X \in R_{m,k})$

5 模型实施

5.1 模型实施

本章中的所有模型，均抽取相同的 10000 个样本作为测试集，对剩余数据采用五折交叉验证进行训练。

作为参照，本文使用基于 2.5 节获得的领域情感词典作为基础分类器。具体操作如下：

1) 扫描每个样本的分词结果，根据词典记录积极词与消极词的个数。

2) 计算每个样本积极词与消极词个数差，结果大于等于零判定为积极，否则判定为消极。

在进行决策树模型的训练时，本文采用了如下参数:

- 对结点再分割的最少样本数：20。
- 每个叶结点的最少样本数：5。
- 决策树的最大层数：15。

图 5.1分别给出了不带附加特征与带附加特征的决策树模型学习曲线。

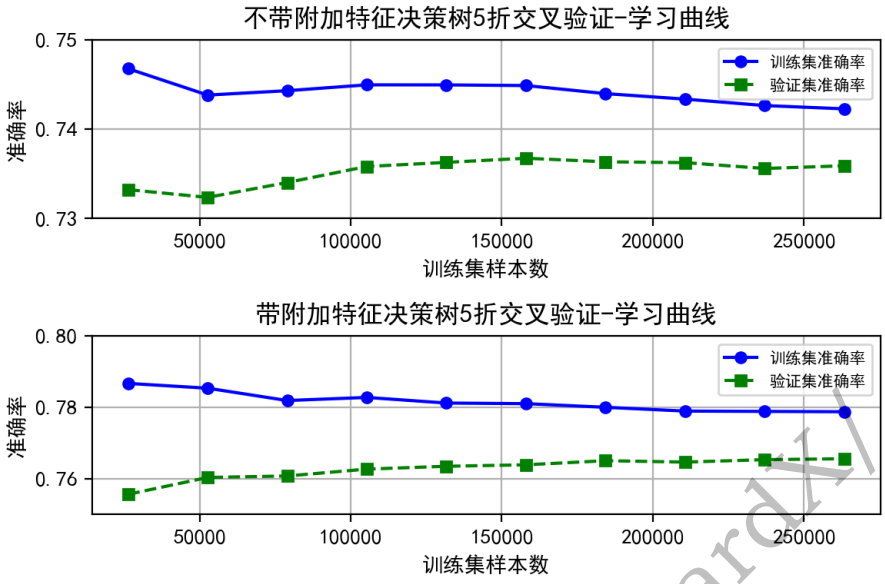

图 5.1: 决策树模型学习曲线

可以看出，训练样本量较少时，两个决策树模型对训练集的准确率都比对验证集的准确率高，但是随着训练样本量的增加，训练集和验证集的误差开始收敛。这说明，随着训练集样本量的增大，决策树的泛化能力增强，反映出的结果便是验证集的准确率在不断提升，向训练集的准确率趋近。另外，对比上下两幅图中的学习曲线，可以很明显地看出，在相同参数下，有附加特征的决策树模型准确率要比无带附加特征的决策树模型准确率更高。这说明，在预测游戏评论的情感倾向时，“游戏时间”，“评论字数”、“手机型号”，“觉得此评论有趣的人数”，“赞同此评论的人数”等评论内容以外的特征，能够提升判断评论情感倾向的准确率。从学习曲线的趋势还可以看出，决策树模型在本文的情感分类任务上基本达到了”偏差-方差权衡”。

对于梯度提升树的训练，采用与决策树模型相似的参数：

- 对结点再分割的最少样本数：20。
- 每个叶结点的最少样本数：5。
- 决策树的最大层数：15。
- 决策树数量：1000。

分别将有附加特征的数据与无附加特征的数据放入梯度提升树模型进行训练。

表 5.1展示了上述模型在测试集上的不同表现。(注：baseline 代表基于情感词典的分类器，DT 代表不带附加特征的决策树模型，DT_OF 代表带附加特征的决策树模型，GBDT 代表不带附加特征的梯度提升树模型，GBDT_OF 代表带附加特征的梯度提升树模型)

表 5.1: 模型评估

模型	准确率	精确率	召回率	F1 分数
baseline	0.698	0.840	0.749	0.792
DT	0.727	0.964	0.724	0.827
DT_OF	0.761	0.935	0.765	0.841
GBDT	0.807	0.928	0.814	0.867
GBDT_OF	0.830	0.931	0.836	0.881

可以看出，DT、DT_OF、GBDT、GBDT_OF 四个模型总体都优于基于情感词典的分类器。

通过 DT 与 GBDT 及 DT_OF 与 GBDT_OF 两组模型对比，可以发现除了精确率略低，GBDT 模型在准确率、召回率、F1 分数的表现上都明显地高于 DT 模型。这说明基于 Boosting 的集成方法对玩家的情感判别有显著效果。

通过 DT 与 DT_OF 及 GBDT 与 GBDT_OF 两组模型对比，可以发现加入评论字数、手机型号、有趣人数、赞同人数、不赞同人数、游玩时间这些非文本的附加特征，可以使模型的情感预测能力得到提升。这证明非评论内容的信息对于评论情感的预测是具有价值的。

5.2 特征重要性

5.2.1 GBDT_OF 模型的特征重要性

在树模型生长时，会在原有叶结点上选取最合适的特征进行再分裂。因此，可以认为，在训练过程中，一个特征被选作分裂特征的次数越多，说明该特征对评论信息情感倾向的识别能力越强。图 5.2根据上述准则展示了 GBDT_OF 模型中重要性排名前 20 的特征。

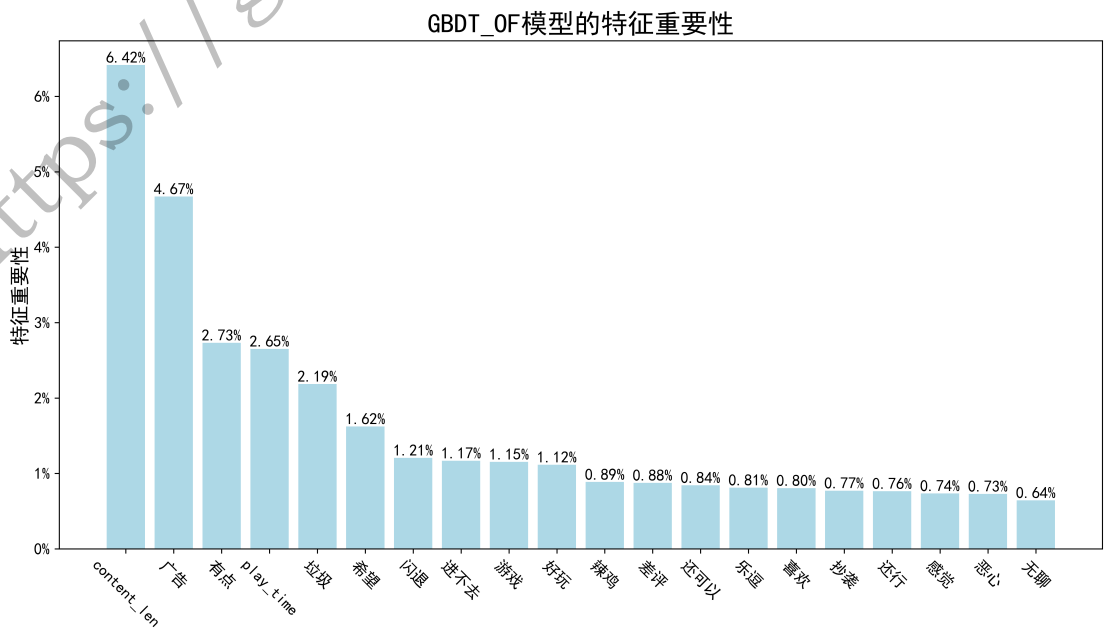

图 5.2: GBDT_OF 模型的特征重要性

从图中可以发现，在附加特征中，游戏时间与评论字数确实对模型的预测提供了较大的贡献；“广告”、“垃圾”、“差评”、“闪退”、“好玩”等表示游戏故障、玩家情绪的词语，同样提升了情感分析的准确度。

5.2.2 关于附加特征——游戏时长的讨论

从图 5.2中可以看出，特征“游戏时长”(play_time) 在所有特征重要性中排名前四，占比 2.65%。尽管该特征的重要性排名已经较为靠前。但是按照人们“对越是喜欢的东西，往往会用得越久”的行为倾向，可以作如下推断：

玩家在一款游戏上花费的时间越久，越能说明他对这款游戏的喜爱，那么他的评论内容也会倾向于正向积极的情感。因此，“游戏时长”是一个非常有力的特征。其重要性应当更为靠前。

产生这样的实验结果，一个可能的原因是，原始数据中特征“游戏时长”包含的空缺值太多。通过统计，发现该特征的空缺率高达 70%。过多的缺失值降低了决策树选取“游戏时长”作为分裂特征的倾向。

为证明上述推断，重新选取“游戏时长”无空缺的样本进行 GBDT 的训练，并展示该模型中重要性排名前 20 的特征，如图 5.3所示。

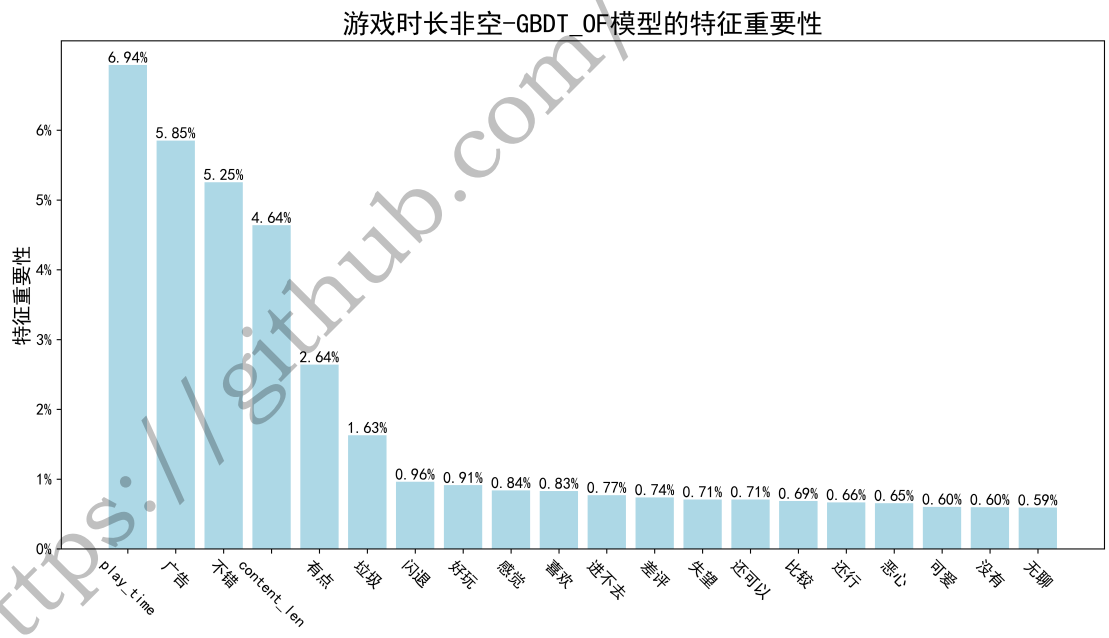

图 5.3: 游戏时长非空-GBDT_OF 模型的特征重要性

很明显，数据经过筛选后，“游戏时长”变成了 GBDT 模型最重要的特征，这证明了上述推断的正确性。

这对游戏行业的实际应用具有启示意义：玩家游戏时长数据对预测玩家情感具有巨大帮助，游戏平台应设法获取该数据。

6 结论与展望

6.1 模型评价

本文基于 TapTap 平台用户游戏评论数据。首先使用 Word2vec 方法构建了电子游戏领域的情感词典，并建立了基于词典的情感分类器。而后运用 TF-IDF 方法建立决策树与梯度提升树两类模型，预测用户情感倾向。实验结果证明，机器学习模型的效果优于词典分类器；Boosting 集成方法能够使原有的决策树模型效果明显提升。

此外，通过加入额外特征并与原模型对比，发现这些附加的相关特征对用户的情感倾向判别是具有帮助的。

最后，本文排列出了有附加特征梯度提升树模型的特征重要性，我们发现重要性占比较为靠前的特征有附加特征及文本中表示游戏故障、玩家情绪的词。通过进一步分析，发现在附加特征中，游戏时长对判别玩家关于此款游戏的情感倾向有着重要的作用。由此得到对应的商业应用启示：想要更为准确的判别玩家对游戏的喜恶，游戏时长是重要的评判标准，应该尽量获取相关数据。

6.2 未来展望

本文提出了基于 Boosting 集成思想的游戏评论情感分析方法，获得了较为良好的效果，但仍有以下不足：

1) 本文使用 TF-IDF 方法对分词数据进行了数字化处理。尽管特征的解释性强，但是产生的特征维度过高，特征矩阵过于稀疏会模型的训练有一定影响。后续工作可以考虑运用深度学习的方法进行处理。

2) 由于编码问题，本文的研究并没有考虑评论中的 emoji 颜文字，在预处理过程直接将其删除。尽管这些颜文字并非严格意义上的中文，但是这种文字在电子游戏领域的价值是很大的，它或多或少地反应了玩家的喜好与厌恶，对情感分析应该有非常重要的效果。后续工作可以考虑将 emoji 颜文字也作为文本特征来处理。

3) 不同的词语表达的情感强烈程度也是不同的。本文在建立基于词典的 baseline 分类器时，默认所有词的权重相同，会对情感分类效果产生一定影响。后续工作考虑对每个词的情感强烈程度进行量化标注。

4) 用户的评论涵盖着游戏的多个方面。一条评论可以包含用户对内容、设计、硬件兼容等各方面的不同情感。本文仅研究了单条评论整体情感倾向，后续工作可以基于方面（aspect）对游戏评论进行情感分析，研究用户对游戏各个方面的情感倾向。

参考文献

[1] 郭博，李守光，王昊，张晓军，龚伟，于昭君，孙宇. 电商评论综合分析系统的设计与实现——情感分析与观点挖掘的研究与应用 [J]. 数据分析与知识发现,2017,1(12):1-9.

[2] 李航. 统计学习方法 [M]. 北京：清华大学出版社,2012.137-152.

[3] 马超，余辉，夏文蕾，管为栋. 政务微博评论中情感极性分析方法研究——以上海公安机构微博为例 [J]. 现代情报,2020,40(03):157-168.

[4] 吴军. 数学之美 [M]. 北京：人民邮电出版社,2014.104-110.

[5] 郑新，张靖. 人工智能时代的教育游戏：发展机遇与趋势 [J]. 数字教育,2020,6(01):27-31.

[6] 张乐，闫强，吕学强. 面向短文本的情感折射模型 [J]. 情报学报,2017,36(02):180-189.

[7] Bais R, Odek P, Ou S. Sentiment Classification on Steam Reviews[J]. 2017.

[8] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[9] Kiran T D V, Reddy K G, Gopal J. Twitter sentiment analysis of game reviews using machine learning techniques[J]. Journal of Chemical and Pharmaceutical Sciences, 2010: 175-178.

[10] Lin D, Bezemer C P, Zou Y, et al. An empirical study of game reviews on the Steam platform[J]. Empirical Software Engineering, 2019, 24(1): 170-207.

[11] Luo R, Xu J, Zhang Y, et al. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation[J]. arXiv preprint arXiv:1906.11455, 2019.

[12] Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing[C]//Proceedings of the 2nd international conference on Knowledge capture. 2003: 70-77.

[13] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining[C]//LREc. 2010, 10(2010): 1320-1326.

[14] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques[C]//Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002: 79-86.

[15] Sirbu D, Secui A, Dascalu M, et al. Extracting Gamers' Opinions from Reviews[C]//2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). IEEE, 2016: 227-232.

[16] Strååt B, Verhagen H. Using User Created Game Reviews for Sentiment Analysis: A Method for Researching User Attitudes[C]//GHITALY@ CHItaly. 2017.

声　　明

本人郑重声明所呈交的论文是我个人在指导老师的指导下进行的研究工作及取得的研究成果，不存在任何剽窃、抄袭他人学术成果的现象。我同意（ ✔ ）/不同意（　）本论文作为学校的信息资料使用。

论文作者（签名）　徐晓杰

2020 年 5 月 15 日